

Clustering Web Pages to Identify Emerging Textual Patterns

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



The impact of the Web on Genre Repertoire

The Web has had a strong impact on the genre repertoire. Novel genres have already emerged such as *personal home pages*, *hotlists*, *FAQs*, and more recently *blogs*, *ezines*, etc. Additional genres are still taking shape, because the Web is still fluid and in constant change. It is often hard to assign a genre label to a web page. Many web pages still remain “unclassified”. These web pages might represent emerging textual patterns.

Experimental Question

Is it possible to detect emerging textual patterns NOW which might develop into new web genres or text types in FUTURE?

Experimental Setting

Web Page Collection

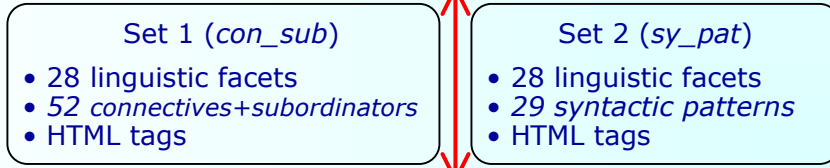
The SPIRIT collection is a random crawl carried out in 2001. It contains single web pages. It is unclassified and represents a genuine slice of the real Web.



K-means

- groups objects on the basis of similarity measures
- is suitable for large datasets
- is easy and fast

2 Sets of Features



2 Cluster Solutions

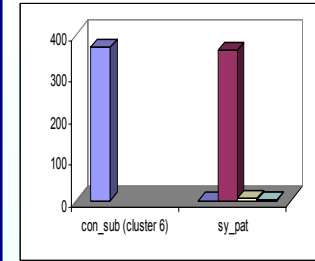


Fig. 2

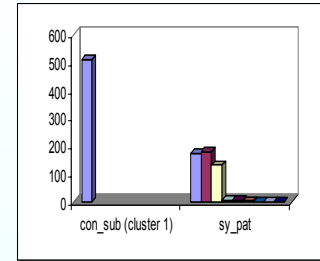


Fig. 3

Results and Discussion

Minority clusters show well-defined textual profiles, and are recognized as text categories by human assessment. These clear-cut clusters confirm that the approach is valid.

Two majority clusters can be interpreted as emerging textual patterns that we labelled as **contact web pages** and **quick information delivery** (Figures 1 and 2). It will be interesting to see whether they will develop into web genres or text types in future.

The largest cluster (Fig. 3), instead, shows a kind of mixed textuality and appears to be still too heterogeneous.

Conclusions

Results are encouraging and the approach seems to be effective in providing hints about emerging textual patterns. Two textual patterns identified in this experiment (**contact web pages** and **quick information delivery**) might become new web genres or text types in future.

Overlap between the 2 Solutions as a Measure of Stability

The extent of the overlap between the two cluster solutions indicates the degree of stability and reliability of the solutions.

Overlap across Minority Clusters

con_sub	sy_pat	Number of files
cluster 4	cluster 15	2
cluster 5	cluster 7	2
cluster 10	clusters 1 & 11	17
cluster 11	cluster 5	1
cluster 12	cluster 12	1
cluster 13	cluster 13	1
		24

Overlap across Majority Clusters

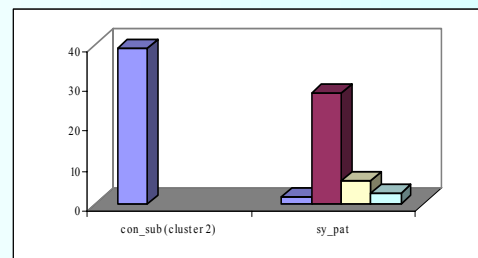


Fig. 1