

The PENG System: practice and experience

Gabriella Pasi
Università degli Studi di
Milano Bicocca, via Bicocca
degli Arcimboldi 8,
Milano Italy
pasi@disco.unimib.it

Gloria Bordogna
CNR-IDPA
via Pasuio 5, 24044 Dalmine
(BG) Italy
gloria.bordogna@idpa.cnr.it

Robert Villa
CNR-ITC
Via Bassini 15, 20133
Milano Italy
peng2@itc.cnr.it

Abstract

The PENG system is intended to provide an integrated and personalized environment for news professionals, providing functionalities for filtering, distributed retrieval, and a flexible interface environment for the display and manipulation of news materials. In this paper we review the progress and results of the PENG system to date, and describe in detail the document filtering part of the system, which is designed to gather and filter documents to user profiles. The current architecture will be described, along with some of the main issues which have so far been found in its development.

1. Introduction

PENG (“Personalised nEws coNtent programminG”) is a European Project, funded under the sixth framework programme. A major contribution of the project is the PENG system, which intends to provide an integrated, personalized environment in which news professionals can work, providing services for the filtering of incoming news streams, searching of distributed databases, and providing an interface which enables a journalist to easily integrate content from all these sources. An overview of PENG is provided in [1] and [2].

As of the writing of this paper, the PENG project is nearing an end, with the software system itself in the final stage of integration. In this paper we introduce the PENG system outlining the main aspects of the project, and in particular the filtering component of the system. Some of the main discoveries found during the project’s life will then be outlined, again from the point of view of the filtering component, before presenting a short conclusion on the lessons learned from the PENG project so far.

The PENG system has been designed for news professionals, such as journalists working in disparate areas as television news, radio news, and magazine publishing. The types of content these different groups of users require are varied, and include text, video, images, or sound files. For this reason, the PENG system was from the start designed with a strong multimedia element, with the intention to allow the filtering and retrieval of various types of data, in addition to conventional text.

2. The Requirements of PENG

The requirements gathering for PENG was carried out based on a group of journalists from Radiotelevisione Svizzera (RTSI) and other journalism experts from the Università Della Svizzera Italiana (USI). This requirements gathering produced some surprises concerning the needs of the PENG system, the first of which was the apparent fearfulness many journalists held of a filtering system blocking access to potentially useful information. This is perhaps not so surprising: journalists are always on the lookout for “new” news stories, although this particular finding had a profound impact on the design of the PENG filtering system (section 5).

Journalists were found to select news based on many different criteria, which may be personal to the journalist, or may be related to the work environment, and include factors such as the audience for which the work is intended. An important factor was that the reliability of information is important, which was reflected in the journalists suggestions for automatic aids in finding the “source” of a news story.

Lastly, journalists are not computer experts, and so a system such as PENG must be easy to use. However, which journalists are not computer experts, they are experts in their own news domain, and require full control over how any “intelligent” systems operate, such as a filtering system.

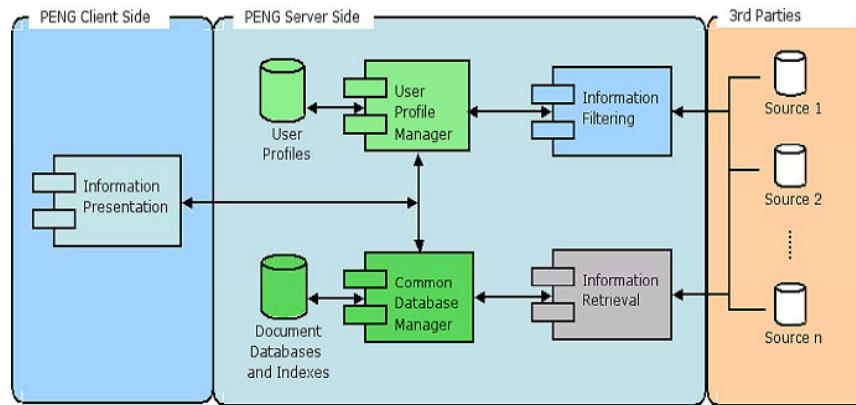


Figure 1: The overall PENG architecture

3. The PENG Software System

The overall PENG software architecture is given in Figure 1, above, and shows PENG's three main software components:

- Information filtering (IF), which gathers new documents from external information sources and then filters this information to relevant users
- Distributed Information Retrieval (DIR), which allows the user to search one or more external information sources
- Information Presentation (IP), which provides a flexible interface to the whole system, IF and DIR

PENG is intended to provide a long term, personalized environment, and therefore an important part of the PENG system is the persistent user profile which is stored for each user. The user profiles are stored in the "user profile manager", which is responsible for maintaining a consistent state for each profile for each of the three components. Alongside this central profile database, we also have a "common database manager", which provides a central repository for documents or other information artifacts gathered by the IF or retrieved by the DIR, allowing search or filtering results to persist over time.

Both the IF and DIR modules access the third party information sources, shown on the right of Figure 1.

It should be noted that all communication between IF, IR and IP is via these two intermediate databases. This includes elements such as search and filtering results. Ranked lists, for example, are stored as part of each user's profile, with the documents themselves stored in the document database module. The databases, plus the IF and DIR, reside on a main PENG server, the interface resides on a separate client machine. More information on this architecture is provided in [2].

This architecture was designed to provide the following advantages :

- a single document representation which can be used consistently by all modules in PENG
- a single user profile representation used by all PENG modules
- a single method of access to both the information artifacts (e.g. documents) irrespective of whether they are returned by the filtering or retrieval

The specification and maintenance of a single user profile was an important consideration in the design – the possibility of multiple modules containing multiple conflicting representations of a user profile was regarded as unacceptable. The maintenance of the consistency of these profiles is the responsibility of the user profile manager.

4. The PENG Interface Component

The PENG interface brings together the other modules to provide a single interface with which the user can access both push (filtering) and pull (retrieval) information. A screenshot of the interface is given in Figure 2, showing a list of filtering results. On the left, the user is able to select different user interests, such as "sport" and "politics".

The interface in Figure 2 shows the following:

- Filtering results, which displays and allows the user to interact with the filtering results
- Search results, allowing the user to search distributed databases
- A "Current work" section, allowing the short-term storage of search and filtering results
- A Personal repository, allowing the long-term storage of documents
- User Profile section, which allows the user to view and edit their user profile

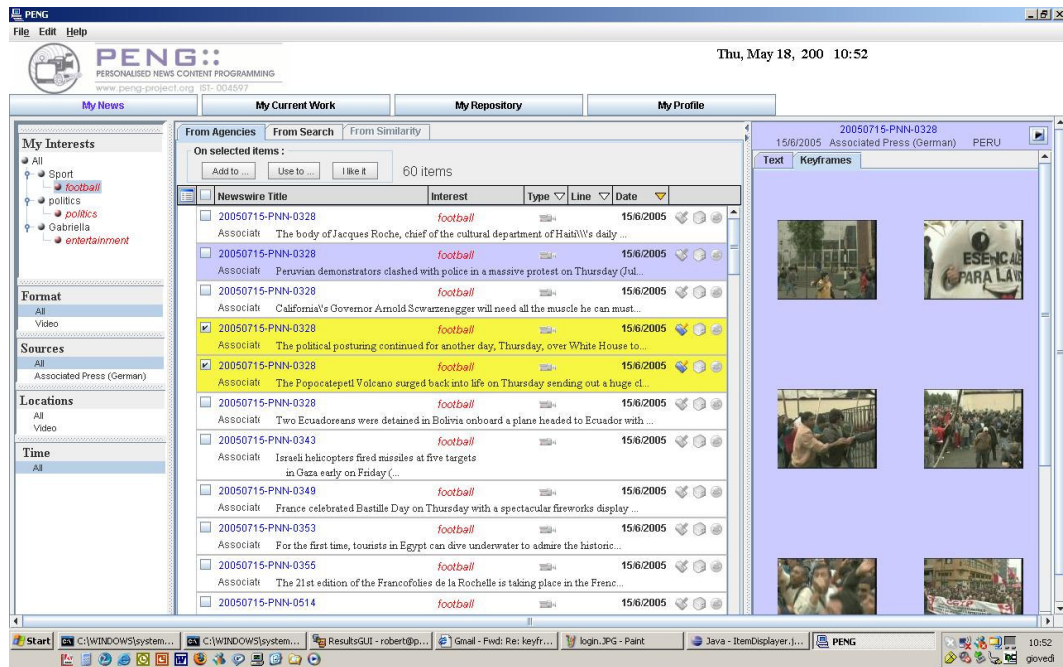


Figure 2: The PENG interface

An important aspect of the interface is its ability to support the journalists task in finding and using documents, and being able to quickly and easily switch between these tasks. To enable this, there is a “personal repository” where a user can store and organize documents which are of long term interest. The “current work” section allows the storage of search and filtering results, and is designed to aid the journalist in their short-term document creation needs.

In the next section we will concentrate on the PENG filtering module, which has been the focus of the work at CNR. The distributed information retrieval functionality is described in [7].

5. The PENG Filtering Module

As described in Section 2, journalists were found to be “fearful” of a filtering system which may cause them to miss important information. This was one of the motivations for the design of the information filtering (IF) module, which goes beyond the functionality of what are normally termed “filters” [5, 6, 12].

In particular, the filtering is intended to:

- Organize the latest information gathered by the PENG system, through the use of a fuzzy clustering algorithm
- Organize the personalized information of each user by filtering to individual user interests
- Rank relevant information by various criteria, such as by time and novelty

The general architecture of the filtering module has three main sub-modules:

- **Gathering:** receives or actively gathers new material from external information news sources such as Reuters and ANSA
- **Classification and clustering:** periodically, the most recently arrived documents will be clustered, to identify topically related groups of documents. The fuzzy clustering algorithm is described in [3]
- **Filtering:** filters new documents or clusters to individual interests within each user profile. An overview of the filtering system is described in [4]

The fuzzy clustering operates periodically, currently once every day, to automatically generate a set of clusters which characterize the days news stream. The clustering is not based on a pre-set 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 “news landscape”, allowing one to identify the main news stories and providing a way of classifying individual documents within a daily structure.

The IF will filter either individual documents or clusters of documents (“category-based filtering”, [3]) to “user interests”. Each PENG user is assumed to have a number of individual interests, which may be general categories such as “sport” or “politics”, or more specific and personalized entities such as “the politics of Tony Blair”. Each interest is defined by the user,

either by example or explicitly. 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's interest.

Many traditional filtering systems, such as SPAM filters [5] or those from TREC [6] 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 profile. From the PENG requirements gathering (Section 2), we know this is exactly what journalists *don't* want. Because of this, the filtering scheme used in PENG assumes that the filtering results are a set of documents judged relevant to an interest, which may be re-ranked based on various criteria. Thus the filtering can be split into two main stages:

- A relevance judgment stage, which computes the "topical" relevance of the document to the user interest. If the document is deemed "relevant" to the interest it will pass to the second stage
- A "merging" stage which will insert a topically relevant document into the existing result list, generating a number of different scores allowing the result list to be re-ranked by different criteria

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. The question the filtering system must ask is not just "is this document relevant?" but also "how is this document relevant compared to the existing results?" This change simultaneously makes the filtering easier (it no longer needs to make a hard classification as soon as a document arrives) and also harder (new documents must be placed into an existing ranked list).

The filtering process itself, and the closely associated ranking of the results for each individual user interest, is carried out based on a number of different criteria described in [4].

6. PENG Architecture: Some Experiences

6.1 The PENG Architecture Implementation

The PENG system was created by four teams spread around 4 different countries in Europe, and used a variety of technologies in its realization. The system itself is made up of two main parts: a server component on which the main common databases, IF and DIR are run, and a client machine on which the PENG interface is run (this follows the client/server split shown in figure 2).

One technical decision which has caused problems was the use of a stand alone java interface rather than a web-based interface: in the current system Java's remote method invocation is used to communicate between the server and client. This, however, is complex, and this extra complexity has impacted negatively on the integration of the overall system.

A second, less technical problem, concerns the common document database, into which IF and DIR results are placed, and from where the interface can access either IF or DIR results. For the IF, this works well: filtered documents are transient entities, which will be broadcast, gathered by the IF system, and then stored for future reference. The same cannot be said, however, for documents retrieved by the DIR module: these documents are likely to already be persistently available. Storing them locally is likely to duplicate the already existing resources, and may lead to publication rights problems for a production system (problems which PENG, currently, does not deal).

6.2. Mapping and Defining Document Schemas

Perhaps one of the most difficult problems PENG has experienced is dealing flexibly with metadata. The system must handle various different kinds of documents (newswire documents, web documents, visual documents, etc) which are marked up using different metadata schemas, and different file formats.

This results in a veritable tower of Babel of metadata. In the current PENG system, these external formats are mapped onto a single database schema. This has resulted in two unforeseen sub-problems:

- The definition of the internal PENG schema has become a problem
- The problem of mapping external schemas into the internal form has become an issue

The latter problem, that of mapping external schemas, is known to be non-trivial [13]. Its difficulty only increases if the target (PENG) schema is only partially or incompletely defined or known, which has often been the case in PENG, a difficulty only increased when new types of documents, with slightly varying schemas, have had to be integrated into the system during the development process.

Recourse to standards such as Dublin core, for common metadata elements, has also been found to be lacking: such schemes are often very broadly defined, so that they may be used in many different situations. For example, an identifier in Dublin core may be a URL, Digital Object Identifier or even an International Standard Book Number (ISBN). This is contrary to the needs of modules such as the filtering and DIR, where the metadata is actually used, and therefore where it

gains its meaning. In these modules, we would like to have consistent, reliably available metadata, to ensure the content of the metadata is *used* correctly.

In the current PENG filtering system these problems are mitigated a little by using more reliably available PENG-generated data, such as the time at which a document is gathered by the filtering module, rather than relying on more unpredictable document date and time metadata.

In practice we have found this common schema development very difficult. A greater emphasis on the creation of a standard “core” metadata set, which should be guaranteed for all documents, followed by a larger set of “unguaranteed”, not always available subsidiary metadata, may have made the problem more manageable. But for this to work, individual modules must also be designed to operate when not all metadata is available, which will have a knock on effect on the techniques which can be used in retrieval or filtering.

5.3. Integration of IR and Database Systems

Another problem, experienced first hand by the design of the filtering module, is the large gap between the databases conventionally used (i.e. relational databases, in the case of PENG, Postgres [14]), and the operation of Information Retrieval (IR) systems.

There have been various attempts to integrate IR systems with databases, such as [9], [10], or [11]. Some of these, such as [10], integrate IR directly into a relational database system. In the PENG system, metadata is stored in Postgres, while an IR system (Lemur, www.lemur-project.org) generates an inverted index which is used to search the document database. While this has been found to work acceptably well, this crisp split is not as flexible as may be liked:

- It's difficult to limit a free text search to only a subset of the common document database (e.g., to search only documents previously seen by a single user, rather than all users).
- Text parsing functionality is duplicated, existing in the IF, DIR and document database

An IR system is not simply a storage engine: it provides textual processing routines which allow a document to be converted into term frequency vectors, for example, and maintains the data structures required for this transformation (such as a Lexicon). Within PENG, the integration of this parsing functionality between IF, DIR and the common databases may provide a cleaner design, although this has to be offset by the greater dependence between these components which would result.

6. Conclusions

The PENG project is, at the time of writing, nearing its final stages of integration, before final user testing. The project itself has generated a number of novel pieces of software and research, including:

- A new fuzzy clustering algorithm [3]
- Multi-criteria text filtering [4]
- Distributed Information Retrieval system [7]

The overall system is currently running as a prototype, although further development on the overall integration is continuing.

The development of any large software system such as PENG is likely to include many twists and turns, and this project was no different. In this paper some of the main problems experienced in the development of the PENG system have been outlined.

References

- [1] G Pasi, “An Overview of the PENG (Personalized News Content Programming Project)”, *EWIMT*, 2004
- [2] G. Pasi and R. Villa, “Personalized News Content Programming (PENG): A System Architecture”, *16th International Workshop on Database and Expert Systems Applications (DEXA'05)*, 2005, pp. 1008-1012
- [3] G. Bordogna, M. Pagani, G. Pasi, F. Invernizzi, and L. Antonioli, “An Incremental Hierarchical Fuzzy Clustering Algorithm Supporting News Filtering”, *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, 2nd - 7th July, 2006
- [4] G. Bordogna, M. Pagani, G. Pasi, R. Villa “A Flexible News Filtering Model Exploiting a Hierarchical Fuzzy Categorization”, *7th International Conference on Flexible Query Answering Systems (FQAS)*, 7th-10th June 2006
- [5] G. Cormack, T. Lynam, “TREC 2005 Spam Track Overview”, *Proceedings of TREC 2005*, ACM, 2005
- [6] S. Robertson, I. Soboroff, “The TREC 2002 Filtering Track Report”, *Proceedings of TREC 2002*, ACM, 2002
- [7] M. Baillie, F. Crestani and Monica Landoni, “Integrated search of distributed news archives: The PENG system”, *ACM SIGIR 2006*, Seattle, USA, August 2006
- [8] G. Amato and U. Straccia, “User Profile Modeling and Applications to Digital Libraries”, *ECDL*, 1999, pp.184 -197
- [9] A.P. de Vries and A.N. Wilschut, “On the integration of IR and databases”, *Database issues in multimedia; DS-8*, January 1999, pp. 16-31
- [10] S. DeFazio, A. Daoud, L. A. Smith, J. Srinivasan, Bruce Croft, and Jamie Callan, “Integrating IR and RDBMS Using Cooperative Indexing”, *ACM SIGIR 1995*, 1995
- [11] W. B. Croft, L. A. Smith, H. Turtle, “A loosely-coupled integration of a text retrieval system and an object-oriented database system”, *ACM SIGIR 1992*, Copenhagen, 1992
- [12] D. Oard, G. Marchionini, “A Conceptual Framework for Text Filtering”, *CS-TR-3643, Univ. of Maryland*, May 1996
- [13] P. Shvaiko, J. Euzenat, “A survey of schema-based matching approaches”, *Journal on Data Semantics*, IV, 2005
- [14] PostgreSQL community site, <http://www.postgresql.org/>, accessed on 29th May 2005