Media-independent correlation of Information: What? How?

Amir Shit * Vimal Kashyap **
*
Department of Computer Science, University of Georgia
415 OSRC, Athens GA 30602-7404
** Department of Computer Science, Rutgers University
New Brunswick NJ 08903

Abstract

A wide variety of data in different media (text, structured, image, audio,...) is now accessible in Internet and Intranet. Limitations of current keyword-based and hyper-media link-based access are all the more apparent when a user wants information without prior knowledge of the locations or media that have data to meet his/her information needs. Two of the several components of the solution being investigated in the InfoShuttle project are discussed here: (a) the use of a wide variety of media, that abstract information content with regards to the location, format, or media of data; and (b) the use of a logical level linking of data called a metadata reference link (MREF) as a hypothetical extension to HTML, to support metadata-based information correlation between heterogeneous digital media data stored in distributed repositories.

1.0 What is media-independent correlation?

In recent years, huge amounts of digital data in a variety of structured, unstructured (e.g., images) and sequential (e.g., audio) formats have been collected and stored in thousands of repositories and on CD-ROMs. Significant advances in managing textual, image, audio, and video databases support efficient storage and access of data of a single type in a single repository. Affordable multimedia systems and a variety of interworking tools (including the current favorite, the World Wide Web (WWW)) allow creation of multimedia documents, and support access and presentation of such data.

However, information needed by a user or application may be stored in multiple media or forms (e.g., structured data, text, image, audio, and video) in different repositories. With a typical user already struggling with information overload, it is unrealistic to expect him/her to be aware of all relevant stored information and the characteristics of the various media. This problem can be substantially alleviated if we can provide a user with the ability to specify higher information request at a higher semantic level independent of the media in which the information is represented. Can this be done with a relatively simple mechanism and without burdening users or administrators with complex models, languages and huge initial effort? One such approach that utilizes symbolic metadata descriptions and their mapping to underlying multimedia data is discussed in the next section.

The inability of correlating information across multiple media representations in current systems has been characterized as a "semantic bottleneck" [Jain]. Here, we argue that identifying relevant information at the instance level using content-independent information (e.g., specifying the URL in an A HREF statement in HTML) has to give way to more content-dependent and media-independent approaches in order to address the problem. We view media-independent correlation as:

- Using symbolic metadata descriptions to capture the information content in data represented using different media-types;
- Establishing relationships between the metadata descriptions;
- Collating and presenting the data retrieved (which may be represented in different media-types).

Rest of the paper is organized as follows. In Section 2, we introduce a variety of metadata that can be used to establish media-independent correlation between data stored in different repositories. Then we introduce the concept of metadata reference link as a potential addition to HTML and sketch an implementation to support this concept by the client and server side extensions in the WWW. Correlations involving content-independent metadata, content-dependent metadata, and using domain specific information are discussed in Section 3, 4 and 5, respectively. Section 6 provides some conclusions.

2.0 Metadata: The key to media-independent correlation

We propose an approach which involves representing information in the various media representations as "symbolic" metadata descriptions. Systems for dealing with alphanumeric or symbolic information efficiently are freely available and do not represent
a technological challenge. Hence, symbolic descriptions are a natural choice. Correlation of information may then be established by querying and computing the symbolic metadata descriptions, using existing, well-developed techniques.

2.1 What is Metadata?

Metadata represents information about the data in individual databases and can be seen as an extension of schema in structured databases. They may represent relationships between individual media objects. These metadata descriptions may be extracted using various mappings/extraction (e.g., see [Hilliard, 1995]) associated with the various types of digital data. In this paper, we consider the following types of metadata (see [Hilliard, 1995] for a classification):

- **Content-independent metadata:** This type of metadata does not capture the content information of the document it describes, e.g., location, date of creation, etc.

- **Content-dependent metadata:** This type of metadata captures the content information of the document. We define three types of content-dependent metadata:
  - **Content-dependent metadata:** This type of metadata depends directly on the document content, e.g., keywords appearing in a document, colors appearing in an image document. One method of representing content-based metadata is using a collection of attribute-value pairs. A discussion of attribute-based access for textual data is discussed in [Schiele, 1995]. The attributes chosen may be media specific (e.g., color) or media independent (e.g., location, relief).
  - **Content-descriptive metadata:** This is a special case of Content-dependent metadata where the content of a document is described in a manner which may not be directly based on the contents of the document. Examples of content-descriptive metadata for images may be found in [Chabot, Kryski] where textual annotations are associated with images and are used to correlate information across image and textual documents.
  - **Domain Specific metadata:** This is a special case of content-descriptive metadata typically represented in an attribute-based manner where the attributes used to characterize documents are domain specific in nature, e.g., relief for the Geographical Information Systems domain. In some cases the metadata may be obtained from domain-specific ontologies and may be represented in a description language like RLQ or DL-Lite.

2.2 Information Correlation using Metadata Reference Link (<A MREF ...>)

How much of the correlation is done automatically by the query processing system? The level of automation usually depends (inversely) on the information content captured in the metadata. How meaningful is the correlation? This, on the other hand, depends (directly) on the information content captured. For query processing systems to adequately address these design considerations, it is desirable to move towards location-independent, media-independent content-dependent methods of correlation specific to the domain of information. We shall show how the type of metadata influences the correlation of information. To discuss the issue using the all too familiar HTML specification, assume a simple (hypothetical) extension to handle the processing of metadata (here MREF stands for metadata reference link and implies using a "metadata-based reference"): 

- `<A MREF KEYWORDS="list-of-keywords"> THRESH=real </A>` Document Description `<A>`

- `<A MREF ATTRIBUTE="list-of-attribute-value-pairs"> Document Description `<A>`

An Implementation of `<A MREF ...>

We now discuss a prototype implementation in progress that is intended to show the viability of using the above concept. Implementation of the metadata reference link would require modifications or additions both to the browser and to the HTTP server. For the browser side, the metadata reference must be supported by implementing a Netscape plug-in which is activated whenever the browser encounters a `<A MREF ...>` statement. The server side modification will require enhancing the protocol to pass the appropriate keywords and attribute-based metadata to the appropriate servers. This may also require the local server to keep track of the metadata and support to forward metadata requests not supported by it to appropriate remote servers. This can be implemented using a simple registration service.

The work on plug-ins on the browser side is nearly complete. Plug-ins are distributed in the form of libraries and are wrapped as a new mime type is encountered in a browsing session. The control of a window in the Netscape frame is given to the appropriate plug-in which can respond to events like mouse clicks. When the Netscape application creates the sub-window, it needs to be sub-classed for the plug-in which results in replacing the window procedure of an existing window class with the customized one. In our plug-in, a window is opened whenever an `<A MREF ...>` initiation is encountered. The window displaying the `<A MREF ...>` link handles a mouse click on the metadata reference link in an appropriate manner. We have also started work on modifying the HTTP server to support enhancements to the HTTP protocol that shall enable us to handle metadata reference.
While the functionality and flexibility of Plug-ins leave a lot to be desired in comparison with what can be achieved with Java on the Web, the restrictions placed on Java have made plug-ins the choice of developing approach over Java when the restrictions become overly restrictive. One of the reasons we chose plug-ins over Java was because internet access is not supported from within an applet.

3.0 Content Independent Correlation

This type of correlation arises when content-independent metadata (e.g., the location expressed as a URL) is used to establish the correlation. The correlation is typically media-independent as content-independent metadata typically does not depend on media characteristics. In this case, the correlation has to be done by the designer of the document and all the query processing system does is to access the document(s) from the relevant locations (which is the content-independent metadata in this case) as illustrated in the following example:

```html
<TITLE> A Scenic Sunset at Lake Tahoe </TITLE>

<p>
Lake Tahoe is a very popular tourist spot and
<a href="http://www2.szpr-er.edu/lake_tahoe.html">some interesting facts</a> are available here. The scenic beauty of Lake Tahoe can be viewed in this photograph:
</p>

<IMG ALIGN=MIDDLE SRC="http://www2.szpr-er.edu/lake_tahoe.jpg"

</center>

The correlation is achieved by using physical links without any higher level specification mechanism being used. This is predominantly the type of correlation found in the HTML documents on the World Wide Web (WWW) and is typically done manually. The meaningfulness of the correlation depends on the assumptions made by the designer of the document and does not involve any sophisticated query processing.

4.0 Content-dependent approaches for Correlation

In this section, we discuss how we can establish correlation across data in different media using more content-based approaches. This type of correlation can be accomplished when content-dependent metadata are used to define relationships across information in different media. We shall discuss two approaches: (a) using content-based metadata; and (b) using content-descriptive metadata.

4.1 Correlation using Content-based Metadata

We present below an example based on a query in [Chab] to demonstrate a correlation involving attribute-based metadata. One of the attributes is color which is a media-specific attribute. Hence we view this interesting case of correlation as media-specific correlation.

```html
<TITLE> Scenic Waterfalls </TITLE>

<p>
Some interesting <a href="#">waterfalls</a> are available here.
</p>

</center>

The portion in bold represents the attribute based query. As may be observed, the onus of determining the URLs and the document satisfying the above conditions is no longer on the designer of the HTML document. One strategy to use the metadata of these types is:

- The `<A MREF ...>` statement is translated into the location of the relevant image and text indices on the underlying server.
- The keywords based condition is passed to the text index which is constructed using the content of the textual documents and is an example of content-based metadata for text documents. The condition based on color is pasted to the image document collection which may have been indexed based on the predominant color of the images. This is an example of content-based metadata for images.
- The results obtained from the appropriate text and image indices are collated together by the server and passed to the browser for display.
Metadata Extraction

The key to the above strategy is the ability to extract content-based metadata from the underlying (text or image) data. Most textual indexing technologies rely primarily on this kind of metadata which is based on the text of the documents [4][5][Meta]. There are generally two flavors of representation/storage of textual metadata.

- An inverted index which keeps track of the documents in which a keyword occurs and its frequency within that document (e.g., WAIS [6][7][8][10]).
- Vectors associated with documents which characterize their position in a multi-dimensional space. An example is LSI [9][10][11] in which both documents and keywords are mapped to the same vector space.

We have written image extractors which extract basic information like major color component, height, width and size for images stored in the GIF or PPM formats. Sometimes, this information can be extracted from the header information stored in these formats and does not require complex image analysis. An example of such an extractor is shown below.

```
<image: Color

Major component (RGB)

Blue
```

4.2 Correlation using Content-descriptive Metadata

In [Key-10] keywords are associated with images and a full-text index is created on the keyword descriptions. Since the keywohds describe the contents of an image, we consider these as content-descriptive metadata. Correlation can now be achieved by querying the collection of image documents and text documents using the same set of keywords as illustrated in this example:

```
<TITLE> Scenic Natural Sights </TITLE>
<p> Some interesting <a href="s<scenic waterfall mountain"> keywords="s<" "scenic waterfall mountain"> information on Lake Tahoe </a> is available here. </p>
```

The portion in bold represents the keyword based query. In this case, after the MREF statement has been mapped to the appropriate indices, the same keyword based query is passed to the textual index for text documents and the index for image documents. The results from the two indices having ranks greater than 0.9 are collated and sent to the browser to be displayed. The metadata extraction for this naive strategy simply bootstraps on available textual indexing technologies and is illustrated next.
This type of correlation is more meaningful than content-independent correlation. Also, the user has more control over the correlation, as he may allow to change the thresholds and the keywords. The keywords used to describe the image are media-independent and hence correlation is achieved in a media-independent manner. We are at present experimenting with techniques for correlation based on combining results from: (a) different types of indices for the same media; and (b) indices on different types of media.

5.0 Domain Specific Correlation

We have illustrated two instances of content-dependent correlation. The first approach using content-based metadata involves media-specific attributes like color for images. The second approach using content-descriptive metadata helps to correlate information across image and text media in a relatively more media-independent manner. However, the latter approach will not give the desired results when the keywords used by the user are different from the keywords in the documents and those associated with the image.

We believe that in order to better handle the information overload on the fast-growing global information infrastructure, we should support correlation of information at a higher level of abstraction independent of the medium of representation of the information itself. We need to model domain specific concepts for multimedia documents. Such metadata should be necessarily media-independent enough they may be mapped to the associated documents via media-dependent attributes. Let us consider the domain of a Site Location and Planning application supported by a Geographic Information System and a correlation query illustrated in the following example:

<TITLE> site Location and Planning </TITLE>

To identify potential locations for a future shopping mall, we present below all regions having a population greater than 300 and area greater than 50 square feet having an urban land cover and moderate relief: <A MREF ATTRIBUTES(POPULATION > 300; AREA > 50; region-type = block; land-cover = urban; relief = moderate)> can be viewed here </A>

The processing of the above query results in the structured information (area, population) and the map of the regions satisfying the above constraints being included in the HTML document. The query processing system will have to map these attributes to image processing routines which determine the color of the pixels at run-time. Thus the responsibility of processing the media-specific part of the correlation is now moved to the query processing system. The issues discussed here are being further explored in the InfoQuill Project at the LSDIS Lab of the University of Georgia. The main issues involved in processing the above query are:

- Construction of the <A MREF> expression using domain-specific attributes and values. An interesting step would be to re-use existing taxonomies and classifications.
- The ability of the underlying server to support querying based on the metadata specified in the <A MREF> expression. The server will either contain the metadata stored in its local metadata repository or will contact other servers which it knows contain the relevant metadata in their repository.
- Heterogeneity of the media in which the underlying data might be stored. Data satisfying the specified constraints can be
found, for example, in independent repositories containing census data and TIGER/LINE database.

Mapping the domain specific attributes relief, land-cover, area and population to the underlying data requires access to processing over additional image and structured databases.

5.1 Use of Domain Specific Ontologies

One approach to raising the level of abstraction of correlation is to use domain specific attributes and values. Our approach here is to re-use existing taxonomies and classification. We have applied this approach to the bibliographic information domain [BRDIVER] where we re-use portions of existing domain specific ontologies such as the Bibliographic Data Ontology [ARPA] and WordNet 1.5 [WordNet]. For the example illustrated above, we re-use two classifications, the Land Use Land Cover classification proposed by USGS and the population area classification by the US Census Bureau.

A classification using a generalization hierarchy

Land Use and Land Cover Classification (USGS)

- Urban
- Residential
- Commercial
- Forest Land
- Industrial
- Evergreen
- Deciduous
- Mixed
- Water
- Lakes
- Reservoirs
- Streams and Canals

A classification using an aggregation hierarchy

Population Area Classification (US Census Bureau)

- State
- County
- City
- Rural Area
- Tract
- Block Group
- Block

5.2 Heterogeneity of the Media of Representation

One of the key things achieved because of media-independent correlation is that a decision maker is completely insulated from the underlying heterogeneity. Consider the example discussed above. Information about the attributes population and area are obtained from a structured database storing Census Data, whereas land-cover and relief information is obtained from the image database storing USGS maps. Boundary information which is necessary for correlation is obtained from the underlying TIGER/LINE Database.
5.3 Mapping of the Domain Specific attributes

There are two basic approaches to map domain specific attributes to the underlying data.

- **Bottom Up**: In this case the metadata are pre-computed and stored in a metadata repository.
- **Top down**: In this case the metadata are associated with parameterized routines and computed at run-time.

5.3.1 The Bottom Up Approach

Consider the example given above. The bottom-up approach consists of executing domain and media specific metadata extractions. An example of a metadata extractor is the one which gets all the regions from the Census database and computes their boundaries from the TIGER/Line database. The boundaries thus computed are used to index into the image maps and the land-cover is computed using appropriate image processing routines.
5.3.2 The Top Down Approach

The top-down approach consists of associating parameterized routines with the attributes. When constraints are specified on these attributes at runtime, the associated parameterized routines are invoked to compute the metadata and return those values which satisfy the constraints. Two alternative schemas for storing the parameterized routines are illustrated below.

<table>
<thead>
<tr>
<th>metadata</th>
<th>function</th>
<th>param_list</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td>compute_area</td>
<td>&lt;minArea, maxArea&gt;</td>
</tr>
<tr>
<td>population</td>
<td>compute_population</td>
<td>&lt;minPopulation, maxPopulation&gt;</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>oid</th>
<th>county</th>
<th>block</th>
<th>land_cover</th>
<th>relief</th>
</tr>
</thead>
<tbody>
<tr>
<td>59716</td>
<td>59</td>
<td>716</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

5.4 Using Domain Specific Attributes for Correlation

The correlation of information based on domain specific attributes can be performed as follows. This strategy is based on the assumption that metadata is computed at runtime.
1. The metadata repository is first queried to determine the metadata exported by the system. This is used to dynamically construct the user interface.
2. The decision maker inputs the constraints on the user interface.
3. Determine the order in which the metadata constraints should be evaluated (the control strategy). The constraints are grouped into two groups: (a) group corresponding to the structured data; and (b) the group corresponding to the instance data.
4. Obtain the set of objects by evaluating the first group of metadata constraints and correlate them by computing the "intersection" of the sets.
5. For each object in the intersection obtained in Step 4 evaluate the second group of metadata constraints and select those objects which satisfy the constraints.
6. Display the objects and properties identified in Step 5 on the User Interface. If the decision maker is not satisfied, s/he can modify the constraints and go back to Step 2.

6.0 Conclusions

In this paper, we showed the use of a wide variety of metadata to establish media-independent correlation between data in different repositories in the WWW-based Intra- and Internet environments. The metadata reference link (MREF) proposed in this paper as a potential extension to HTML can un-tackle us from physical link and provide a powerful ability to correlate data at a logical level regardless of their location, format and media. Examples of using a variety of metadata to link data using the MREF construct are also given.

This work is part of the InfoQuilt project. Demonstration of use of a variety of metadata for structured, text and image data is supported by the InfoQuilt prototype system. It is now being extended to support metadata reference link as discussed in this paper.

Acknowledgements

We wish to thank Srilekha Madumbai for implementing the image extractors and Wayne Yu for implementing support for the MREF statement.

References

