# A Continuum between Serendipitous Browsing and Query-based Search for Multimedia Information Access

Julien Ah-Pine, Jean-Michel Renders, Marie-Luce Viaud

# A Continuum between Serendipitous Browsing and Query-based Search for Multimedia Information Access

Julien Ah-Pine[1], Jean-Michel Renders[1] and Marie-Luce Viaud[2]

[1] Xerox Research Centre Europe
6 chemin de Maupertuis, 38240 Meylan, France
firstname.lastname@xrce.xerox.com
[2] Institut National de l'Audiovisuel
4 avenue de l'Europe, 94366 Bry-sur-Marne Cedex, France
mlviaud@ina.fr

**Abstract.** This paper deals with information seeking in a multimedia database. In such a context, we assume that the search process is complex, dynamic and multi-faceted. We suppose moreover, that the information need with respect to a topic, can change during a same search session, going from a simple lookup search to a thorough discovery of connected subtopics. We propose a system that aims at addressing these challenges. It couples serendipitous browsing and query-based search in a smooth manner. The main features of our proposal that we want to underline are the following ones. The system offers two levels, global and local, of visualizing the context of the information seeking task and it also allows to view and search the data using either monomodal or cross-modal similarities. Furthermore, the system integrates a new relevance feedback model that takes into account the multimodal nature of the data in a flexible way and a combination of two parameters, the locality and forgetting factors, that allows to design adaptive metrics in the interactive search process. These functionalities are aimed at offering tools to the user in order to solve information seeking tasks efficiently. A preliminary user-centered evaluation of our system, that we also present in this paper, gives encouraging results.

## 1 Introduction

In the Information Seeking field (see for example [1]), we can distinguish different strategies for accessing and exploring multimedia databases, in order to acquire and discover knowledge to solve some user tasks. One strategy is *browsing and navigation:* the aim is to browse a large digital library in order to have a general overview of the different themes and the underlying structure using a tool that groups together similar objects and visualize the similarity relations between them; the user can then explore these clusters, by zooming into particular areas, visiting specific documents and jumping to their neighbors. Another strategy is *query-based search:* the aim is to quickly find relevant objects with respect to a

given query using a tool that takes into account the user feedback to bridge the semantic gap between the user's query and the multimedia objects. In this use case, the key features rely on avoiding redundancy and visualizing the similarity relationships between the retrieved objects so that the user can have a more rapid understanding of the different subtopics. But a more general "mixed-strategy" scenario happens when the user wants to have a mix between serendipitous search and query-based search. Indeed, it is often very hard for the user to formulate an unambiguous query, which is the direct translation of her information needs. It also happens that the user does not know exactly what she is looking for: she has a general question in mind, but absolutely no idea of the answer and in which direction she has to search. The ideal process the user wants to be involved in is a discovery process, where she could incrementally precise her requirements depending on what the system is able to propose her interactively, where she could understand the direction she is currently investigating with respect to the global picture, and where she can go back to explore new directions, being aware of the boundaries of this discovery process.

The system we propose in this paper exactly aims at addressing these complex needs, with a "mixed-strategy" approach. It offers some continuum between the browsing behavior and the query-based search behavior. In this paper, we focus on digital libraries that contain multimedia objects that are constituted of texts and/or images, even if most of the proposed methods and tools could be extended to other modes (speech, music, ...), once adequate monomodal similarity measures have been defined. The context of this paper is also related to information fusion in multimedia information processing and, especially, cross-media techniques that can combine visual and textual aspects in order to bridge the gap between these two modes when exploring, exploiting and searching in databases of hybrid objects.

The rest of this paper is organized as follows. In section 2, we give an overview of the global architecture and the main novelties of our system and we detail each component. In section 3, we present some results of a preliminary user-centered evaluation based on the Cognitive Walkthrough method. Then, in section 4, we analyze some related works before concluding in section 5.

## 2 Description of the System

### 2.1 Global Architecture and main functionalities of the system

The architecture of the system we propose is depicted in Fig. 1. It consists of several interlinked components: the Graphical User Interface, the monomodal Search Engines, the Ranker/Scorer and the Graph Layout Map Builders. These components combined together allow to achieve the following functionalities that we think, are valuable for multimedia information seeking tasks:

• **interlinked multi-scale visualization and navigation:** the system offers (at least) two levels of visualizing the context of the information seeking task. One visualization is the 2D *global map* of the whole multimedia corpus, emphasizing the underlying structure of this corpus The structure is typically charac-

terized by different clusters and sub-clusters, with mutual positions indicating how these clusters relate to each other. The second map, called *the local map*, synthesises the history of the current session, by representing all objects the user has to interact with, on a single 2D map that again emphasizes the underlying structure of this set of objects (clusters), as well as their mutual similarity relationships. These two maps are processed by the Graph-Layout Map Builders component that we detail in subsection 2.4. Objects of the local map are linked to their counterpart in the global map; the latter clearly indicates what are the areas that are represented on the local map. Having at least two maps makes it easier for the user to be aware of the boundaries of her search, to understand the different landscapes at different scales, and to better control the exploration (global visualization) and exploitation (local development) phases.

• **multimodal views of the data:** on the global map, the user can have different views of the data: she can switch to purely textual, purely visual or hybrid similarities, so that there are actually 3 static maps that co-exist. This allows the user to change the global map according to the modality she is interested in at each step of the session.

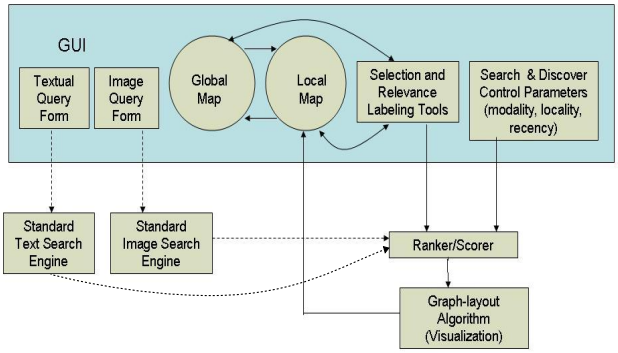
• **flexible multimodal relevance feedback:** on the local map, the user can label the text and the image part of a same multimedia object differently. In that way, she can associate relevant texts with relevant images that best correspond to her current multimodal information need.

• **adaptable search/development metrics:** our relevance feedback technique implemented in the Ranker/Scorer component described in subsection 2.3), contain several parameters that allow the user to design metrics that adapt to her information need at each iteration. First, after giving feedback, the user will typically promote different kinds of similarity for the next step: her search can rely on textual, visual or hybrid similarities. The Graph-Layout Map Builders component takes this into account so that on the local map, the metrics that define local neighbors around some nodes could be visual, while being textual or hybrid around other nodes. Next, as the system aims at providing the user with a continuum between a browse-based search and a query-based search, it allows the user to tune a suitable combination of a locality factor and a forgetting factor, that will weight accordingly all the accumulated information[3] in the current session. More particularly, locality allows some selected objects to have more weight than others, in order to "develop" the similarity graphs locally and to give a new direction to the search. The forgetting factor assumes that the user is naturally more prone to give more importance to what she interacts with recently, rather than what she developed initially[4]. On the map, objects impacted by the forgetting factor are indicated by their fading level or by their decreasing size.

[3] Mainly the initial query, if it exists, and all proposed objects that the user labeled as relevant or not relevant

[4] The Ostensive Model introduced in [2] makes the same assumption. See section 4 for further details.

**Fig. 1.** Architecture of the proposed system




Before detailing each component, let us introduce the following definition: a *session* is a sequence of interactions between a user and the system, that corresponds to the same information need or task[5]. These interactions include visualizations and proposals from the system side, query formulation and/or object selection and/or relevance feedback from the user side.

### 2.2 The Graphical User Interfaces

The GUIs performing the global and local maps contain 3 parts (see Fig. 2 and Fig. 3). The central part represents the maps, the left part shows some detailed view of the data: when the mouse is located on an item, the associated text and image are displayed. However, the right part of the interface is different for both maps:

• For the global map, the right part includes two standard query interfaces (one for textual query such as depicted in Fig. 2, one for image query) that are typically used at the beginning of the session, in order to generate an interesting subset of objects for further developments as the "page zero" of the local map. The search engines are standard ones, typically returning the $k$ nearest neighbors of a given query (the latter objects are then highlighted on the global map). Note that the use of query forms is optional[6] as the user can simply select one or more objects of the global map to "develop" them in the local map.

• For the local map, the right part is dedicated to the parameter settings of the adaptable search/development metrics (feedback, modality, locality and forgetting factors) as it is shown in Fig. 3. Labeling of the retrieved elements (the text and/or the image of elements more precisely) is done by selecting the corresponding items with the mouse and by clicking on the "+" or "-" button of the right panel (see Fig. 3). "+" and "-" respectively represent relevant and irrelevant labels. The items which are neither annotated "+" nor "-" are considered

[5] A task is constituted of two subtasks, a closed problem and an open problem. See subsection 3.2 for more details.

[6] Dotted arrows on the diagram indicates that the textual and visual search engines are optional and that they are typically used only for iteration 0.

as neutral. They have a null weight and they remain displayed on the local map with a grey color. We assume that they can be of interest for the user but not immediately. Hence, the user can label those neutral objects positively or negatively later on in a session. Then, the type of modality is a value to set among textual, visual or hybrid mode and the locality and forgetting factors are real numbers to set between 0 and 1.

The user can zoom in/out or move the map panels with the mouse roll. To launch a local map, the user selects one or several objects of the global map by clicking on them with the mouse and activates the adequate menu item. A new window appears and the local discovery/search process can start. On the local map, chosen positive items are put in red whereas negative items are first put in green but finally disappeared at the next iteration. Once the items have been labeled, clicking on the "submit" button launches the retrieval process with feedback. Results appear instantaneously on the local map and the corresponding items are highlighted on the global map.

### 2.3 The Ranker/Scorer Component

The Ranker/Scorer component is the core of the system: it generates at each iteration a ranked list of objects, that are considered to have the largest probability of being relevant, given the information accumulated[7] up to that moment and the different search/discovery parameters[8] of the current iteration. This component has also to deal with the issue of fusing the textual and visual modes, when needed; in what we propose, this could be partly realized by defining a cross-media similarity measure based on a mix of real and pseudo-relevance feedback[9].

We propose the formula given in eq. 1 for computing a new relevance score for each (unlabeled) object, $x$, of the database based on the accumulated feedback information and the control parameters chosen at the current iteration $t$. It can be seen as a non-trivial extension of Rocchio's method [6] to the more general case of interactive multimedia information seeking.

$$f^{t+1}(x) = \tag{1}$$
$$\gamma_T^t \left[ \sum_{y \in \mathcal{T}_+^t} \frac{\alpha_T^t(y)}{\sum_{y' \in \mathcal{T}_+^t} \alpha_T^t(y')} \left( S_T(y,x) + \lambda_T \frac{\sum_{z \in \mathcal{B}_T^t(y)} S_T(y,z) S_I(z,x)}{\sum_{z' \in \mathcal{B}_T^t(y)} S_T(y,z')} \right) \right.$$
$$\left. - \sum_{y \in \mathcal{T}_-^t} \frac{\beta_T^t(y)}{\sum_{y' \in \mathcal{T}_-^t} \beta_T^t(y')} \left( S_T(y,x) + \delta_T \frac{\sum_{z \in \mathcal{B}_T^t(y)} S_T(y,z) S_I(z,x)}{\sum_{z' \in \mathcal{B}_T^t(y)} S_T(y,z')} \right) \right]$$

[7] Relevance feedback labels and potential initial query.

[8] Modality(ies) selected for searching, forgetting factor value, locality factor value

[9] Our proposal is an interactive extension of the trans-media pseudo-relevance feedback introduced in [3–5] for the non-interactive case.

$$+\gamma_I^t \left[ \sum_{y \in \mathcal{I}_+^t} \frac{\alpha_I^t(y)}{\sum_{y' \in \mathcal{I}_+^t} \alpha_I^t(y')} \left( S_I(y,x) + \lambda_I \frac{\sum_{z \in \mathcal{B}_I^t(y)} S_I(y,z) S_T(z,x)}{\sum_{z' \in \mathcal{B}_I^t(y)} S_I(y,z')} \right) \right.$$
$$\left. - \sum_{y \in \mathcal{I}_-^t} \frac{\beta_I^t(y)}{\sum_{y' \in \mathcal{I}_-^t} \beta_I^t(y')} \left( S_I(y,x) + \delta_I \frac{\sum_{z \in \mathcal{B}_I^t(y)} S_I(y,z) S_T(z,x)}{\sum_{z' \in \mathcal{B}_I^t(y)} S_I(y,z')} \right) \right]$$

In eq. (1), $f^{t+1}(x)$ is the new relevance score of the (unlabeled) multimedia object $x$ provided at iteration $t+1$. The subscripts $T$ and $I$ respectively correspond to text and image modality. $S_T$ and $S_I$ are then the textual and visual similarity matrices.

Let denote $mod^t$ the search modality(ies) chosen by the user at iteration $t$. $mod^t$ can take the value $T$ or $I$ or $H$ (hybrid: $T$ and $I$). In the sequel, we introduce the notations with respect to the text modality only. However, since text and image play symmetric role, one can deduce the corresponding definition for the image part by simply replacing the subscript $T$ with $I$, the set notation $\mathcal{T}$ with $\mathcal{I}$ and "text" with "image" in the text (and vice-versa).

$\gamma_T^t$ reflects the weight given by the user to the text modality at iteration $t$. More precisely, we have $\gamma_T^t = 0$ if $mod^t = I$ and $c_T$[10] otherwise.

$\mathcal{T}_+^t$ is the set of objects whose textual part was labeled as relevant by the user up to step $t$. On the contrary, $\mathcal{T}_-^t$ is the set of texts that were labeled as irrelevant up to iteration $t$.

$\alpha_T^t$ and $\beta_T^t$ are weights that give the importance of texts within $\mathcal{T}_+^t$ and $\mathcal{T}_-^t$ in order to compute the new relevance scores vector $f^{t+1}$. These weights take into account different parameters. First, the user can select a special subset of the items annotated at the current step $t$. These selected texts, $\mathcal{S}_T^t$, correspond to the text part of the nodes of the graph around which the system should develop new elements. In comparison to other labeled items, the selected objects are given an extra weight $loc_T^t \in [0,1[$ specified by the user. The greater the locality value, the more the user wants to focus on the newly selected objects. Second, the user can also explicitly mention to the system what is the importance to be given to previously annotated items. This is the forgetting factor $forg_T^t \in [0,1]$. With such a factor, we assume that the weight of an annotated item will decrease with time. Thus, the older the labeling of an object (its text or image part), the lower its weight. This effect is even stricter as the forgetting factor increases. The "recentness" of the labeling is something not so trivial. Let assume that, at the current iteration $t$, the user decides to go back to the results provided at iteration $t' < t$ and select some of these items. This might mean that the user wants to pursue another direction in her information seeking. Therefore, we assume that the objects that were annotated from step $t'+1$ up to $t-1$ are not important anymore. Hence, we give a null weight to these items[11]. More

[10] $c_T$ being a strictly positive pre-defined constant.

[11] This case corresponds to the third case in eq. (3).

formally, we compute the weight vectors for the annotated objects as follows[12]. Note that by default, we take $\forall y \in \mathcal{T}^t = \mathcal{T}^t_+ \cup \mathcal{T}^t_- : \beta^t_T(y) = \alpha^t_T(y)$, since this setting works better according to some preliminary experiments. $\forall y \in \mathcal{T}^t_+$, we have the following definition:

$$\alpha^t_T(y) = \begin{cases} \frac{1}{1-loc^t} & \text{if } y \in \mathcal{S}^t_T \\ (1 - forg^t_T)^{m_T(y)} & \text{if } y \notin \mathcal{S}^t_T \text{ and } m_T(y) \geq 0 \\ 0 & \text{if } y \notin \mathcal{S}^t_T \text{ and } m_T(y) = -1 \end{cases} \quad (2)$$

where:

$$m_T(y) = \begin{cases} t - date_T(y) & \text{if } \mathcal{S}^t_T = \emptyset \\ \min\limits_{z \in \mathcal{D}^t_T(y)} (date_T(z) - date_T(y)) & \text{if } \mathcal{S}^t_T \neq \emptyset \text{ and } \mathcal{D}^t_T(y) \neq \emptyset \\ -1 & \text{if } \mathcal{S}^t_T \neq \emptyset \text{ and } \mathcal{D}^t_T(y) = \emptyset \end{cases} \quad (3)$$

where $date_T(y)$ is the iteration number when the text of object $y$ was annotated and $\mathcal{D}^t_T(y) = \{z \in \mathcal{S}^t_T : date_T(z) \geq date_T(y)\}$. In other words, given $y \in \mathcal{T}^t$, $\mathcal{D}^t_T(y)$ is the set of selected texts $z$ that were annotated after $y$.
Notice that a locality factor $loc^t_T$ equal to 0 amounts to give no extra weight to selected objects[13]. On the contrary, a locality factor very close to 1 will result in discarding non-selected items. Indeed, when this factor tends to 1, the non-null contributions in the different terms of eq. (1) come only from the selected objects, due to the weighted average effect.

With respect to eq. (1), positive and negative text pseudo-relevance feedback are respectively introduced through the terms weighted by $\lambda_T$ and $\delta_T$. To be more precise, it is a trans-media pseudo-relevance feedback which considers as relevant the visual part of texts that are very similar to the texts fed back as relevant by the user; but this feedback mechanism discounts a pseudo-relevant object by the factor $\lambda_T$ and by the specific (normalized) textual similarity between the pseudo-relevant text and the corresponding labeled texts whose it is the neighbor. Accordingly, we denote $\mathcal{B}^t_T(y)$, the set of texts that haven't been annotated yet and which are the nearest neighbors of (the text part of) $y$. Similarly, the system considers as irrelevant the visual part of texts that are very similar to the texts fed back as non-relevant by the user. Likewise, this dual negative view of the pseudo-feedback mechanism consists of the terms weighted by the discount factor $\delta_T$ in eq. (1). To be consistent, neighbors of labeled objects that are themselves labeled are never considered as pseudo-relevant objects.

### 2.4 The Graph-Layout Map Builders

This component is the one that produces as outputs the different maps for visualizing globally or locally the objects of interest.

Global maps are computed off-line. We first apply a sequence of several force directed layout algorithms to generate the maps, we then use the LinLog energy

[12] As mentioned beforehand, we define the weights for the text modality only but the ones corresponding to the image modality are similar.
[13] In this case, it does not make sense to select any object.

model [7] as the final stage. The basic material consists of thresholded similarity matrices[14]. A standard agglomerative hierarchical clustering algorithm is then applied to identify clusters in the 2D space. Cluster naming techniques allow then to extract the most representative keywords of each cluster.

The local map layout is a dynamic process: results are appended to the map at each "interactive query" performed by the user. Regarding dynamic representations, one additional constraint has been established by the visualization community: the problem of preserving the user's mental map [8]. The objective is not to loose the user by constantly changing the map layout from one iteration to the next one: new objects are added by slightly perturbing the previous layout and using the similarity metrics promoted by the user at the current iteration, while already present objects keep their mutual similarity relations, as a result of all previous interactions. This is realized by increasing the inertia of existing nodes and by using the Fruchterman-Rheingold layout algorithm [9], that appears to be the most adequate for this kind of task. Optionally, a clustering algorithm could be applied as well in the 2D local map, in order to avoid redundancy in the results given at the next iteration and to favor quick local exploration: only the most relevant objects of each cluster will be displayed on the map (see for example [4]). This could be considered as an indirect way of realizing diversity-based re-ranking and can be particularly valuable during the early stages of the search process.

## 3 Preliminary Evaluation of the System

### 3.1 Evaluation Methodology

It is difficult to assess an interactive information seeking system. Indeed, as the system is mainly designed to improve the efficiency of user feedback, traditional retrieval measures such as precision and recall are not really interesting here; other measures based on satisfaction and increase in work efficiency to solve the task are more important. Previous works already conducted user-centered evaluations in similar contexts [10]. It has been shown that interactive methods are well-appreciated by regular user of information retrieval systems. In the following study, our goal is to have a preliminary user-based evaluation of our main contributions in multimedia information seeking namely, the interlinked multiscale and multimodal visualization and navigation and the flexible multimodal searching and relevance feedback.

To this end, we chose the Cognitive Walkthrough Inspection methodology [11, 12] to perform the evaluation. This method involves an experimented user and an evaluator and allows to observe the user behavior during the realization of a complex task. The evaluator has pre-specified scenarios and tasks to be realized, while the user[15] is already familiar with similar tasks and already using standard

[14] With a user-adjustable threshold.

[15] In our case, she is an expert archivist who is used to seek information in a multimedia database with classical systems.

retrieval tools and graphical user interfaces. The procedure is the following one: the evaluator explains the goals of the task and the functionalities associated to the different sub-components of the GUI (maps, buttons, parameters,...). The user has then to choose the sequence of elementary actions in order to solve the task. But, during the evaluation, the evaluator manipulates himself the tool and asks for the functional action to be executed at the next step: the reason for this is to get rid of specific ergonomic aspects (design of buttons, choice of colors,...) and to solely focus on addressing the following points: **achievability** (is the set of elementary functions sufficient to solve the task?), **efficiency** (does the system promote the most efficient paths?), **predictability** (is the user able to predict the effect of the launched action?), **obviousness** (how intuitive is it?), **proactivity** (after the action, is the feedback good enough to encourage her to continue?) and, finally, **confidence** (is the user more confident about the obtained results?). What is eventually measured is *(i)* that the user is able to understand the link between the sequence of actions and the final goal of the task, and *(ii)* that she is able to memorize the corresponding actions and settings. At the end of the evaluation, the user is asked extra questions, related to the comparison with her existing tools, in terms of complementary or new possibilities.

### 3.2 Design of the Evaluation Scenarios

The multimedia collection used for the experimentation consists of text/image objects extracted from a subset[16] of pages from the French Wikipedia that are related to the general theme "Tourism in France". From each selected Wikipedia page, we extracted several multimodal objects, namely the images present in the page with their associated texts (image caption if any, text of the surrounding paragraph and sequence of titles and subtitles leading to this paragraph). Note that, due to our construction mechanism, the relationship between an image and its associated text could be noisy or very vague. This collection[17] is made of more than 50,000 text/image objects.

A task consists in solving two subtasks related to a same topic: a specific search (closed problem) and a discovery analysis (open problem). The specific search subtasks were designed in such a way that it is very unlikely to obtain the information directly by a single textual query and that combining both modalities in a flexible way is essential. The different topics presented to the user were related to: *Eiffel Tower, surfing, old stamps, Charles de Gaulle* and *Nantes*.

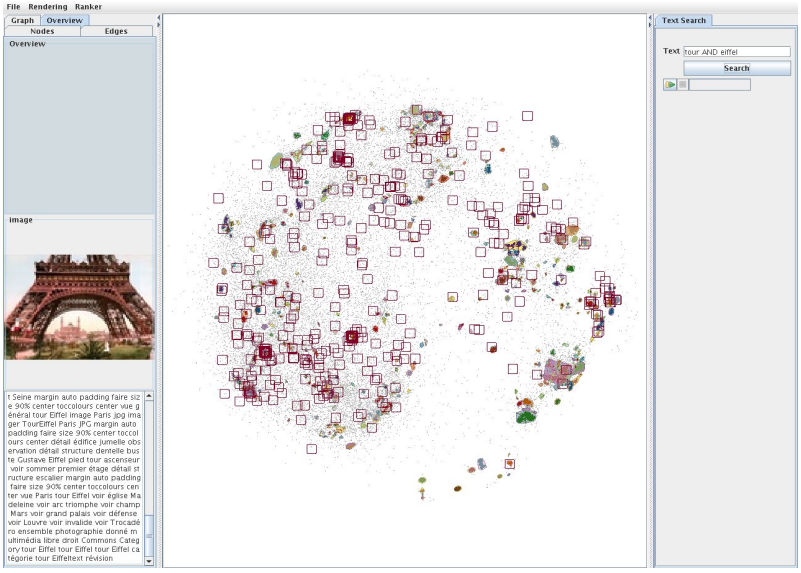
### 3.3 Description of a Particular Evaluation Scenario

In this subsection, we report some retrieval results for the *Eiffel Tower* scenario. For this topic, the user had to retrieve old pictures representing the Eiffel Tower

[16] The selection is based upon the categories of Wikipedia.

[17] This collection was constructed for the purpose of the *Infom@gic* project. See the acknowledgments section.

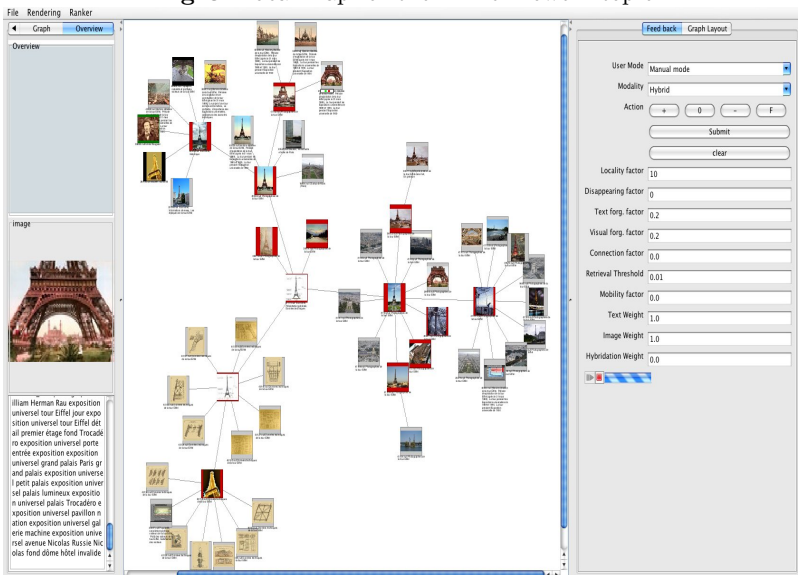
**Fig. 2.** Global map (with objects relevant to "Eiffel Tower" highlighted)




and dating from the beginning of the 20th century (closed subtask). She also had to explore the collection in order to gather different multimedia objects that cover all potential subtopics related to the Eiffel Tower, as if she wanted to find as much material as possible to make a multimedia presentation on that topic (open subtask). The evaluator let the user free to solve these subtasks sequentially or in parallel, but the user actually found it more efficient to solve them simultaneously.

The user started with a general textual query "Eiffel Tower" using the basic text search engine, whose results were highlighted on the global map (see Fig. 2). She observed that a lot of items were surrounded and their distribution spread all over the global map. After a quick observation, the user mentioned that many of the highlighted objects were not relevant for the specific task. The reason is that the Eiffel Tower is often used as a generic French emblem. After zooming in some particular areas presenting a high density of highlighted results, the user picked an object whose image represents the Eiffel Tower, even if not completely relevant to the search subtask (drawing instead of picture). The latter is the black and white drawing of the Eiffel Tower in the center of Fig. 3. From this chosen element, the user started a local deployment with the "hybrid modality". The user asked to set the forgetting factor to 0.2. After 8 iterations during which 12 objects were labeled relevant and 20 irrelevant, the user obtained the results presented in Fig. 3. During this sequence, the user used the locality feature that allows to deeper focus on a set of selected objects. To be more precise, this focus was deployed on the bottom left corner of Fig. 3 which shows technical drawings of the Eiffel Tower. This resulted in a first set of objects, relevant to the second subtask but not to the first one. Next, the feedback provided by the

**Fig. 3.** Local map for the "Eiffel Tower" topic



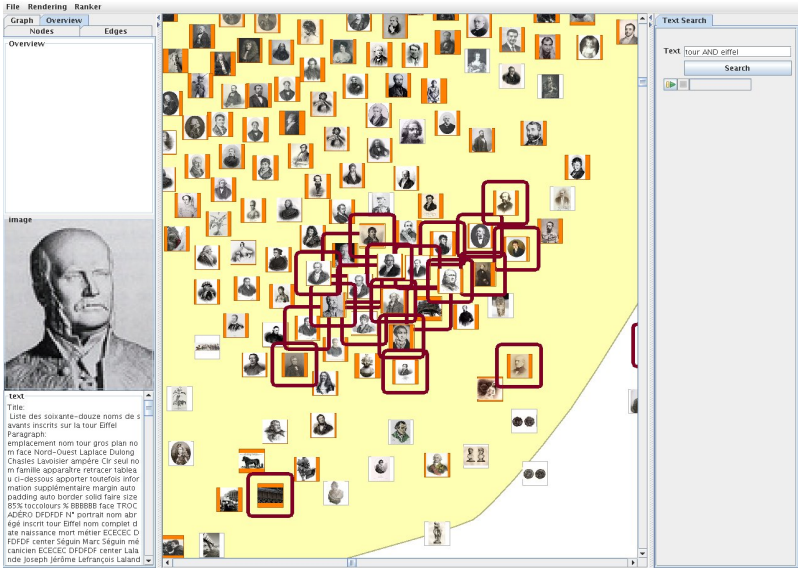

user gave rise to the right branch which shows landscape pictures of Paris with the Eiffel Tower; once again, this set of objects was considered as relevant to the second subtask, but not to the first one. But the branch deployed on the upper left part of the map appeared to be more related to some historical events with a certain relationship with the Eiffel Tower: postcards from the 1900 universal exhibition, portrait of an engineer, etc. Among the objects that constituted this branch, the user finally found what she was looking for regarding the first subtask: two pictures of the monument with a blue and white sky on the upper left, which correspond to pictures of the Eiffel Tower dating from the beginning of the 20th century. The user noticed here that the effect of the forgetting factor was effective since the strong visual contribution of the first drawing (the item around which the user originally chose for local deployment) was progressively lowered iteration after iteration. To conclude the second subtask, the user further analyzed the global map and particularly areas presenting a strong density of results for the textual query "Eiffel Tower"; this resulted in the discovery of 3 other interesting clusters, that are relevant to the task. Fig. 4 shows an example of cluster displayed on the global map not covered by the local search: it displays portraits of famous scientists whose names are written on the Eiffel Tower.

### 3.4 First Conclusions drawn out of the User Evaluation

Based on the user's reactions collected during the evaluation (including global comments at the end), we address the different points raised in subsection 3.1 according to the Cognitive Walkthrough Inspection methodology:

• Achievability and predictability: in general, the user succeeded in finding satis-

**Fig. 4.** Zoom on the Global map (with objects relevant to "Eiffel Tower" highlighted)




fying results for the search and discovery parts of each task. She had no trouble to perform the list of actions needed to obtain the results. However, some local deployments were very noisy. This is often due to the wrong association between text and image in the multimedia collection[18] that can be encountered in some cases[19]. In those cases, the user would have liked to mark a positive feedback only on a selected portion of the text rather than on the whole text. Besides, the user's comments on the parameter settings, are balanced. The control on the local deployment is rich, innovative and interesting but more training is needed to really understand all its possibilities. In fact, it is not obvious to anticipate the results when using the forgetting factor, especially when both modalities are involved with different values.

• Efficiency, obviousness and proactivity: regarding the use and the selection of different modalities, it appeared that this new feature was very useful for the search in general and for cases where the image is not well-described by its associated text in particular. The use of textual queries for generating a "page zero" and initiating a search process even with objects without any text associated is particularly valuable. With respect to the use of two linked maps, the user exploited both of them for all tasks. The connection between the local and global maps during the search process turned out to be very intuitive and helpful: this feature allows to have a better perception of the content of the corpus.

[18] This is a side effect of the way we pre-processed the French Wikipedia corpus.

[19] In the case of the "Eiffel Tower" topic for instance, objects corresponding to other monuments appeared because there is a Wikipedia page that lists the most visited monuments in Europe, so that the same text is associated to very different images.

Regarding the usability, the user was globally comfortable with using the maps. Navigating in the global map results appeared easy, especially with the display of the text and image in the left part of the panel just by moving the mouse on the items. Selection and launch of the local map was also easy. But, the local map parameters' setting was not obvious. The selection and labeling of the texts and/or images were all right. The use of the focus mode was intuitive.

We can formulate the following preliminary conclusions from this evaluation:
• Using different modalities and particularly cross-media techniques that allow to combine visual and textual information efficiently is necessary. It allows to provide faster ways to achieve relevant results particularly when the information need is difficult to express in terms of queries and when the different modalities of the same object do not match from a semantic viewpoint[20]. In that perspective, the possibility to change from one modality to the other one for visualizing and searching is interesting.
• Using one global map and one local map jointly allows the user to better control the exploitation/exploration trade-off. The local map allows the user to express her information need more precisely while the global map allows her to better understand the different boundaries of her search and discover non-expected subtopics.
• While using the local map, the user can progressively express her information need by selecting relevant texts and/or images and discarding negative examples, in a flexible manner. As far as our multimodal feedback technique is concerned, this flexibility provides efficient ways to achieve interesting results since the user is really free to associate relevant texts with relevant images that best correspond to her multimodal information need.
• The use of extra features that are integrated in our feedback model such as forgetting and locality factors are encouraged though we should not loose the user by asking him to tune a lot of parameters. The locality factor clearly allows to have a continuum between browse-based and query-based search since the user can discover many subtopics related to a broader topic and focus on some of them at any time of a session. The forgetting factor also allows to achieve this continuum as it models the fact that the user is more aware of her last annotations rather than her first ones. Furthermore, the forgetting factor allows to decrease the importance of texts and images with different rates; this feature appeared to be particularly interesting in our evaluation scenarios.

## 4 Related Work

The literature covering interactive multimedia retrieval is very vast. However, our proposal particularly concerns information seeking in a text/image corpus. In that context, the paper [13] presents a relevance feedback approach which integrates semantic (keywords) and low-level feature in the context of image retrieval. Their method is an extension of the Rocchio technique [6] which relies on a semantic network derived from the keywords associated to the images.

[20] For example, images that have poor or noisy textual descriptions.

However, their system only targets basic query-based search and the semantic network is updated using the feedback provided by the user, with no possibility to judge independently texts and images. Focusing on text/image collections, there have been many works in the context of ImageCLEFPhoto evaluation campaigns in multimodal retrieval. Particularly, the paper [14] tackles such a task in the case of interactive search. Their combination method is based on a hierarchical late fusion approach which is different from our technique [3]. Systems addressing video retrieval are also related to our proposal. Particularly, the work presented in [15] shows several common aspects with our work. The authors use a multimodal similarity (or dissimilarity) space for representing the multimedia objects; they then propose to apply a one-class SVM in order to learn a classifier that separates relevant from non-relevant examples.

Concerning the visualization part of the system, a state of the art of visualization methods and tools developed for multimedia information is given in [16]. Some systems propose a multi-scale view of objects for browsing and interactively searching within a multimedia corpus. The most closely related work to our proposal is the following one [17]. The projection methods employed are distinct. In [17], they use non-linear embedding algorithms whereas we rather use graph-layout methods. Then, the main difference between the two systems is that we propose not only a multi-scale but also a multimodal view of the data.

Finally our proposal is also very close to the Ostensive Model introduced initially in [2]. This approach considers the information retrieval process as dynamic and it is a relevance feedback model that integrates a temporal notion to relevance. In fact, the forgetting factor that we have introduced acts exactly the same as the so-called ostensive relevance. More specifically, our system shares many common points with the following work [10], which addresses content-based image retrieval from a multimodal perspective and using the Ostensive Model. However, some important differences with our approach are the following ones: first, in our system there are two interlinked multi-scale maps whereas in [10] only one map is employed; second, the multimodal nature of the data is not emphasized in the feedback and the search processes in [10]; third, our system explicitly allows the user to annotate many candidates at each step whereas in [10] the user is asked to select only one relevant candidate; last, the combination of textual and visual information are different in both systems.

## 5 Conclusion

In this paper, we have introduced the architecture and the key components of a system for accessing information in a multimedia digital library. The general aim of our proposal is to design a system that offers some continuum between serendipitous browsing and query-based search. We have detailed the key features and algorithms of our proposal, namely the multi-scale and multimodal navigation and the adaptive multimodal relevance feedback technique. The next steps of this work will consist in better evaluating the performances of our system from the user viewpoint since the study we have presented is a preliminary,

yet encouraging, evaluation. Then, we would like to further investigate the joint use of multi-scale maps in information seeking tasks.

**Acknowledgments.** This work was funded by the *Infom@gic* project, part of the "Pôle de compétitivité Cap Digital de Paris".

## References

1. Marchionini, G.: Exploring search: from finding to understanding. Communications of the ACM **49** (2006) 41–46
2. Campbell, I., Van Rijsbergen, C.: The ostensive model of developing information needs. In: Proc. of CoLIS 2. (1996) 251–268
3. Clinchant, S., Renders, J.M., Csurka, G.: XRCE's participation to ImageCLEF 2007. In: Working Notes of CLEF'07 Workshop. (2007)
4. Ah-Pine, J., Cifarelli, C., Clinchant, S., Csurka, G., Renders, J.: XRCE's participation to ImageCLEF 2008. In: Working Notes of CLEF'08 Workshop. (2008)
5. Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.M.: Crossing textual and visual content in different application scenarios. Multimedia Tools Appl. **42**(1) (2009) 31–56
6. Rocchio, J.: Relevance feedback in information retrieval. In: The SMART Retrieval System. (1971) 313–323
7. Noack, A.: Visual clustering of graphs with nonuniform degrees. In: Proc. of Int. Symp. on Graph Drawing (GD'05), Springer-Verlag (2005) 309–320
8. Misue, K., Eades, P., Lai, W., Sugiyama, K.: Layout adjustment and the mental map. Journal of Visual Languages & Computing **6** (1995) 183–210
9. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Softw., Pract. Exper. **21** (1991) 1129–1164
10. Urban, J., J., J.M., van Rijsbergen, C.J.: An adaptive technique for content-based image retrieval. Multimedia Tools Appl. **31**(1) (2006) 1–28
11. Huart, J., Kolski, C., Sagar, M.: Evaluation of multimedia applications using inspection methods: the cognitive walkthrough case. Interacting with Computers **16**(2) (2004)
12. Polson, P.G., Lewis, C., Rieman, J., Wharton, C.: Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. Int. J. Man-Mach. Stud. **36**(5) (May 1992) 741–773
13. Lu, Y., Zhang, H., Wenyin, L.: Joint semantics and feature based image retrieval using relevance feedback. IEEE Transactions on Multimedia **5** (2003) 339–347
14. Rahman, M., Desai, B., Bhattacharya, P.: Multi-modal interactive approach to imageCLEF 2007 photographic and medical retrieval tasks by CINDI. In: Working Notes of CLEF'07 Workshop. (2007)
15. Bruno, E., Moenne-Loccoz, N., Marchand-Maillet, S.: Design of multimodal dissimilarity spaces for retrieval of video documents. IEEE Trans. Pattern Anal. Mach. Intell. **30** (2008) 1520–1533
16. Goëau, H., Thièvre, J., Verroust-Blondet, A., Viaud, M.L.: State of the art on advanced visualisation methods (2007) Report D7.2 of the Vitalas EC project FP6 - 045389.
17. Nguyen, G.P., Worring, M.: Interactive access to large image collections using similarity-based visualization. J. Vis. Lang. Comput. **19**(2) (2008) 203–224