Web-Specific Genre Visualization

Ivan Bretan, Johan Dewe, Anders Hallberg, and Niklas Wolkert
Telia Research AB, Vitsandsgatan 9, SE–123 86 Farsta, Sweden
+46 8 713 1000, ivan.p.bretan@telia.se

Jussi Karlgren
Swedish Institute of Computer Science, Box 1263, SE–164 28 Kista, Sweden
+46 8 752 1500, jussi.karlgren@sics.se

Abstract: User interfaces to WWW search engines typically present results as ranked lists of documents. Such lists give users little help in understanding document variation: we propose a richer representation of retrieval results in the search interface. Fundamental to us is the notion of document grouping. We use both stylistic genre-based document categorization and statistical content-based clustering, and organize documents along these criteria in a highly interactive visualization front-end to WWW search engines, enabling quick overview and incremental query refinement.

Introduction

The vast majority of user interfaces to WWW search engines are still based on an exceedingly simple interaction model where a linear list of hits, i.e. document items, is sorted after so-called "relevance" with inner workings and metrics hidden and all but incomprehensible to most users: "This is appealing in its simplicity, but users are often frustrated as they do not know what the results mean, nor can they control aspects of the search." [Shneiderman 1997] Usage problems stem partly from lack of overview and means of organizing the presentation of documents. These issues are addressed in a user interface framework known as Easify [Fig. 1].

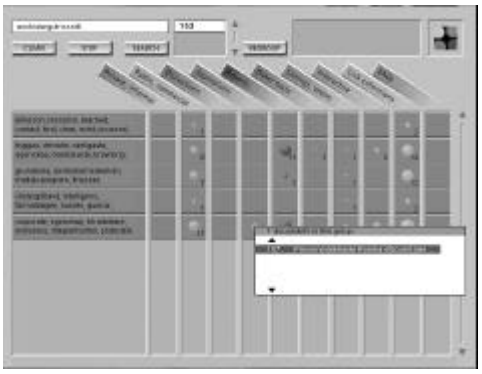

Figure 1: The Easify search interface.

Figure 2: The genre palette.

Documents can be grouped by topic as shown by Scatter/Gather [Cutting et al. 1992] or a variety of other criteria, such as site address, geographical location or title similarity. In the framework presented here, the notion of genre or stylistic variation is one of two main grouping criteria. Easify takes genre classification further than other similar tools by establishing a genre palette tailored to the typology of WWW documents, and applying automatic document classification algorithms accordingly. The second dimension made use of in our interface is statistical content-based clustering. Incremental refinement of the search process is supported by means of giving users increased control over pruning the search space through direct manipulation. Envision [Nowell et al. 1997] is another example of a search interface relying on multi-dimensional document visualization.

Stylistic Genre Classification vs. Content-Based Clustering

Stylistic items can be found on any level of linguistic abstraction: lexical, syntactic, or textual; each is of little import in itself, but taken together their variation indicates systematic differences [Biber 1989, Karlgren 1998]. We make use of dozens of items to classify documents into genres: sets of documents with a consistent tendency to make the same stylistic choices. Useful genres must be based on differences known and recognized by readers. To this end, we have created a genre palette and collected a test corpus tailored for our trial users and the WWW. We have interviewed 102 users on their perceptions of what types of material they find and interact with on-line. These impressions are used to define a palette of genres [Fig. 2] both reasonably consistent with what users expect and conveniently computable using measures of stylistic variation [Dewe et al., 1998]. Our test corpus was used to train a categorization tool. A relatively large number of textual features are calculated for each individual text and combined into simple if-then categorization rules using the C4.5 machine learning tool set [Quinlan 1993]. The features we use here are rather lexical in nature, for ease of processing: the relative frequency of certain classes of words such as personal pronouns, emphatic expressions, downtoning expressions, etc. We add more general textual and genre-specific features: relative number of digits, or average word length, for instance. Others yet are vectored specifically to the web material we have been using for training: number of images or proportion of HREF links in the document, among others. The genre determination algorithm in its current state makes sub-optimal classifications too often. Flexible genre determination would help here – a document should be allowed to fall into several genres rather than exclusively one.

Grouping of documents based on textual content, as opposed to style, is based on traditional statistical term-frequency based metrics. Since the emphasis is on a high degree of interactivity, a quick and rudimentary clustering must be used for the initial document sets. We assume that a small number of clusters in the interface is desirable. Initial clustering can be achieved by defining the first clusters on a few (10-50) randomly selected documents. The clustering itself is a variant of the standard metric: a hierarchical agglomerative group-average algorithm [Jain & Dubes 1988]. After deciding the first *n* clusters (with *n* user-adjustable) the following documents are each routed to one of them. A simple assign-to-nearest algorithm is used to decide cluster membership. Cluster headings are currently sets of significant keywords, and could definitely be improved upon.

Visualization of the Document Space

Following query submission, documents found relevant by the underlying search engine are sorted into *nm* slots, where *n* is the user-defined number of content-based clusters and *m* is the system-defined number of genres (currently *m*=10, following the results of the genre survey and viable screen real estate usage). The group of documents in a slot is presented in the shape of a “bubble” growing as search results become available. Although the number of slots may well be in the range of 30-90, the complexity of navigation is reduced through the clear two-dimensional lay-out with genre headings spelled out along the x-axis and content clusters along the y-axis (Fig. 1). When the cursor is moved across a bubble, a pop-up window is displayed with a list of all the documents contained within the group. This list can be presented using title, URL, size, date and other attributes. Clicking on a single document brings up the abstract, and double-clicking sends the URL to the default web browser to be displayed in its entirety. In order to “zoom in” on a subset of the resulting documents, bubbles can be dragged onto the “regroup” area in the top right corner of the main screen. When regrouping is applied, a second screen appears where only the documents contained in the selected groups are candidates for creating new bubbles. In the subselection, distribution according to genre remains the same as in the superset case, but the content-based clusters may differ.

References

[Biber 1989] Biber, D. (1989). A Typology of English Texts, *Linguistics*, 27 (3).

[Cutting et al. 1992] Cutting, D.R., Karger, D.R., Pedersen, J.O., and Tukey, J.W. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. *Proceedings of the 15th SIGIR Conference*, Copenhagen, ACM Press.

[Dewe et al., 1998]. Dewe, J., Karlgren, J., and Bretan, I. (1998). Assembling a Balanced Corpus from the Internet. *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Copenhagen.

[Jain & Dubes 1988] Jain, A.K and Dubes, R.C. (1988). *Algorithms for Clustering Data*, Prentice Hall.

[Karlgren 1998] Karlgren, J. (1998). Stylistic Experiments for Information Retrieval. Strzalkowski, T. (ed.) *Natural Language Information Retrieval*, Tomek, Kluwer.

[Nowell et al. 1997] Nowell, L.T., France, R.K., and Hix D. (1997). Exploring Search Results with Envision. *Proceedings of CHI '97*, Atlanta, ACM Press.

[Quinlan 1993] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

[Shneiderman 1997] Shneiderman, B. (1997). Designing Information-Abundant Web Sites: Issues and Recommendations. *International Journal of Human-Computer Studies*, Academic Press, 47 (1).