

A Framework for Filtering and Packaging Hypermedia Documents

Lucimar C. Martins, Tatiana A. S. Coelho, Simone D. J. Barbosa,
Marco A. Casanova and Carlos J. P. de Lucena

Informatics Department, Pontifical Catholic University of Rio de Janeiro
Marquês de São Vicente, 225 – 22453-900 – Rio de Janeiro – Brazil
{lucimar, tati, sim, casanova, lucena}@inf.puc-rio.br

Abstract. Despite great effort in attempting to develop systems that personalize both content and presentation, there are still some important challenges related to information filtering, packaging and formatting that adapt to user's goals, interests and presentation preferences. This paper addresses these issues by proposing a three-level framework that achieves a high degree of separation of concerns. The framework dissociates the packaging process from the filtering and formatting processes, and thus facilitates the implementation, user testing and fine-tuning of the system representations and algorithms.

1 Introduction

The feeling of being “lost in hyperspace” is familiar to most Web users. The excessive amount of information confuses us to the extent of not knowing where we are and forgetting what we were looking for when we started browsing.

Many applications related to e-learning, e-commerce and information retrieval have been designed so as to adapt both content and navigation access to the users' supposed knowledge [7,8], to their preferences and goals [1,2,3,9,11,13], to their tasks and receptivity [4,5,15]. Many of these references illustrate the increase in efficiency that certain adaptation techniques bring about [12,17].

Despite great effort in attempting to develop systems that personalize both content and presentation, there are still some important challenges, some of which are related to:

- acquisition and representation of relevant document and user information;
- information filtering that takes into consideration such information;
- information packaging and formatting that adapt to user's presentation preferences, his current browsing device and network conditions.

In this context, the primary contribution of this paper is to propose a three-level framework that achieves a high degree of separation of concerns, by dissociating the *packaging* process from the *filtering* and *formatting* processes. We also illustrate a

possible use of our framework through an instantiation called *MyNews*, which is a personalized electronic newspaper.

A high degree of flexibility, provided by the framework, is necessary because the success of adaptive systems relies heavily on the choice of representations and algorithms adequate to the underlying domain and application. The evaluation of this success must be done empirically, by means of user testing, which may require further fine-tuning. The final goal is to achieve a high level of user acceptance and satisfaction, with minimal redesigning and code rewriting.

This paper is organized as follows. Section 2 presents an overview of the proposed framework. Section 3 describes the user, document and packaging models. Section 4 discusses the process of information filtering. Section 5 describes information packaging. Finally, Section 6 presents the conclusions and suggests directions for future research.

2 Framework

The idea of separation of concerns [18] provides a major motivation for the definition of a framework for personalized filtering and packaging of hypermedia documents. Our goal is to be able to analyze issues related to document and user modeling, information filtering and information packaging, as independently as possible.

This approach makes it easier to build a variety of applications in diverse domains, by deriving multiple instantiations of the framework using different algorithms and configurations. This characteristic is essential to adaptive applications, since user satisfaction can only be empirically verified, and a great amount of tweaking and fine-tuning may thus be necessary, according to the successes or limitations of each configuration.

In order to illustrate the proposed framework and personalization processes, we will use an instantiation of the framework, called *MyNews*, for the domain of electronic newspapers. *MyNews* creates a personalized newspaper that selects adequate content and presentation structures to be delivered to its users.

In the following subsections, we will present the architecture of the proposed framework and a brief description of its hotspots.

2.1 Architecture

The overall goal of the architecture presented in this section is to outline the adaptation process the available content goes through, from being requested by the user to finally being presented to him. This process takes into account document, user, packaging and formatting models.

Figure 1 presents a functional architecture of a document personalization system. A user starts by requesting a personalized view of the set of documents contained in the repository. This request is processed by the *Interface Subsystem* and dispatched to the *Filtering Subsystem*, which is responsible for creating an ordered set of

documents, based upon the user and document models. These models will be discussed in Sections 3.1 and 3.2, respectively.

The *Packaging Subsystem* is responsible for regrouping, reordering and restructuring the filtered set of documents, based on the packaging model described in Section 3.3.

The *Formatting Subsystem* creates the final layout and visual design of each document, according to the formatting model. This task relies heavily on the expertise of graphics designers and will not be further discussed in this paper.

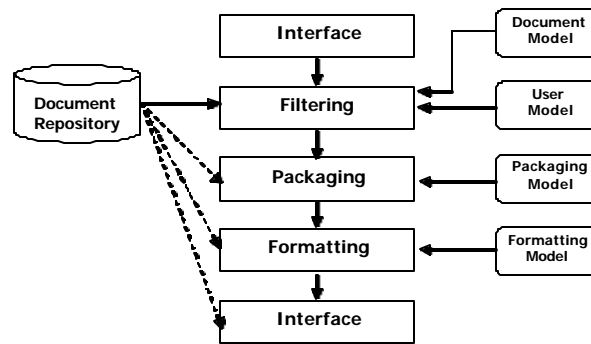


Fig. 1. Functional architecture.

2.2 Hotspots

Our framework contains several hotspots to achieve a high degree of separation of concerns. Whereas the basic architecture is invariant, the models and algorithms may differ from one domain or application to another.

The framework hotspots are as follows:

- **document model:** defines the documents' structure, including metadata that help classify the documents in various ways.
- **user model:** allows the definition of constraints related to each metadata, i.e., acts as a definition of views upon the document model.
- **filtering algorithm:** constructs a query by combining the document and user models, and returns the ordered set of documents computed as most relevant to the user.
- **packaging model:** defines which portions of the document's structure should be presented, and in which arrangements or presentation structures.
- **packaging algorithm:** uses the packaging model to regroup and rearrange the filtered set of documents, as well as to restructure the documents.
- **formatting model:** defines how each document should actually look like when presented to the user, including platform dependencies.
- **formatting algorithm:** uses the formatting model to create the final presentation layout of the documents.

3 Models

3.1 Document Model

The document model is any hypermedia model that defines the documents' structure, including metadata attributes that reflect characteristics of the documents. The definition of each metadata domain must include comparison and metrics operators that permit defining precise filtering algorithms.

In the context of *MyNews*, documents are structured into: title, subtitle, authors, summary, image and whole text. The metadata domains are:

- the *semantic metadata domain* is represented by a labeled directed graph, in which a node label represents a theme in the domain, and an arc label represents the strength of the connection between two nodes. The distance between two nodes A and B is the weighed sum of the arcs in the shortest path between A and B. Figure 2 presents a sample semantic graph;
- the *importance metadata domain* is the integer interval [1,10] and indicates the importance of a document, where 1 means the highest importance. This metadata determines the relation between a document and its corresponding theme;
- the *temporal metadata domain* is a set of timestamps and allows the computation of a document's obsolescence.

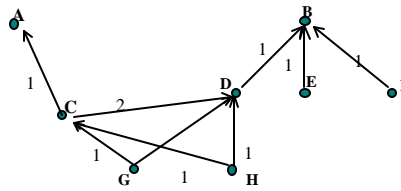


Fig. 2. Part of the semantic metadata domain with the themes of an electronic newspaper.

3.2 User Model

The user model must contain information about the preferences and goals of a user when using the system. It depends on the document model. The user model may be provided directly by the user through a questionnaire, or implicitly inferred by the system, which then needs to monitor his interaction behavior.

In the *MyNews* example, the user model consists of a triple (S,I,T) , where:

- S is a list of themes that interest the user, taken from the themes listed in the semantic metadata domain;

- I indicates the minimum importance of the documents to be retrieved.
- T is a set of constraints on the obsolescence of the documents.

Suppose that the user is interested in theme A and that he wishes to view all news pieces (any importance value between 1 and 10) that were inserted in the last 3 days. The user model could then be represented by the triple $(\{A\}, 10, \{t < 3\})$.

3.3 Packaging Model

The packaging model specifies:

- the groups into which the filtered set of documents will be partitioned;
- the order of the groups and the order of the documents within each group;
- the document views, which define the components that will be retained and passed to the formatting process, in addition to references or links to other document views.

A *document view* specifies a different structure for a document, that may reorder or even omit its components.

In *MyNews*, the document views are represented as in Figure 3.

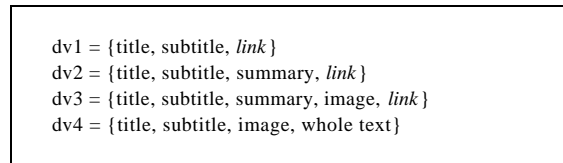


Fig.3. Document views used in *MyNews*.

In the example, we define a very simplistic packaging model as follows. The model contains the following groups:

- the *front page* contains news pieces with rank less than 6 (see Section 4 for the definition of rank);
- for each theme T in the user model, there is group, called *thematic group T*, containing all filtered news pieces whose theme is T .

The front page is always the first group and the thematic groups are listed by alphabetical order of their theme. The news pieces within each group are listed by rank order, as computed by the filtering algorithm.

For the first three documents in the front page, the packaging model retains the components defined in view dv3 of Figure 3; for all other documents in the front page, it uses view dv1. The thematic groups use view dv4 for their documents. In addition, for the documents in the front page, the hyperlinks in views dv1 and dv3 point to the corresponding news pieces in the thematic groups.

An alternative packaging model for *MyNews* would, for example, use only view dv1 for all documents and hyperlink each view to the document's full version in the document repository. A second and more sophisticated model would take into

account limitations of the platform and indicate to use views dv1 and dv2, instead of views dv3 and dv4, respectively, if the platform has a small presentation area.

4 Filtering

The filtering algorithm creates an ordered subset of the set of available documents, based on a *ranking function* that maps each document into its *rank*. according to their metadata values and the goals and interests represented in the user model. Intuitively, the filtering algorithm correlates the documents' metadata with the user model, creating an ordered subset of documents that supposedly interest the user the most.

The definition of the ranking function is an interesting issue. In order to verify that a certain document is indeed relevant to the user, a great deal of user testing must be done, and the ranking function will probably need to be fine-tuned for best results.

Returning to our running example, recall from Section 3 that the user model is the triple $(\{A\}, 10, \{t < 3\})$. Consider a ranking function that combines the semantic and importance metadata and is defined as $P[U](x) = \sqrt{(D_s[U](x))^2 + (D_i[U](x))^2}$, where $D_s[U](x)$ corresponds to the shortest distance between the theme labeling document x and the themes of interest to the user, and the value $D_i[U](x)$ represents the importance of document x .

Table 1 shows a possible result of the filtering process, where the rank column contains the document rank, computed by the above function, and all other columns correspond to the document metadata and document components.

Table 1. A sample set of documents resulting from the filtering process.

Document	rank	theme	importance	obsolescence	title
Doc2	2.24	C	1	3-Dec-2001	Wall Street...
Doc1	3.16	A	3	5-Dec-2001	Dolly...
Doc4	3.16	A	3	3-Dec-2001	Crisis in...
Doc53	5.10	A	5	3-Dec-2001	Research...
Doc18	7.21	D	6	4-Dec-2001	Christmas sales...
Doc7	10.05	A	10	1-Dec-2001	New clones...

5 Packaging

Packaging comprises three major processes, grouping, ordering and document restructuring, and is driven by the packaging model.

The first step partitions the set of documents returned by the filtering process into several groups, according to the criteria defined by the packaging model. The second step reorders the groups, and the documents within each group. It may directly follow the documents' rank order or it may be recomputed as a function of both the rank and

some of the documents' attributes. For instance, the documents may be presented in reverse chronological order, indicating the recency of events, independently of their relative importance. The final step selects which components of each document will be passed to the formatting process.

Thus, the packaging process results in a sequence of tuples containing: a group descriptor, a sequence of documents, the corresponding document view and the "target" document views, in case the selected view contains links.

In the *MyNews* example, using the packaging model of Section 3.3, the documents shown in Table 1 will be rearranged as follows:

- *front page*: contains documents doc2, doc1, doc4 and doc53, in this order;
- *Thematic Group A*: contains document doc1, doc4, doc53 and doc7, in this order;
- *Thematic Group C*: contains document doc2;
- *Thematic Group D*: contains document doc18.

The documents in the front page retain the following components:

- the first three documents - docs2, doc1 and doc4 – follow view dv3 in Figure 3, that is, they retain the title, subtitle, summary and image, if any;
- the last document, d53, follows view dv1, that is, it retains the title and subtitle.

The documents in Thematic Groups A, C and D follow view dv4, that is, they retain the title, subtitle and the whole text.

This result may be represented as the following sequence:

```
[
    (front pagea [doc2, doc1, doc4], dv3, dv4),
    (front pageb [doc53], dv1, dv4),
    (thematic group A, [doc7], dv4),
    (thematic group D, [doc18], dv4)
]
```

Figure 4 schematically shows the final result of the packaging process, including the hyperlinks between the filtered documents. Note, however, that the document views used in Figure 4 do not imply any formatting scheme. They are used just for illustrative purposes.

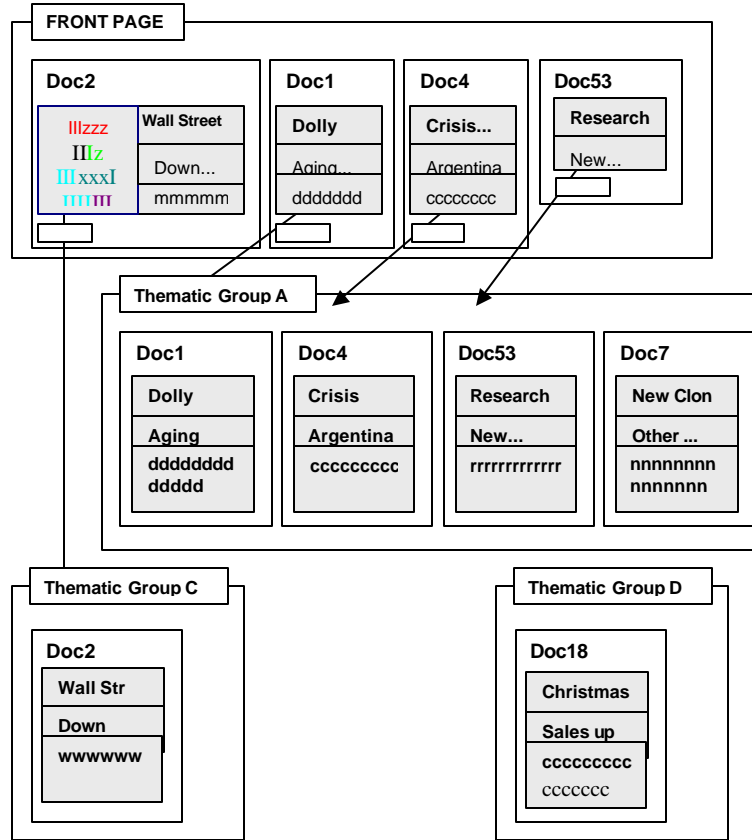


Fig.4. Packaging of the documents in Table 1.

6 Conclusions

In this paper we have defined a framework that helps design personalization systems. Our major contribution was to organize the framework into three levels that achieve a high degree of separation of concerns, by dissociating the packaging process from the filtering and formatting processes.

The framework facilitates the implementation, user testing and fine-tuning of the system representations and algorithms. The filtering process, for example, has three hotspots - the filtering algorithm, the document model and the user model - that can be instantiated in different ways. Likewise, the packaging process has two hotspots - the packaging model and the packaging algorithm.

If we consider the filtering process, our approach is similar to other approaches in the literature. It aims to select the documents (topics and news) that are most relevant to each user.

In our framework, we introduce the packaging level, which allows us to define the abstract organization of the selected information, by selecting and organizing relevant elements of the filtered documents.

Our framework may be used as a resource for analyzing existing approaches to personalized systems. Ossenbruggen's work on *Cuypers* [16] decomposes what we call formatting level according to their algorithms for automatic generation of the final multimedia presentation. If we can design algorithms to generate an adequate semantic structure based on communicative intentions, our packaging process would be able to provide the input for their prototype. Our packaging process would be thus positioned before the *Cuypers Engine*.

Our work is somewhat similar to the IMMPS proposal [6], which defines a Standard Reference Model for Intelligent Multimedia Presentation Systems. Their *design layer*, however, encapsulates our packaging level and, to some extent, our formatting level as well.

The approach described in [3] proposes an electronic newspaper that allows the personalization of the content and presentation detail of the news item based on receptivity (a dimension used for estimating the amount of information that a user might read). In place of our packaging process, their approach provides only two different types of pages for organizing information: index pages and news pages. Our instantiation, *MyNews*, maintains the distinction between the packaging and formatting processes, allowing editors and designers to experiment more easily with alternative solutions.

We are currently implementing the framework described in Section 2, with the *MyNews* instantiation [15]. We plan to investigate an alternative framework where the packaging process drives filtering. Also, we plan to extend the framework to address the problem of processing specific user requests, such as keyword searches.

Acknowledgments

Lucimar Martins would like to thank CAPES for supporting her work. Tatiana Coelho and Simone Barbosa thank CNPq for supporting their research.

References

1. Ardissono, L., and A. Goy. *Tailoring the Interaction with Users in Electronic Shops*. In J. Kay, ed.: UM99 User Modeling: Proceedings of the Seventh International Conference. Wien New York: Springer-Verlag, 35-44, 1999.
2. Ardissono, L., Console, L., and I. Torre. *On the application of personalization techniques to news servers on the WWW*. In: Lecture Notes in Artificial Intelligence N. 1792. Berlin: Springer Verlag, pp. 261—272, 1999.

3. Ardissono, L., Console, L., and I. Torre. *Strategies for personalizing the access to news servers*. Working Notes of the Adaptive User Interfaces. Spring Symposium of AAAI (Technical Report SS-00-01), pp. 7-12, Stanford, CA, 2000, AAAI Press.
4. Billsus, D. and M. Pazzani. *User Modeling for Adaptive News Access*. User Modeling and User-Adapted Interaction 10(2-3), 147-180, 2000.
5. Billsus, D., Pazzani, J., and J. Chen. *A Learning Agent for Wireless News Access*. Proceedings of the 2000 International Conference on Intelligent User Interfaces , 2000, Pages 33 - 36.
6. Bordegoni, M., Faconti, G., Maybury, M.T., Rist, T., Ruggieri, S., Trahanias, P., and M. Wilson. *A Standard Reference Model for Intelligent Multimedia Presentation Systems*. Computer Standards & Interfaces, 18(6-7):477-496, December, 1997.
7. Bradley, K., Rafter, R. and B. Smyth. *Case-Based User Profiling for Content Personalization*. Book: AH. 2000.
8. Brusilovsky, P. and D. W. Cooper. *ADAPTS: Adaptive hypermedia for Web-based performance support system*. Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW. 1999.
9. Cingil, I., Dogac, A., and A. Azgin. *A broader approach to personalization*. Communications of the ACM 43(8):136-141, August 2000.
10. Fayad, M., Schmidt, D., and Johnson, R. *Building Application Frameworks*, Wiley Computer Publishing, 1999.
11. Fink, J., Kobsa, A., and J. Schreck. Personalized Hypermedia Information Provision through Adaptive and Adaptable System Features: User Modeling, Privacy and Security Issues. Proc. of the Workshop Adaptive Systems and User Modeling on the World Wide Web of the 6th Int. Conf. on User Modeling, Chia Laguna, Sardinia, June 1997.
12. Hof, R. D., Green, H., and Himmelstein, L. *Now it's YOU WEB*. Business Week, pages 68-74, October 5, 1998.
13. Kamba, T., Bharat, K., and M. C. Albers. *The Krakatoa Chronicle – An Interactive, Personalized, Newspaper on the Web*. In Proc. of the Fourth International World Wide Web Conference, pp. 159-170, Nov 1995.
14. Khan, L., and D. McLeod. *Audio Structuring and Personalized Retrieval Using Ontologies*. In Proceedings of IEEE Advances in Digital Libraries, Library of Congress, Washington, DC, May 2000.
15. Martins, L.C. *Personalização de Visões sobre Documentos Hiperídia*. Technical report in preparation. Informatics Department, Pontifical Catholic University of Rio de Janeiro, 2002.
16. Ossenbruggen, J.V., Geurts, J., Cornelissen, F., Rutledge, L., and Lynda Hardman. *Towards Second and Third Generation Web-Based Multimedia*. In The Tenth International World Wide Web Conference, pages 479-488, Hong Kong, May 1-5, 2001
17. Parsaye, K. *PQ: The Personalization Quotient of a Website*. Published by personalization.com. NovuWeb, Inc. 2000.
18. Tarr, P., Ossher, H., Harrison, W., and S. Sutton. *N Degrees of Separation: Multi-Dimensional Separation of Concerns*. Proceedings 21st International Conference on Software Engineering (ICSE'99), May 1999.