PENG: Integrated search of distributed news archives

Mark Baillie
mb@cis.strath.ac.uk

Fabio Crestani
fabioc@cis.strath.ac.uk

Monica Landoni
monica@cis.strath.ac.uk

Computer Information Sciences Department
University of Strathclyde, Glasgow, UK

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Design, Human Factors

Keywords: Personalisation, News search

1. THE PENG SYSTEM

News professionals, such as Radio, TV and Newsprint journalists and editors, now have at their disposal a large and varied collection of digital information resources. News Agencies such as ANSA, Reuters and AP can, for example, provide live feeds of breaking stories directly into a newsroom. Journalists can also search and browse a variety of online news archives, digital libraries and web repositories when researching and compiling a report. However, to utilise this wealth of digital information, it is expected that the busy news professional is proficient in a number of systems or interfaces. The aims of PENG, a European Union funded Specific Targeted Research project, is to address this issue of news content management allowing the news professional to access information under a single interface. PENG integrates several key tasks, including personalised filtering, retrieval, and presentation of multimedia news, into a single system. In this poster we provide an overview of PENG, describing our approach to constructing a dedicated retrieval and content management system for a specific user group. We aim to show how detailed knowledge of a user group and the information tasks they perform can be used to inform the design of retrieval and filtering system components.

Requirements

The design of PENG is based on a study, undertaken as part of the project, of the work practices of journalists from different mediums. The study investigated how journalists gathered information, and how they exploited current systems to complete their daily work tasks. Several important requirements came out of this study. Firstly, journalists required a high level of control over the operation of a system due to the fear of missing important information. Ideally, they wish to view all potentially relevant documents across a number of distributed archives and information providers. However, given the vast quantity of digital information readily available, journalists are now forced to consider some form of automated content management or filtering.

Journalists used a range of criteria when gathering information for a task and the selection of these criteria was found to be dependent on the task itself, the journalist, the work environment and also the intended audience. Importantly, these criteria were not static but constantly changing as the journalist and environment changed. In particular, it was identified that the resources the journalist searched for information (i.e. news archives, digital libraries or web resources), is also dependent on the journalist and their individual tasks and needs.

Finally, the notion of trust is important in assessing information. Those interviewed believed it vital to be able to identify the original document source, for example, in order to determine the accuracy of content, political motivation, etc. Although, reliability on the whole is difficult to judge and in some scenarios overridden by other factors including the speed in which a news report is available. For example, journalists who compile radio reports, indicated that easy and rapid access to information was as important as reliability. As a consequence, developing a system for journalists requires to take into account these preferences and practices as well as providing the technical infrastructure to support information access.

System Functionality

PENG combines filtering, searching and presentation within a unified framework. During the filtering or *push* phase, news streams are monitored for relevant information based on a personalised profile defined for each user. Journalists can set up dynamically updating filters (known as interests) which aid in the identification of potentially relevant breaking stories. Using fuzzy clustering techniques [2], the push phase also automatically and dynamically organises the incoming newswires based on the information needs of the journalist and also topically, assisting content management.

PENG also allows journalists to search topics across a large number of local and distributed news archives, digital libraries and web repositories within a single service, called the *pull* phase. This is achieved using the latest Distributed Information Retrieval (DIR) technology, allowing a journalist to search all potentially available resources using a single system accurately and efficiently.

Finally a *presentation* phase displays the results of both *push* and *pull*. Apart from notifying the journalist of relevant breaking stories and providing an interface to the search service, the user interface module allows the journalist to take notes, store retrieved information for report compilation, edit and prepare documents, and update his/her profile information. These activities were shown to be essential for using the information retrieved by the system in accomplishing the journalists' work tasks. The interface also provides support for the application of techniques such as news multimedia document visualisation and summarisation.

Personalisation

For each user, a profile is defined containing information that can be adopted for personalising both filtering and search based on the information needs of that user. The user profile is dynamic, designed to change over both time and task to meet the changing user needs (by learning/training techniques). Information retained in the profile includes a measure of the journalist's trust in specific information resources and news agencies, filters represented by fuzzy clusters, as well as previous query history and interaction. By utilising these statistics held in the journalist's profile, each phase of PENG can be personalised. This is a key motivation, to move away from the traditional paradigm of providing results for an "average" user. PENG is instead motivated to personalise filtering and retrieval specifically to the individual information needs of each journalist. For example, it was highlighted during the requirements study that local journalists consider regional stories of more importance than national or world news during the compilation of reports in a working day. Therefore, news stories reporting local issues should be given more importance when specified.

2. DISTRIBUTED IR PHASE

The *pull* phase allows the journalist to provide background context on breaking news stories, deepen their knowledge of the story and/or assist during the compilation of news reports or articles. The input to the retrieval component can be either *pushed* news documents explicitly selected to "deepen" the topic by the journalist or from ad-hoc querying. In the latter case, automatic queries are formulated from *pushed* documents, minimising workload for the journalist, by extracting query terms based on term importance. If available, queries are also expanded using terms from the journalist's interest determined by the filtering module during the *push* phase. The refined query is then used to search multiple distributed collections available to PENG. Searching distributed resources within an integrated framework involves three tasks: resource description acquisition, resource selection and data fusion[3]. We now discuss how PENG addresses each issue in turn.

Resource Description Acquisition

Journalists interact with a variety of resources and an integrated system must search across resources for a single information need. This means that we must obtain a description of the resource to be searched and this is an important stage because the perceived quality of such representations will impact on resource selection accuracy and ultimately retrieval performance. PENG uses Query-based Sampling (QBS) for the acquisition of resource description information [3]. Our approach is based on measuring the Predic-

tive Likelihood (PL) of the journalist's information needs given the estimated resource description. This provides an indication of the description quality and indicates when a sufficiently good representation of the resource has been obtained [1]. Integrating PL as part of the QBS algorithm, performance was improved both in terms of efficiency and effectiveness when compared to currently adopted threshold based stopping method, minimising overheads while maintaining performance. Our approach is fundamentally different to existing work which measure the quality of an estimate against the actual resource. This requires full collection knowledge which is not readily available except in an artificial environments and is not realistic for journalists who are searching real information resources. PL requires that only a set of queries are available for evaluating each resource description. In PENG, we mine the journalist query logs to obtain representative queries.

Resource Selection and Data Fusion

The goal of resource selection is to search only those collections that hold relevant documents given a query request. In PENG, we rank collections by combining two evidence sources (using simple weighted averages): (1) an estimation of collection relevance with respect to a query using CORI [3], and (2) a user specified *trust* score for each resource. Trust scores are an estimate of the quality of information held in each resource. Applying *trust* addresses a key concern of journalists who often use such criteria when researching a story. To illustrate, using *trust* alongside relevance, a digital library of refereed academic articles can be given more importance than a collection of unpublished web articles even though the resource has been given a higher relevance score, thus in turn reflecting the current users needs. After ranking the collections the top k ranked are searched by asking for a decreasing number of documents form each collection based on the position in the ranking. The returned document results are then fused using CORI.

3. EVALUATION

The prototype system will be evaluated operationally on a sample of European journalists to investigate the effectiveness of PENG in supporting real-life tasks over a number of user studies focusing on task-based evaluation. As well as assessing the effectiveness of individual components we hope to be able to investigate the effectiveness of a task-based system, specifically designed for the needs of one user group, against standard retrieval technology.

4. ACKNOWLEDGEMENTS

5. REFERENCES

[1] L. Azzopardi, M. Baillie, and F. Crestani. Adaptive Query-Based Sampling for Distributed IR. In *ACM SIGIR 2006*, Seattle, WA, USA, August 2006.

[2] G. Bordogna, M. Pagani, G. Pasi, L. Antoniolli, and F. Invernizzi. An Incremental Hierarchical Fuzzy Clustering Algorithm Supporting News Filtering. In *IPMU 2006*, Paris, France, 2-7 July 2006.

[3] J. P. Callan. *Advances in Information Retrieval*, chapter Distributed Information Retrieval, pages 127–150. Kluwer Academic Publishers, 2000.