Preprint of paper; final version available as:

ATTFIELD, S., BLANDFORD, A. & CRAFT, B. (2004) Task Embedded Visualisation: The Design for an Interactive IR Results Display for Journalists. *Proc. IEEE Information Visualisation 2004*. 650-655.

Task Embedded Visualisation: The Design for an Interactive IR Results Display for Journalists

Simon Attfield, Ann Blandford, Brock Craft
University College London Interaction Centre
4th Floor, Remax House, 31/32 Alfred Place
London, WC1E 7DP, Great Britain
s.attfield@cs.ucl.ac.uk, a.blandford@cs.ucl.ac.uk, brock@craft.org

Abstract

There is need for user-centred visualisation research to engage with the activity context, information needs, knowledge and abilities of target user-groups. With a focus on the work of journalists, we first argue for information-retrieval results visualisation as a suitable browsing framework for journalists' frequently ill-defined needs and high-recall searches. We then describe the design and rationale for a histogram-based visualisation for journalists. We also describe the integration of this idea within a system that structures searching as a two-step query-and-filter operation. This approach is intended to support initial exploratory browsing and refinement in a way that is sympathetic to the systematic focusing that naturally occurs during complex, unstructured task performance. Further, the use of an enduring 'base' results set is intended to encourage structural familiarity with the broader results and therefore to enhance navigation.

1: Introduction

The landscape of visualisation research provides a rich source of concepts and tools for the representation of interactive, abstract information spaces. Several approaches have been proposed for the representation of document collections and collection sub-sets as tools for supporting document retrieval. Notable approaches include systems that represent document collections in terms of hierarchical classification schemes, such as Hearst's Cat-a-cone [1], systems that represent document content in terms of the distribution of query term occurrences, such as TileBars [2] and work by Byrd [3], and systems that represent topic clusters within a document collection, such as SPIRE [4 & 5] and Bead [6]. Some work has also been specifically concerned with the visual representation of news report document spaces—the topic of this paper—such as Galaxy of News [7] and BreakingStory [8].

Graphical information-retrieval displays and interactive dynamic querying hold the promise for the smoother integration of technology with the tasks that bring users to need information in the first place. However, the starting point for this paper is that, amidst explorations of the visualisation solution space, design is often based around intuited and implicit ideas of users and task situations, rather than these being empirically grounded and explicit. This is in contrast to user-centred research in Information Science, in which researchers have increasingly emphasised the need to understand the 'real-life' contexts for information seeking and information use in order to inform information systems design (see for example [9] & [10]).

We argue that there is a need for user-centred visualisation research that engages with:

- the activity or work context of user-groups that might bring them to use a system;
- the specific problems that this broader activity evoke—and for which a system might provide support;

- the typical behaviours, knowledge and abilities that target user-groups bring to a system interaction;

Whilst in practice, homogenous user-groups may not always be easy to identify (library users, for example, are notoriously diverse), there are identifiable groups organised around particular kinds of activities, such as professions, who have particular kinds of needs and abilities.

The research described in this paper forms part of a wider endeavour to explore the design of integrated information retrieval and authoring systems with particular emphasis on the needs and work of journalists. This has involved field and lab studies examining the information seeking and information manipulation behaviour of journalists, and also how this relates to the wider task of writing (See [11] & [12]).

In this paper, we describe the design of an information retrieval results visualisation intended to form part of an integrated system we are developing with newspaper journalists in mind. In section 2, we argue that visualisation is a particularly appropriate approach to results-display, given journalists' typical search skill and needs. In section 3, we describe a histogram-based results visualisation and present its rationale. In section 4, we describe a two-step approach to low-focus and high-focus searching intended to support easy, exploratory search refinement whilst simultaneously promoting users' familiarity with a base results set; we explain why this is useful for our particular target user-group. Finally in section 4, we summarise the paper and discuss future work.

2: Journalist end-user searching and the need for visualisation

Over the past 20 years, electronic news cuttings (ENC) databases have become increasingly important as research tools within news organisations. Initially, systems appeared in news libraries and were predominantly accessed by news librarians acting as search intermediaries to journalists. But before long, access at the journalist's workstation saw journalists performing many of their own searches. Today, searching an online cuttings archive is a daily activity for many news journalists.

Journalists search ENC archives frequently, but they also tend to be unsophisticated searchers, and certainly lack the training and skill of a professional librarian. A study by Nicholas [13] compared journalists' searches to those of their news librarian counterparts, and found that the journalists utilised far fewer search operators, used fewer search terms—typically only two—and searched cuttings archives predominantly by subject only (rather than additionally specifying author or date fields, for example). From our own observations, knowing how to refine a search to improve results precision presents journalists with particular difficulties.

Surprisingly though, Nicholas also found that the journalists were frequent system users and were often satisfied with their search outcomes. To reconcile low search expertise with relatively high levels of satisfaction, Nicholas [14] suggests that large results sets, which simple, poorly refined searches typically return, offer the journalists a browsing framework that is well-suited to their frequently ill-defined needs. As Nicholas puts it, "an over-refined search would not provide sufficient noise to feed off" (14, p.230).

The general suggestion, which has been made by other researchers (see for example [15]) is that high-recall searches, as Nicholas observed, are useful for users engaged in complex, constructive and uncertain information work, who are consequently often unsure of what they are looking for—this being particularly true during the early stages of an assignment [16][17]. Such users often wish to explore an information space in order to develop their understanding and to explore assignment opportunities and possibilities [12].

The question of designing search results displays for journalists using news cuttings archives, then, in part becomes one of supporting this high-recall search strategy. This need becomes even more evident when one considers that ENC archives, which may store articles from thousands of publications—many of which may cover the same story on the same day—are typically vast. At the time of writing, LEXIS NEXIS Professional, one of the most widely used commercial news cuttings services, was advertised as offering content from over 35,000 publications, in some cases going back 30 years [18]. LEXIS is designed such that searches that produce more that 1000 hits result in an error message and an instruction to the user to refine the search. It has been estimated that as much as 90% of LEXIS searches by journalists at News International result in this message [19].

Dealing with large results sets, then, is a real problem for journalists. But traditional text-based results displays are poor at supporting high-recall searches. Users are forced to explore their results by navigating through an ordered series of local textual displays, each providing a view of only a small set of

documents. And the larger the results set, the longer and more unmanageable the series. Navigating large, textually displayed results can be time-consuming, and overwhelming to the user; not least since the approach does not lend itself to providing overviews through which global features and areas of particular interest can be comprehended quickly and easily.

3: A histogram-based overview

The issues we have raised so far argue in favour of visualisation as a basis for displaying search results to journalists, but leave open what a useful display would be. In addressing this question, we have focused on two primary considerations:

The first consideration relates to the episodic, temporal character of news and its related reporting. Documents on ENC archives usually report, or relate in some way, to an event, and this is nearly always an event occurring around the time of publication. This is obviously true for news reports, but is also the case for opinion articles and usually for features (which typically provide background about a person or issue in the news).

Date of publication, then, represents an informative dimension. Documents which are about the same event will naturally cluster on this dimension, giving structure to a results set through which a user can orientate their browsing. Further, by visually locating document representations on this dimension, a user can superimpose their own event schemas or scripts as an aid to navigation (*e.g.* a report of someone's conviction comes about a week before a report of their sentencing and after a report of their arrest, a report of an air crash comes a few months before the report of the findings of the crash investigation *etc*.). In this way, representing documents on a temporal dimension might leverage a journalist's domain knowledge to dramatically narrow a large document space.

The second consideration is simplicity. One of the main reasons that journalists lack sophisticated searching skills is that, despite good intention, computer training is always secondary to meeting tight publication deadlines; often they receive no training at all. One of the challenges for information systems design is to provide meaningful collection overviews in which patterns can be easily recognised [20] and we regard this as particularly true for journalists. Hence, there is a need for a familiar representational form with meaningful axes that are simple in derivation and so easy for the untrained user to understand at a glance.

The design that we have chosen uses a histogram to display a summary of search results in the form of a frequency distribution on a date-of-publication scale. A sample screen-shot mock-up is shown in figure 1. In the display, horizontal bars are formed from squares, each of which represents a single article published during a given month that matches the user's query. Each document is intended to be selectable from the representation to display the full-text article.

The distribution of the 273 documents represented in the example corresponds with the results of a real search for articles within The Times/Sunday Times archive that appeared between 1997 and 2002, and match the phrase query "train crash". In addition to it being evident from the display that hardly a month went by during that period when an article containing the phrase "train crash" did not appear, it is also evident that there are also some peaks in articles containing this phrase. Notably, five of these peaks (indicated on the figure) correspond to five major UK rail disasters that occurred during this period: Southall, Paddington, Hatfield, Selby and Potters Bar.

The display shown in figure 1 provides a structured graphical overview of a moderately large IR results set that is intended to help the user orientate their exploration around emergent reporting clusters, and also by exploiting their own schemas of how news stories unfold over time. This representation, we argue, is far more suitable as a basis for the exploring of high-recall search results than a traditional text-based results display.

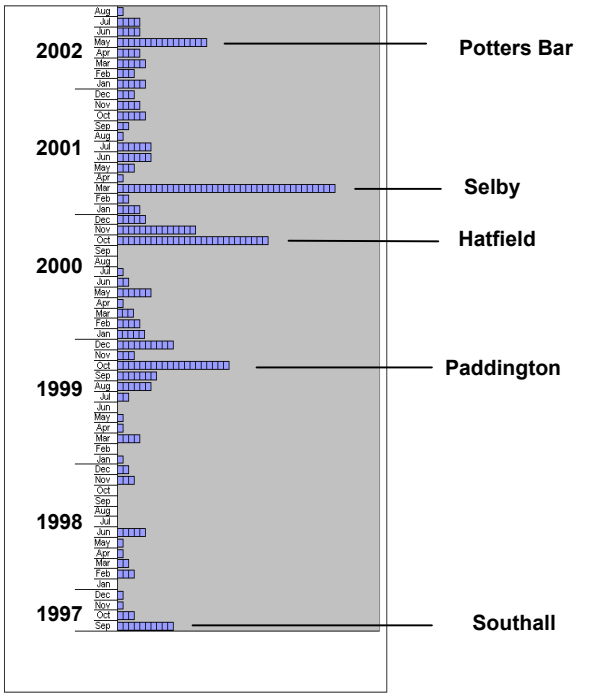

Figure 1. Results for the query 'train crash' represented on a time-based histogram. The peaks in reporting relate to major UK crashes.

4. Search refinement by secondary filtering

The design presented so far engages with the idea that journalists tend to perform high-recall searches and that they value browsing large results sets. However, during the course of a reporting assignment a journalist may evolve specific information needs as a function of the systematic focusing that occurs during the performance of complex, unstructured tasks such as writing assignments [12]. However, as discussed above, a lack of search knowledge makes refining searches problematic.

The solution that we propose is a design that structures search as a two-step process. The principle is that, in the first instance, users can submit a high-recall query around a topic of interest and the results are displayed in overview to support exploratory browsing (as in figure 1). Adjacent to this display is a panel with controls to enable additional filtering of the results of the type proposed by Schneiderman [21]. The entire interface, (including a full text document being displayed) is shown in figure 2. According to our design, rather than causing documents to disappear from the summary, colour coding is used to highlight documents that additionally match the filter combination.

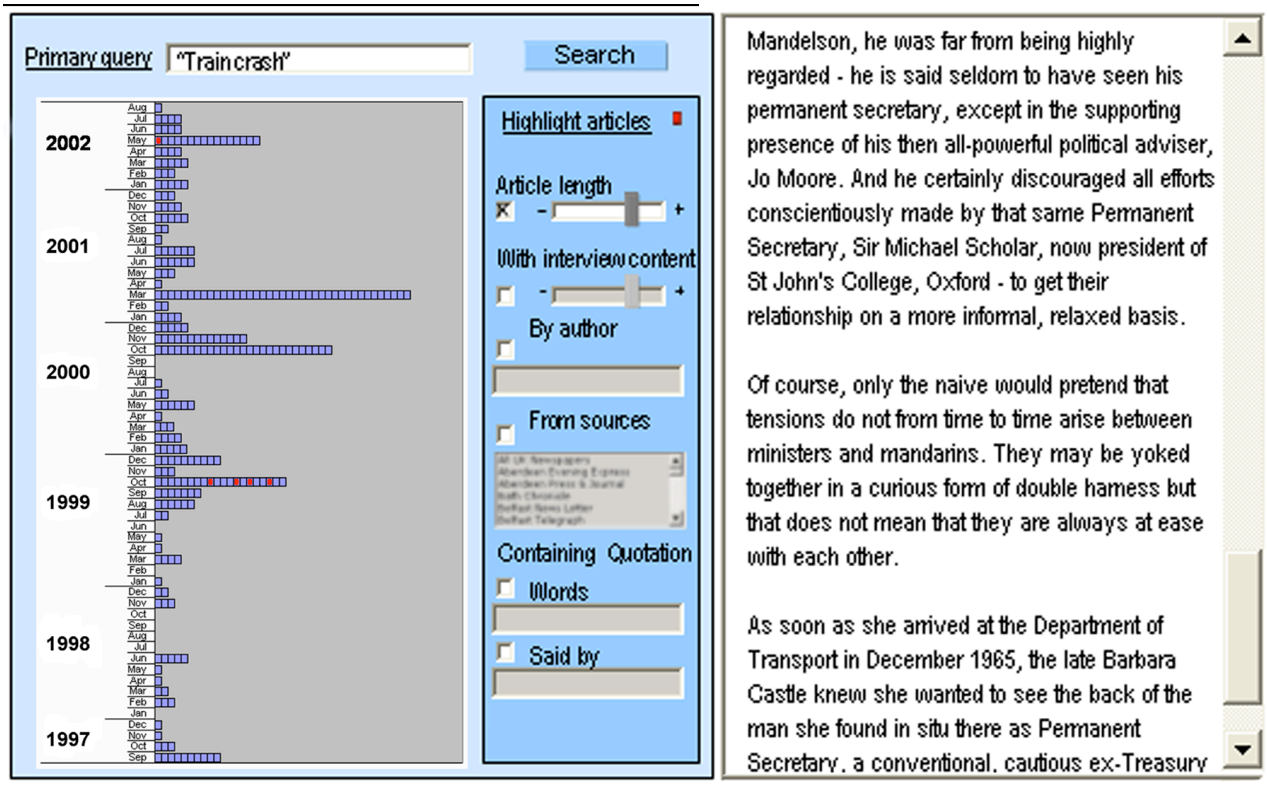

Figure 2. Integrating the results visualisation within a system that

The rationale for this two-step search arises from two issues:

First, by providing visual controls for search refinement, users with a limited knowledge of Boolean search operators should nevertheless be able to refine and produce more complex searches. Further, tight-coupling between the filter controls and the display would enable users to instantly see the effects of different refinements. This should encourage the exploration of refinement possibilities.

Second, superimposing focussed enquiries over a single, more enduring 'base' search, should promote structural familiarity, and therefore enhanced navigation, with respect to the broader results set. One reason this could be useful is that, in complex information work users are not always able to categorically judge the relevance or 'value' of information as it is encountered [17]. This is particularly true of journalists since they work in the context of wider, dynamic situational factors that frequently cause revisions to be made to task constraints mid-assignment [11]. For example, new information can come to light about an event being reported, and revisions can be made to editorial decisions about the approach or angle being adopted. As a consequence of this, journalists often wish to return to documents to review information. By maintaining the representation of a single broader results set with which the user can become familiar, document relocation should be made easier. In addition, of course, document representations can be coded to indicate whether they have been read before.

In terms of what filtering to provide, our research has suggested a number of possibilities. Early in an assignment a common journalist's goal is to quickly develop a deeper knowledge of the background of an issue to enable better interpretation of recent events or perhaps in preparation for interviewing. Typically, this is done by submitting a broad topic search and then browsing the results for two or three feature length pieces. If the focus of research is a person, then they may look for the last big interview. Some journalists also tend to focus on articles from particular sources or by known specialist journalists. In figure 2, we show filters in the control panel that will draw the user's attention to documents above a user-defined length threshold, 'interview-content' threshold, and those by a given author or from given source.

A more specific need that often occurs is to find a quotation. This often occurs in relation to a broader comparison that is being made between something said recently, perhaps by a politician, and what has been said in the past. In some circumstances the speaker may be defined, in others they may not. At election times, newspapers will even hand-generate and index documents of quotations to support this particular need. To address this need, figure 2 shows a filter for identifying documents incorporating particular terms within quotations.

5 Summary and future work

The visualisation we have described in this paper forms part of a wider project aimed at exploring the design of integrated information retrieval and authoring systems with an emphasis on the work of journalists. Journalists are high-recall searchers and they like to browse, which makes visualisation a very appropriate mode of results representation. Based on the temporal nature of news and the need for simplicity, our design uses a time-based histogram showing a frequency distribution of documents from which full texts can be selected for display. We have shown how, on such a display, documents naturally cluster around surges in reporting and we argue that this would help the journalist to orientate exploration, and to exploit their knowledge of how news stories unfold over time.

We have also proposed a two-step search process according to which the user performs a broad topic-level search and then superimposes refinement using dynamic visual filters to control document icon colour coding. This approach, we argue, would help the low-skill searcher explore refinement possibilities whilst also promoting familiarity and hence navigation of the primary search over the course of a work assignment.

Our approach has been to base visualisation design on our target user-group's work, the nature of their information needs and their search behaviour and ability. Early design briefings with News International and BBC News have received positive responses and have opened the door for future collaboration.

Future work includes integration of the visualisation into an existing prototype information retrieval and authoring tool and performing evaluations with the target user-group. In addition to assessing general acceptability, we intend to evaluate issues of scalability of the display and also the claims we have made in this paper. These include: assessing the extent to which the display offers a useful tool for orientating exploration; the extent to which users can leverage their own news story schemas for navigation; the advantages the tool provides in supporting low-skill users to refine their searches; and the extent to which the two-step

search approach promotes search familiarity through measures such as search recall and recognition.

Acknowledgements

The work reported here is supported by EPSRC grant GR/S73723.

References

[1] M.A. Hearst & C. Karadi, "Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results Using a Large Category Hierarchy", *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* July 1997, pp. 246-255.

[2] M.A. Hearst, "TileBars: Visualisation of Term Distribution Information in Full Text Information Access". *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, 1, May 1995, pp. 59-66.

[3] D. Byrd, "A Scrollbar-Based Visualization for Document Navigation", *Proceedings of the 4th ACM International Conference on Digital Libraries,* August 1999, pp. 122-129.

[4] J.J. Thomas, "Information Visualization: Beyond Traditional Engineering", *Human-Computer Interaction and Virtual Environments*, National Aeronautics and Space Administration, (NASA Conference publication 3320), 1995.

[5] J.A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur & V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents", *Proceedings of the IEEE Symposium on Information Visualization,* IEEE, October 1995, pp. 51-58.

[6] M. Chalmers & P. Chitson, "Bead: Explorations in Information Visualisation". *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* June 1992, pp. 330-337.

[7] E. Rennison, "Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes", *Proceedings of the ACM Symposium on User Interface Software and Technology*, November 1994, pp. 3-12

[8] J.A. Fitzpatrick, J. Reffell & M. Aydelott. "Breakingstory: visualizing change in online news", *Interactive posters: Proceedings of ACM CHI 2003 Conference on Human Factors in Computing Systems*, 2, April 2003, pp. 900-901.

[9] T. Wilson, "Models of Information Behaviour Research". *Journal of Documentation*, 55(3), 1999, pp. 249-270.

[10] C.C. Kuhlthau & S.L. Tama. "Information Search Process of Lawyers: A Call for 'Just for me' Information Services", *Journal of Documentation*, 57(1), 2001, pp. 25-43.

[11] S. Attfield & J. Dowell, "Information Seeking and Use by Newspaper Journalists", *Journal of Documentation*, 59(2), 2003. pp. 187-204.

[12] S. Attfield, A. Blandford, & J. Dowell, "Information Seeking in the Context of Writing: A Design Psychology Interpretation of the 'Problematic Situation'", *Journal of Documentation*, 59(4), 2003, pp. 430 - 453.

[13] D. Nicholas, "An Assessment of the Online Searching Behaviour of Practitioner End Users", *Journal of Documentation*, 52(3), 1996, pp. 227-251.

[14] D. Nicholas & H. Martin, "Should Journalists search Themselves? (And What Happens When They Do?)", *Proceedings of Online Information*, 1993, pp. 227-234..

[15] C. Cole, B. Mandelblatt & J. Stevenson, "Visualizing a High Recall Search Strategy Output for Undergraduates in an Exploration Stage of Researching a Term Paper", *Information Processing and Management*, 38(1), 2002, pp. 37-54.

[16] C.C. Kuhlthau. "A Principle of Uncertainty for Information Seeking", *Journal of Documentation,* 49(4), 1993, pp. 39-55.

[17] P. Vakkari. "A Theory of the Task-based Information Retrieval Process: A Summary and Generalisation of a Longitudinal Study", *Journal of Documentation,* 57(1), 2001, pp. 44-60.

[18] *LexusNexis Professional product information*
url: http://207.24.42.51/page_63.html
[Accessed March 24, 2004]

[19] Iley, L. (2004) (Information Services Manager, News International) personal communication, 1/3/2004.

[20] B. Shneiderman, D. Feldman, A. Rose, & X.F. Grau, "Visualizing digital library search results with categorical and hierarchical axes", *Proceedings of the 5th ACM Conference on Digital libraries,* June 2000, pp. 57-66.

[21] B. Shneiderman, Dynamic queries for visual information seeking, *IEEE Software*, 1994, 11(6), pp. 70-77.