The PENG System: Integrating Push and Pull for Information Access

Mark Baillie[1], Gloria Bordogna[2], Fabio Crestani[3], Monica Landoni[3], and Gabriella Pasi[4]

[1] Dept. Computer Information Sciences, University of Strathclyde, Glasgow, UK
[2] CNR IDPA, Dalmine, Bergamo, Italy
[3] Faculty of Informatics, University of Lugano, Lugano, Switzerland
[4] DISCO, University Milano-Bicocca, Milano, Italy

Abstract. This paper describes the PENG project that integrates personalized push and pull technologies to access relevant information. PENG integrates several key tasks, including personalized filtering, retrieval, and presentation of multimedia news, into a single system. In this paper we provide an overview of PENG, describing our approach to constructing a dedicated retrieval and content management system for a specific user group. We also report critically on the results of a user and task based evaluation.

1 Introduction

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 aim of PENG, an EC funded Specific Targeted Research (STREP) Project[1], was to address the 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 paper we provide an overview of PENG, describing our approach to constructing a dedicated retrieval and content management system for a specific user group. We show how detailed knowledge of a user group and the information tasks they perform has been used to inform the design of retrieval and filtering system components.

[1] More information at `http://www.peng-project.org/`

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 351–360, 2007.

2 The PENG System

The design of the PENG system was motivated by a user study undertaken as part of the project, investigating the work practices of European journalists and editors from different news mediums i.e. TV, Radio and Print. The study investigated how journalists searched, verified, processed and then used information gathered from a wide variety of electronic resources, and in particular how they exploited current systems to support these daily work tasks. Although we cannot report here in full the results of this study, several important requirements came out of this study. Firstly, it was highlighted that journalists require 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 (disparate) news archives and information resource providers, in contrast to what is offered by current systems.

A common theme highlighted during the study was that journalists use a range of criteria when gathering information for a task. These criteria are not static but constantly changing as the journalist and environment changed. For example, across the journalists surveyed, documents were judged by the accuracy of their contents, the reliability and verification of the information source, the accessibility of the information (in terms of speed, cost, etc.), the timeliness of the information, and also the proximity of the information to the journalist (i.e. local news concerning local issues). These findings mirrored previous studies that have highlighted the dynamic and multidimensional nature of relevance, where many factors beyond topicality and aboutness influence how a user assesses information [1,2,3].

One important criteria in particular was the notion of *trust*: the interviewed declared how important it was to identify the original source of the document in order to determine the accuracy of its content. This was considered a vital step when assessing information, and reflected what has previously been cited as one of the key elements of journalism: the verification of a news source [4]. Given the nature of information retrieved from the web, where the original publisher and source of content can be difficult to identify, and where many documents do not go through a strict refereeing or editorial process, the journalist has to be vigilant [5]. To address this issue of source verification, journalists often restrict their search for information to a number of (trusted) resources. Therefore, which resources the journalist searched for information were of particular importance (i.e. news archives, digital libraries or web resources). For example, web search engines were often used because of ease of use and speed, while internal databases were used for checking personal details of sources. Overall, the type(s) of resource a journalist would access at any given time was dependent on the journalist and their individual tasks and needs.

With regards to the requirements analysis, three important phases were identified during the daily workflow of a journalist; information *push*, *pull* and *presentation*. To address these three phases, PENG combines information filtering, searching and presentation within a unified framework which also provides support for personalisation.

PENG has a typical client server architecture. There are five main components on the server:

Filtering module: It provides support for the filtering of incoming stream of news from various resources reporting breaking stories.
Distributed Information Retrieval (DIR) module: It provides support for the search of multiple disparate third party information resources.
Common database manager module: It co-ordinates communication and functionality across the system, also maintaining both databases.
Document database and indexes: It provides a central repository for documents or other information gathered through filtering, as well as important results required for the DIR (such as query history), allowing data to persist over time.
User profile database: It manages a persistent profile for each user stored in this database.
User interface: It provides support for all the functionality of push, pull and presentation of the system, in addition to several others aimed at facilitating the user information gathering, organisation and composition tasks.

This architecture was designed to provide the following advantages (i) a single document representation which can be used consistently by all modules in PENG, (ii) a single user profile representation used by all PENG modules, and (iii) a single method of access to both the information artefacts (e.g. documents) irrespective of whether they are returned by filtering or retrieval. We are now going to describe the functionality of the two main modules of the PENG system, the filtering and DIR components, responsible for the push and the pull of information. Due to space limitations the other modules of the system, though important, will not be presented here.

3 The Information Filtering Module

The filtering component implements the push phase of the PENG system. From the requirements analysis, journalists were found to be "fearfull" of a filtering system which may cause them to miss important information. This was one of the prime motivations for the design of the filtering component, which goes beyond the functionality of what are normally termed filters. In particular, the filtering component was designed to (i) organise the incoming new feeds pushed directly into the PENG system through the use of (fuzzy) clustering, (ii) organise the personalised information of each user by filtering information with respect to each individual users interests, and (iii) rank relevant information by using a variety of criteria alongside relevance, such as timeliness and novelty. With this motivation, the internal architecture of the filtering module was divided into four main sub-modules: 1) Gathering, that receives or actively gathers new material from pushed external information news feeds such as Reuters and ANSA; 2) Clustering, that periodically identify topically related groups of recently arrived documents, thus providing an overview of the current scenario of recent news;

3) Filtering new documents or clusters to individual interests within each user profile by applying personalized multicriteria evaluating content based relevance of news, actuality and novelty of news and trust of the news sources; 4) User monitoring of user actions (carried out in the PENG user interface) and records these actions for later use by the training sub-system. The two most important sub-components are described in the following.

The *personalised filtering system* can filter either individual documents or clusters of documents (i.e. category-based filtering) to user interests [6]. Each PENG user may have a number of individual interests. Each interest is defined by the user, either by example or explicitly by typing a textual description. These user interests provide the second method of organization of the data within PENG, providing a view of the incoming stream of filtered data personalized to each user, and each user interest. Many traditional filtering systems carry out a hard classification of the input stream of documents, classifying each document as it arrives, as either relevant or not relevant to a (user) profile. From the PENG requirements gathering, we posit that this is exactly what journalists do not want. Because of this, the filtering model used in PENG applies a multi-criteria decision for each individual user interest, thus reflecting the multiple criteria used by journalists to assess relevant information. The following measures are computed and used by the filter: *aboutness*: it is a usual measure of the content-based similarity of a new document to the interests of the user; *coverage*: it is a measure of the inclusion of the user interests in the contents of the latest news; *novelty*: it is a measure of the new information offered to the user by an incoming document; *reliability*: it is a measure related to the user trust in the resource from which the news is coming; *timeliness*: it is a measure of the usefulness of a news item to the user-specified time-window. The personalised filtering can be split into two main stages. The first stage computes the relevance judgment of a news to the user interest. This stage first applies a pre-filtering phase based on the consideration of the trust score specified in the users profile. Then the relevance score is generated by combining aboutness and coverage. The value so obtained can then be thresholded to determine the final relevance, i.e. the selection condition of the document to a user interest. If the document is deemed relevant to the interest it will pass to the second stage, i.e. a merging stage which inserts a topically relevant document into the existing result list, generating a number of different scores allowing the result list to the re-ranked by different criteria. The relevance from the previous stage can be combined with either the novelty, trust or timeliness values, to produce different ranking the output of the filter customisable to the users. Such a scheme can be considered as the maintenance of a single ranked list over time, where the job of the filtering system is to place each new document in the ranked list, relative to the other existing documents. In other words, the filtering system must now determine not just whether the document is relevant, but also how is this document relevant relatively to existing results.

A *fuzzy clustering module* operates periodically, currently once every day but potentially every few minutes, to automatically generate a set of clusters which

characterise the incoming stream of news. The clustering is not based on a preset classification, and so may vary from day to day, depending on the content of the daily news. This is intended to provide journalists with an overview of the current "news landscape", easing the identification of relevant breaking news stories, and also providing a means to classifying individual documents within this daily structure. The output of the algorithm is a fuzzy hierarchy of the news reflecting the nature of news, which may deal with multiple topics. The algorithm computes a membership degree score (between [0,1]) for each item (news) to each generated fuzzy cluster allowing the documents to be ranked within a cluster, easily supporting flexible filtering strategies such as the selection of the top ranked news within a cluster of interest.

The generated fuzzy hierarchy represents the topics at different levels of granularity, from the most specific ones corresponding to the clusters of the lowest hierarchical level (the deepest level in the tree structure representing the hierarchy), to the most general ones, corresponding with the clusters of the top level. Since topics may overlap one another, the hierarchy is fuzzy allowing each cluster of a level to belong with distinct degrees to each cluster in the next upper level. To generate such a fuzzy hierarchy, we have defined a fuzzy agglomerative clustering algorithm based on the recursive application of the Fuzzy C-means algorithm (FCM). The algorithm works bottom up in building the levels of the fuzzy hierarchy. Once the centroids of the clusters in a level of the hierarchy are generated, the FCM is re-applied to group the newly identified centroids into new fuzzy clusters of the next upper level. In this way, each level contains fuzzy clusters that reflect topics homogeneous with respect to their specificity (or granularity), so that, in going up the hierarchy, more general topics are identified [7]. The algorithm can also operate an updating of a generated hierarchy of clusters with new news arriving on the stream. This incremental modality can eventually add new clusters when the content of new news is too different from that represented in current clusters of the hierarchy. More specific characteristics of the clustering algorithm are the automatic estimation of the number of clusters to generate, the efficient management of sparse vectors of documents features, and the use of a cosine similarity [7]. Finally, when the clusters are identified their labelling takes place with the aim of summarizing the main contents of the most representative news of the clusters. The summarization criteria identify the index terms with highest share among the top ranked news of the cluster and with the highest discrimination power among all the clusters. The balance of these two criteria makes it possible to generate unique labels for the overlapping fuzzy clusters.

4 The Distributed Information Retrieval Module

The pull phase allows the journalist to find background context on breaking news stories, deepen their knowledge of the story and/or assist during the compilation of news reports or articles. In PENG we facilitate search across a wide range of remote and local information resources. To enable search across a number of

distributed resources within an integrated framework, we had to address four main research problems: automatic query generation and refinement, resource description acquisition, resource selection and data fusion [8].

The input to the retrieval component can be either from standard ad-hoc querying or pushed news documents explicitly selected to "deepen" the topic by the journalist. In the former case, the journalist enters keywords into the system when searching for information. In the latter case, automatic queries are formulated from pushed documents selected from the filtering. This process minimises the workload for the journalist, by extracting query terms based on term importance, and also provides a longer query than is typically submitted to a search interface potentially providing better retrieval accuracy. If available, queries are also expanded using terms from the journalist's user interest determined by the filtering module during the push phase. The refined query is then used to search distributed collections available to PENG.

We investigated a number of solutions for *automatically generating queries* from pushed news feeds. We investigated the use of representative and discriminative terms for query expansion, which has been found to be an effective technique for query expansion in centralised retrieval. The assumption for using query expansion with representative and/or discriminative terms is: user's with little topic familiarity in the topic, representative terms for the topic will be able to locate documents that are very general e.g. overview documents. In comparison, discriminative terms can be used to find detailed documents about a topic, for those user's with previous knowledge of that subject area. To extract either discriminative and representative terms, a topic language model is formed from the pushed news documents. The Kullback-Leibler Divergence measure is then applied to determine a term's contribution to the topic model [9]. Terms in the topic model are then ranked according to how representative or discriminative they are, and then used as a input query to the DIR module. For those users with low familiarity of the topic, the top ranked representative terms are used, while a user with high topic familiarity the set of discriminative terms used. Across an exploratory analysis of using both sets of queries for various user contexts, within the 2005 HARD track of TREC, it was discovered that using discriminative queries provided improved retrieval accuracy when compared to other query expansion techniques for users of varying topical knowledge. As a result, we have adopted this approach for generating automatic queries from all pushed news documents.

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*, 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 [8]. Our approach is based on measuring the Predictive 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 [10]. 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 actual information resources. PL requires that only a set of queries are available for evaluating each resource description. In PENG, we mine the journalists query logs to obtain queries that are representative to the typical information needs of the journalists. Past queries are stored in the user profiles database, alongside a record of the meta-data of the current documents searched and queried by each user (stored in the document database).

Finally, 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 [8], 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 the CORI algorithm.

5 Evaluation

The evaluation of the prototype system followed a task-based and user-oriented methodology in the context of a formative design evaluation framework. The evaluation involved 9 professional journalists, and 13 postgraduate students of journalism. Professional journalists, considered in this context expert users, were asked to complete forms describing typical information filtering and information search tasks carried out during their daily work activities. These forms gathered, over a period of 3 months, contained information about the nature of each task, the way it was carried out using any system available to the journalists, and the information found to be relevant for the task. During the same period material from the newswires and information repositories the journalists typically accessed was logged and copied into a separate storage. The documents reflected the nature of the tasks defined, resulting in a mixture of various multimedia and also multilingual documents (English, Italian, German and French) to reflect the working environment for this sample of users. Information extracted from the forms was used to design specific information filtering and information

search tasks that both students and professional journalists carried out using the prototype.

To simulate appropriately the information filtering task, newswires and other pushed information was delivered in a time delayed fashion. The information available through the search was also the information that was originally available to the journalist at time $t - k$, as the users were then required to carry out at time t with the prototype the same filtering and search tasks performed by the experts at time $t - k$ with different systems. Users were then asked to judge the performance of PENG in relation to a number of evaluation qualitative dimensions pertaining to ease of use, learnability, satisfaction, likeability, and general attitude to the system and compare that with the systems they used during their everyday work. Quantitative data was also collected related to rate of task completion, time to carry out tasks, error and recovery rates. Additional information related to the performance in these tasks was also collected using the "think aloud" technique, direct observation and interviews[2]. Each user involved in the evaluation experiments was asked to carry out one filtering and one search task per session for a total of 3 sessions spread out over a period of two months.

The results in general indicated that the *students* found search tasks intuitive and were comfortable with using the PENG prototype. When considering the prototype against the system they would have normally used, Google, the overall usability of the prototype was comparable and helped them retrieve relevant documents with 75 % finding the system easy to use and browse. Approximately 50% of the user group believed that the prototype helped in completing their tasks faster; general consensus being that the ability to search simultaneously a lot of various news sources and resource was an advantage over a generic search engine. In particular, a cited advantage of the prototype was the importance given to news agencies and resources used for the region of Europe the students lived (i.e. Switzerland). This was potentially an indication of the trust model of the resource selection algorithm placing more weight on the geographical proximity of some resources in comparison to others. Also, the ability to search local repositories (both personal and shared in the PENG system) was considered a useful feature. One limitation of the prototype during the pull phase, however, was that the accuracy of the retrieval results varied across tasks. In particular the search for named entities such as people or specific places was often variable. A number of students asked specifically for Boolean operators to be available while formulating a query. Possibly as some queries were not returning documents with the people or places expected result, the users did not feel in control, hence the request for more advanced search features. Users were confused by the ability to search multiple language resources. In terms of the push phase, the students found overall the filtering tasks complex or even too complex for them to cope with the extra complications of not really understanding initially the meaning and implications of filtering. While interacting with the PENG retrieval module

[2] The interviewers did not belong to the design and development team in order to avoid any bias.

students could relate to their previous experience with search engines, they found it hard to deal with the filtering component as they were lacking of previous experience in this area. This was probably one of the underlying factors that influenced the less positive scores to the usability of PENG filtering component.

In comparison to the student responses, the *professional journalists* on the whole felt PENG was an obstacle for completing tasks. Professional journalists expect a high quality of service, in particular radio news journalists who are very dependent on the speed that they can find relevant information. As a result of these high expectations, the speed of the search was a negative comment even for a prototype version of the system. While the professionals agreed with the feature of searching agencies, repositories, personal archives being an advantage in the prototype, there was also emphasis that this should be extended to include local newspapers' and local authorities' archives. Even with a prototype system, the variable accuracy and speed made the journalists very sceptical about using the system in future to perform new tasks. Also, a general observation was the lack of obvious Boolean search operators traditionally used for search in library style systems but nowadays substituted by natural language interfaces. This could be justified as above for students with the users' need to feel more in control and possibly relate to users' previous extensive experience with more traditional interfaces for searching. A total of 71% negative responses of users not feeling satisfied about task completion indicates how journalists did consider filtering as an hindrance more than a useful functionality. As filtering is in nature passive and transparent to users it was difficult for them to understand how the system could be used to accomplish their tasks. Indeed the opinions expressed by journalist were inevitably biased by their everyday experience using Open Media, a professional tool described by them as extremely effective and flexible. Overall filtering proved not popular nor used as journalists had no frame of reference and lack of trust that it was not hiding news.

In general the opinions expressed by journalists and students were in line. If anything the journalists were more confident in their choice of scores, especially in negative terms, than the students. They were also more experienced in using a variety of sources and tools for finding relevant information, in particular they tended to compare PENG to the professional tool in use. Journalists were quite critical about PENG usability from the very beginning when performing search tasks and complained they could have performed better without PENG.

It is without doubt that being compared versus a professionally designed and engineered tool such as Open Media did not help PENG, a tool still in its prototype version. Even more so as the final experiments had to be rushed in order to take place on time for the completion of the project. The lack of time unfortunately affected particularly the filtering as it is the most time consuming of the modules and resulted in journalists not being able to perform properly some of the proposed tasks, as commented by the group locally in charge of running the evaluation experiments. In particular the filtering suffered from the confusion and lack of understanding in both user groups of what filtering really was for, that resulted in confused/mistaken expectations. This had emerged and

was reported earlier on, while collecting task descriptions, and these findings confirmed that both user groups, students and journalists, felt equally confused when defining and later on performing filtering tasks. While search engines have educated users on retrieval tasks, filtering systems are not yet as popular and effective in educating users.

6 Conclusions

We have highlighted one solution to the information access and seeking problems that journalists currently face. PENG is an initial attempt at modelling and integrating the push, pull and presentation phases of a journalists workflow. In this study real users and in particular professional journalists were involved and this allowed us to have very valuable feedback as opposed to the tradition student lab based approach used in literature, where students are conveniently involved in evaluation of retrieval and filtering systems even if they are not necessarily representative in terms of genuine needs, skills and motivations of the final users. Indeed involving busy professionals added a level of complexity both in practical and conceptual terms.

References

1. Barry, C.L.: User-defined relevance criteria: an exploratory study. J. Am. Soc. Inf. Sci. (JASIS) 45, 149–159 (1994)
2. Schamber, L., Eisenberg, M., Nilan, M.S.: A re-examination of relevance: toward a dynamic, situational definition. Inf. Process. Manage. 26, 755–776 (1990)
3. Ruthven, I.: Integrating approaches to relevance. New Directions in Cognitive Information Retrieval 19, 61–80 (2005)
4. Kovach, B., Rosenstiel, T.: The Elements of Journalism. Random House (2001)
5. Pharo, N., Jarvelin, K.: Irrational searchers and ir-rational researchers. Journal of the American Society for Information Science and Technology 57, 222–232 (2005)
6. Bordogna, G., Pagani, M., Pasi, G., Villa, R.: A Flexible News Filtering Model Exploiting a Hierarchical Fuzzy Categorization. In: Larsen, H.L., Pasi, G., Ortiz-Arroyo, D., Andreasen, T., Christiansen, H. (eds.) FQAS 2006. LNCS (LNAI), vol. 4027, Springer, Heidelberg (2006)
7. Bordogna, G., Pagani, M., Pasi, G., Invernizzi, F., Antoniolli, L.: An Incremental Hierarchical Fuzzy Clustering Algorithm Supporting News Filtering. In: IPMU. Information Processing and Management of Uncertainty in Knowledge-Based Systems (2006)
8. Callan, J.P.: Distributed information retrieval. In: Advances in information retrieval, pp. 127–150. Kluwer Academic Publishers, Dordrecht (2000)
9. Baillie, M., Azzopardi, L., Crestani, F.: An evaluation of resource description quality measures. In: ACM SAC 2006, ACM, New York (2006)
10. Baillie, M., Azzopardi, L., Crestani, F.: Adaptive query-based sampling of distributed collections. In: Crestani, F., Ferragina, P., Sanderson, M. (eds.) SPIRE 2006. LNCS, vol. 4209, Springer, Heidelberg (2006)