# IRAIA: A PORTAL TECHNOLOGY WITH A SEMANTIC LAYER COORDINATING MULTIMEDIA RETRIEVAL AND CROSS-OWNER CONTENT BUILDING

Kurt Englmeier and Josiane Mothe
*German Institute for Economic Research (DIW), Berlin, Germany, and*
*Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse, France.*

Abstract: This paper presents a technology for information portals that supports multimedia retrieval while focusing at the same time on different ecologies of information provider and content owner environments. This means it addresses cross-media information provision and cross-owner content building. The technology was developed under the FP5 project IRAIA[1] and was applied to two different areas: economic information and cultural information related to puppetry[2]. The first area puts more emphasis on multimedia retrieval while the latter one has special requirements in respect to cross-owner content building and content delivery. IRAIA enriches the implementation of multimedia, cross-media, and cross content owner retrieval with a design flexibility that enables an information provider to adapt the delivery channel as well as the content building process to requirements specific to its service. Leading through IRAIA applications in the areas mentioned above we present IRAIA's system design and architecture, its human-computer interaction modes, and its powerful semantic layer that links the different facets of cross-media delivery and of cross-content building.

Key words: Context-aware computing, cross-media information provision, cross-owner content building, intecreation.

[1] The project IRAIA (IST-1999-10602) was co-funded by the European Commission under the 5th framework programme.

[2] The IRAIA information service related to puppetry is accessible through www.epuppetry.com

# 1. A RATIONALE FOR INFORMATION MEDIATING

The ultimate purpose of any presentation of information is to communicate properties of the information to a human. In the realm of economic information this may range from information items such as time series or texts to the whole information space comprising a database's content. Cultural information may more include images, video and audio documents.

The framework of information provision in IRAIA follows the discussion in Wiss and Carr (1998) on the presentation of information spaces and is extended by a semantic dimension which is extremely important in the realm of multimedia and cross-content owner information. The framework emerges from the following three basic aspects:

Attention: the user must be in the position to focus attention on certain elements of the information presented. The human perception, however, includes the ability to register semantically presented context, i.e. the arrangement of keywords describing an information space.

Abstraction: user perception is focused on important parts of an information space when the amount of information becomes too large preventing the user to perceive all significant information elements at the same time. Filtering removes a global context if it becomes too complex. Abstraction groups information to higher-level elements that retains the focus of perception on the global context. Both are functions of information hiding that confront the user only with as much detail as needed at a certain time. In abstraction, groups of elements can be completely replaced by a single element which represents the entire collection. The new higher-level element may be different from the lower-level elements in some way, and it may reveal its underlying structure only on closer inspection. Again, this relates to both the information's purely graphical and its semantic presentation.

Affordances: they relate to querying an information space and presenting the parts relevant to a query. From this aspect emerged the rationale of a "language" for affordances in portal applications. In the context of information ambience addressed here, this language is highly related to the significant vocabulary used in the economic analysis or in the creation of artifacts for puppetry.

From these basic principles we developed functional levels for the framework of information presentation as used in IRAIA. Taking up the discussion in Card (1996) the corresponding functional levels are the following:

Infosphere: it is the total amount of information available that needs to be presented. Here typically abstraction techniques apply.

The workspace is a subset of infosphere the user currently needs. Attention techniques can be used to develop a task adequate workspace from the infosphere. Abstraction is used to arrange the workspace. The concept of workspace essentially depends on the rationale of the users' intentions regarding the data collection. The workspace may be constructed using a museum as metaphor which means as design rationale. Navigating such a workspace is equivalent to a virtual visit of a museum as an informal activity that offers education, entertainment, leisure, and so on. Knowledge acquisition is one of the visit's most relevant aspects.

The museum metaphor is suitable if the users should be guided through certain sections of the collection for learning purposes, for instance. This design of a workspace can be completely inappropriate if the user is only looking for on spot of the information space like the latest employment figures of the chemical industry. It makes no sense in this case to start with the explanation of employment models.

The third level corresponds to Card's "sensemaking tools". These "help users to understand information by associating and combining it." They are related to attention techniques that help the users perceive the focus points of their interest. The affordance aspect is important in these tools and makes it easy for the user to manipulate the data via sensemaking tools. In the context of IRAIA, this level presents the semantic surroundings of a document actually focused, i.e. the annotated concepts.

Finally, the document level discusses presentation of the elementary units of data such as time series and related texts or images and related puppet plays, for instance. This functional level relates to the multimedia presentation of data.

In the following sections we present the concrete realization of these levels in IRAIA.

## 2. DESIGN AND PRESENTATION OF THE IRAIA INFOSPHERE

Information arises from data when they are combined, arranged, and presented accordingly. Only a suitable combination of related time series and texts is in the position to convey the information that is contained in these separate and otherwise imperceptible components. An infosphere providing suitable abstractions that express the content of these components in a

comprehensive and uniform way is of outstanding importance when it comes to proliferate information that has to be composed by distributed and heterogeneous data. Our model for the construction of an infosphere is related to models of retrieval environments for interacting with large data collections (Agosti et al. 1992, Krause 1996). In our context an infosphere is a retrieval environment where the users explore data collections within a semantic coordinate system derived from taxonomies of the respective information domain. These taxonomies exist for a variety of application areas. They are a solid basis for a controlled and structured vocabulary and therefore pretty appropriate for content semantics defining a domain-related context.

## 2.1 Multimedia retrieval

The aspect of context-oriented retrieval has to be stressed alongside semantic infrastructure, scalability, and interoperability of federated repositories (Englmeier, 2000; Ramakrishnan, N. and Grama, A.Y., 1999).

*Figure1.* Searching and navigation in a semantic coordinate system for economic information. Selected concepts make up the initial query profile. While realizing their retrieval strategy the users usually perform iterative steps of defining a query and analyzing the retrieved results. In IRAIA, the documents are annotated solely with entries from the hierarchies.

The success of a design for an infosphere depends on the capability of its underlying approach to what extend it helps to reduce information overload in using the functional levels of the system. Now the challenge is to find an appropriate representation of an infosphere together with suitable workspaces that emerge while analyzing the data of a specific domain

(Meisel and Sullivan, 2000). It can be pretty helpful if the system provides the users with semantic elements that help to describe concisely the things they expect to find while roaming an infosphere. These semantics shown in a suitable way enable the users to find easily the locations of the documents related to the context of the users' search problem, i.e. to draft easily the relevant workspace.

A semantic coordinate system endows the users with a concise as well as comprehensive vocabulary. Hierarchically arranged and grouped along major content facets this vocabulary acts as a stable coordinate system easy to comprehend and memorize[3]. The users are thus much more in the position to localize themselves effortlessly. Successfully searching and navigating now means guided travelling from information to information just by changing the semantic coordinates, i.e. by pointing to relevant concepts. This structure on the other hand enables to pinpoint the semantic location of any kind of information.

For the ambience of economic information we produced a powerful taxonomy that merges two of the most important structures in this field: eurostat's NACE[4] nomenclature and the industry systematic of the ifo institute for economic research. The unified taxonomy creates a semantic coordinate system that enables exact and automatic positioning of coherent documents even if they are of different types. It also provides users with the necessary orientation while exploring the infosphere or the actual workspace. Like in using languages it helps users as a passive vocabulary to identify the topics of their information problem.


*Figure 2.* A sample of query results reflects a workspace. Its corresponding documents are presented together independent from the type of data and even the underlying query language.

[3] The coordinate system itself can be presented simultaneously in different languages. This ensures that the domain model mentioned later features multilinguality.

[4] Nomenclature des Activités dans la Communauté européenne -systematic of the economic activities of the European Union

Grouping large data samples along easily discernible aspects of an infosphere tackles the problem of ambiguity which the traditional search engines cannot master. This means that at each navigation step a user gets only a workspace, precisely tailored to the selected entries in the respective concept hierarchy, and taking into account multimedia documents actually retrieved. This also means that the sample of retrieved document (i.e. the workspace) form a contextually coherent group and are presented as such.

*Figure 3.* An example from a presentation of cultural information related to puppetry: Similar to the portal for economic information the infosphere here is structured along the categories "attributes", "roles", and "characters". The document is presented together with its semantic surroundings. The interface shows the annotated concepts that are thus already familiar from the initial query formulation. Modifying the set of annotated concepts is thus tantamount with repetitive query formulation. The user can always invoke the respective concept window if concepts are required in addition to those annotated to the document.

Independent of their data type the users get the information that belongs together for answering their query. Of course, different types of data are handled by their corresponding tools. These "sensemaking" tools provide for a suitable information presentation and include in some cases also manipulation functions such as statistical methods in the case of economic information, for instance.

It is very important to point out that these presentation features work the same way for an infosphere regarding cultural information. The presentation tools are also in the position to render images, videos, and audio which is much more relevant for this area.

## 2.2 Cross-content delivery

The application of IRAIA to the ambience of cultural information focuses on an infosphere that is built around material related to puppetry such as

plays, descriptions of puppets, abstracts of plays, and useful literature like tales. Unlike the IRAIA services for economic information the digital collection of puppetry is available for free and lives from the free contributions of its community. This trait reflects best the capabilities of IRAIA's design of cross-owner content building: It supports self-organizing and self-sustaining service established for and through the community active and interested in puppetry. In the area of economic information the aspect of content building addresses more the integration of the data collections from the different economic research institutes and national statistical institutions. In the area of puppetry IRAIA went a step further and regards all the creative talents such as authors or puppeteers as content owners.

Unlike other systems preserving cultural content the puppetry application of IRAIA does not base on already existing (physical) collections of one or a couple of content owners – primarily museums or archives. This means it cannot resort to an approved information arrangement, too, i.e. to a systematic that categorizes the collection's objects and structures their presentation. The arrangement of physical objects lends itself for a comprehensive and often complete organization of virtual museums and the like. The availability of this organizational groundwork is quite an advantage in the design of infospheres. The essential disadvantage of this kind of information arrangement lays in its static and proprietary design. An approach for one infosphere can be barely applied to another one. This means it is not suitable for an information service that bases on the works of a volatile and highly disconnected group of content owners. There's no organizational groundwork other than the semantic structure reflecting themes and concepts that appear in puppetry.

## 2.3 The link to inspirational desire

For the ambience of cultural information related to puppetry IRAIA provides an adaptation opportunity that brings serendipity into focus. Overt expressions of serendipity combined with practicality are rare in the realm of information retrieval engineering. The adaptability designed for the query feature in IRAIA can achieve the marriage of technical solutions and inspirational desire. It meets a collection's structural requirements while evoking a sense for exploring an immediate information ambience. Its flamboyance inextricably weds the stunning and the practical.

The goal here is to be a source of inspirations for the ideas of puppeteers – including professionals and amateurs. The things puppeteers are after can be best clarified by the following example from a talk with the ensemble of a

puppet theatre: “We were looking for a new play for children that should have some educational effect through a certain level of cruelty as it occurs likewise in many of Grimm’s fairy tales. We looked for children’s books in libraries, bookshops and even toy shops and run into a book in English telling the story of a mother and cannibal looking for a child to eat. The design idea for puppet came from a picture in a festival announcement showing a puppet representing a fierce devil.”

The cornerstones of this user interactions are the following:

- The access to a digital archive must enable the creative talents to go directly to the spot they are looking for while bypassing the vast majority of the collection.
- The interaction mode must cope with the requirement that users show up with a very vague idea that steers their searching and navigating. However, this idea is the only steering element.
- The structure of guidance provided by the system must be sensitive enough to let the users keep their vague idea as the steering element. Any too rigid structure may cause an unacceptable information overload.
- There is no such thing like a precise answer to this vague query. Anything that relates to it can be useful. Thus, any kind of information found by serendipity is welcomed.

The characteristic outlined so far demonstrates the expectations of a user community requiring a platform that supports actively their work - expectations that determine the rationale of an information platform in the specific context of fostering creative talents. Such a platform must be different from an application that digitizes an archive or museum with their functions of information mediating as a leading design rationale. These users clearly do not want a portal realizing a virtual tour through a puppetry museum. This may be expected by someone who wants to be informed of puppetry as a whole or at least of more comprehensive blocks of this theme. This user expects to get an overview of the virtual collection she or he is entering, to be guided through the collections and to have learned something from this tour.

The situation is different when it comes to the provision of economic information. Here the users are less interested in serendipity, they have a much more precise conception of their information problem and want to get quickly to the relevant spot in the infosphere. Thus the IRAIA technology can be adapted to both types of retrieval strategies.

# 3. A SEMANTIC LAYER FOR A MULTIMEDIA AND CROSS-CONTENT INFOSPHERE

In the end, every design model for an information providing system goes back to the good old quest to find an efficient matching between the representation of documents and the user's information need. A query usually manifests an information need. Satisfying an information need usually is the result of an iterative process of querying and analyzing the retrieved results. Through this process the information need itself is often refined. Thus only a extensive sequence of these iterative steps representatively reflects an information need. The querying and navigation process itself can be regarded as a probabilistic inference process that compares document representations (descriptors) based on different forms of linguistic and/or statistical evidence on the one with representations of information needs (queries) based on similar evidence on the other side. The framework for the design of the matching processes is generally known as the inference network model. Different representations of documents and queries (words, features, phrases, additionally assigned words and structured combinations of them) can be combined in a consistent probabilistic network. (Callan et al., 1995)

## 3.1 Content Matching

The task to be carried out corresponds to document categorization. While parsing the concept hierarchies the entries of their nodes are treated as profiles. These profiles are composed of phrases because the entries of the concept hierarchies usually consist of a couple of words describing a concept. From the lexical analysis results an index list of words that is in this form a parsed representation of the controlled vocabulary. Alternatively, a profile is composed of a set of phrases automatically extracted during a learning stage. This latter method is used when a training document set is available that is to say when the system is provided with documents and their associated correct entries. The salient difference is, that like in text parsing the underlying morphology of the concept trees is preserved. This allows the expansion of the meaning of a node by propagating the content from its ancestors. Thus, a node together with its ancestors can be regarded as a pseudo-document that may be helpful in the further matching process.

This process compares, roughly speaking, the content of the concept hierarchies with those of the documents and decides which tree nodes are the most prominent ones for this document. The entries of each hierarchy are

ranked according to the belief that those are relevant to the document. This annotation process serves to create references from the document to the hierarchies of one or more thematic domains.

A threshold ensures that only those entries are selected that contribute to an appropriate abstraction of the document. It's obvious that the ancestors of an entry get less weight than one of its subordinal concepts (in terms of specialization). This bases on the assumption that only most specific terms are in the position to capture the specifics of a document's content and these terms can be found towards the respective ends in the ramification of an hierarchy. References from a document to the most specific entries of a hierarchy lead to a higher precision.

## 3.2 Evaluation and Annotation

The following evaluation calculates the belief in a profile (a node in a concept hierarchy) due to the occurrence of a concept $c_i$ of a document d:

$$bel_{Eh}(d) = \sum_{c_i} \left( \frac{tf_i}{s(d)} \cdot \frac{s(h)}{tfE_i} \right) \cdot e^{\frac{s(Eh \cap d)}{s(Eh)}}$$

where

*Eh= an entry or node from the concept hierarchy h, represented by a profile,*

*$C_i$ = a concept from the profile Eh,*

*$tf_i$ = frequency of the concept $C_i$ in the document d,*

*$tfE_i$ = frequency of the concept $C_i$ in the node Eh,*

*s(d) = size of the document d (number of concepts),*

*s(h)= size of the concept hierarchy,*

*s(Eh) = size of the node Eh,*

$s(Eh \cap d)$ *= number of terms from Eh that occur in Eh.*

$\frac{tf_i}{s(d)}$ *measures the importance of the concept in the document,*

$\frac{s(h)}{tfE_i}$ *measures the importance of the concept in the concept hierarchy and*

$\frac{s(Eh \cap d)}{s(Eh)}$ *measures the rate of occurrence of the concepts from the concept hierarchy node in the document (coverage).*

This formula is derived from the well-known *tf.idf* weighting function. The values are normalized and remain between 0 and 1. Concept phrases for a given document are ranked by this function. A concept phrase is selected for annotation if its value is above a certain significance threshold (let's say

0.4, for instance). Due to practical reasons only the three entries highest in

$$ann(d) = \bigcup_{h} \begin{cases} bel_{Eh}(d) \text{ with } bel_{Eh}(d) >= \delta_0 \\ \{\} \text{ otherwise} \end{cases}$$

ranking as well as above the threshold are chosen.

where

$\delta_0$ = *significance threshold*

*h = concept hierarchy*

Calculations are performed for all concept hierarchies. In principle, this evaluation can be applied at the entire document or parts of it as titles, paragraphs, and the like. We emphasize once again that the matching process bases exclusively on concept phrases. These entries of concept hierarchies (as well as the controlled vocabulary derived from) define the semantic context of the specific domain-related information space where the users' browsing and navigating is taking place. This also means that users are interested in a document primarily because of being related to this context, albeit the aspects covered by this document reach far beyond this topic.

## 4. CONCLUSION

Perhaps one of the largest surprises in computing at the end of the 20th century was the extent to which entertainment became a major driving application. The power of the consumer market far outweighs in the meantime that of the scientific or even defense markets. This is a landmark indicating that computing found its way into a broad use of the society through entertainment. It seems only natural, then, that interactive entertainment should be considered as a valid application area for providing cultural as well as scientific information and learning. This starts with intecreation when content owners, users, and experts rather than technology people are the major player in designing the presentation of content that should be both informative and entertaining.

IRAIA followed this rationale in the development of a platform technology that supports cross-media information provision and cross-owner content building. Its user interface design emerges from the paradigm of intecreation, the interactive and cooperative development of content for an information ambience – a feature fostering the creation of information repositories that lay ground for virtual communities and new ways of content and knowledge exchange.

The outstanding element of the IRAIA interface is its flexible and appropriate representation of an infosphere together with suitable workspaces that emerge while analyzing the data of a specific domain – an interface that provides for precise queries as well as for serendipity.

From the puppeteers' point of view IRAIA is a platform technology for "boxes of inspirations". Viewed from a more global angle it is a mining system. Unlike typical data mining systems that restrict mining to numerical data, IRAIA's multimedia character enables to link heterogeneous data types on a semantic level. So it enables mining simultaneously in factual data ("hard facts") as well textual data ("soft facts"). This characteristic is of an information system is also required for other applications areas as presented here when it comes to providing access to huge collections that are made up by a number of heterogeneous databases.

## 5. REFERENCES

Agosti, M.; Gradegnio, G.; Marchetti, P.: A hypertext environment for interacting with large databases. *Information processing and management*, 28 (1992), 371-387.

Callan, J.P.; Croft, W.B.; Broglio, J.: TREC and TIPSTER experiments with INQUERY. Information Processing and Management, 31, 327-332, 343. 1995.

Card, K. S.: Visualizing retrieved information: A survey. IEEE Computer Graphics and Applications CG&A, 16(2), pp. 63-67, March 1996.

Englmeier, K.: The link between data and tasks - the crucial challenge in designing user interfaces for information retrieval systems, in: Human-Computer Interaction. Communication, Cooperation, and Application Design. : H.-J. Bullinger; J. Ziegler (eds.), Vol. 2. Proceedings of HCI International '99. 1999.

Krause, J.: Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung. ("Schalenmodell"). IZ-Arbeitsbericht Nr. 6. 19, 1996.

Meisel, J.B. and Sullivan, T.S.: Portals: the new media companies. info – the journal of policy, regulation and strategy for telecommunications information and media. 5(2), 2000.

Ramakrishnan, N. and Grama, A.Y. (1999): Data mining: from serendipity to science. IEEE Computer. Vol. 32, no. 8, 1999.

Wiss, U. and Carr, D.: A Cognitive Classification Framework for 3-Dimensional Information Visualisation. Technical report of the department of computer science and electrical engineering of the Lulea university of technology. 1998:04.

## ACKNOWLEDGEMENT

Research outlined in this paper is part of the projects IRAIA and eStage that were supported by the European Commission under the Fifth Framework Programme (IST-1999-10602 and IST-2000-28314). However

views expressed herein are ours and do not necessarily correspond to the IRAIA or eStage consortium.