Crossing textual and visual content in different application scenarios

Julien Ah-Pine, Marco Bressan, Stephane Clinchant, Gabriela Csurka, Yves Hoppenot, Jean-Michel Renders

Noname manuscript No.
(will be inserted by the editor)

Crossing textual and visual content in different application scenarios

Julien Ah-Pine · Marco Bressan · Stephane Clinchant · Gabriela Csurka · Yves Hoppenot · Jean-Michel Renders

Received: date / Accepted: date

Abstract This paper deals with multimedia information access. We propose two new approaches for hybrid text-image information processing that can be straightforwardly generalized to the more general multimodal scenario. Both approaches fall in the trans-media pseudo-relevance feedback category. Our first method proposes using a mixture model of the aggregate components, considering them as a single relevance concept. In our second approach, we define trans-media similarities as an aggregation of monomodal similarities between the elements of the aggregate and the new multimodal object. We also introduce the monomodal similarity measures for text and images that serve as basic components for both proposed trans-media similarities. We show how one can frame a large variety of problem in order to address them with the proposed techniques: image annotation or captioning, text illustration and multimedia retrieval and clustering. Finally, we present how these methods can be integrated in two applications: a travel blog assistant system and a tool for browsing the Wikipedia taking into account the multimedia nature of its content.

Keywords First keyword · Second keyword · More

1 Introduction

Information, especially digital information, is no longer monomodal: web pages can have text, images, animations, sound and video; we have audiobooks, photoblogs and videocasts; the valuable content within a photo sharing site can be found in tags and comments as much as in the actual visual content. Nowadays, it is difficult to visit a page within a popular social network without finding a large variety of content modes surrounded by a rich structure of social information such as profiles, interest groups, consumer behaviour or simple conversations. This major shift in the way we

Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, France
Tel.: +33-76615050
Fax: +33-76615099
E-mail: FirstName.lastName@xrce.xerox.com

access content and what type of content we access is largely due to the connected, easily accessible, global nature of the internet. The democratization of the tools of production and delivery have also strongly contributed to this situation, e.g. low cost camera-phones combined with accessible publishing tools. This scenario poses a strong need for tools that enable interaction with multimodal information.

Multimodal information processing strongly needs trans-media similarity measures that can effectively bridge the gap between the co-existing modes. Such tools will enable applications whose relevance can already be perceived today: accessing an image repository from a textual query, automatic illustration, content creation and document generation, captioning, multimodal summarization, hybrid categorization and indexing, content personalization, etc. In this article, we explore such measures and propose two new approaches for hybrid text-image information processing, an important particular case of the more general multimodal scenario.

Within multimodal information processing an important family of approaches is constituted by the so-called “trans-media” pseudo-relevance feedback approaches (section 3). The basic idea is to first use one of the media types to gather relevant multimedia information and then, in a second step, use the dual type to perform the final task (retrieve, annotate, etc). These approaches can be seen as an “intermediate level” fusion as the media fusion takes place after a first mono-media retrieval step based on mono-modal similarities (see 2.1 and 2.2).

We describe two methods, both falling in the “trans-media” pseudo-relevance feedback category. In our proposed approaches, similarly to [21], we need to define similarities between an aggregate of objects and a new multimodal object. In our first approach (section 3.1), we build a model inspired by the mixture model [38] of the aggregate. In our second approach (section 3.2), we define trans-media similarities as an aggregation of mono-modal similarities between the elements of the aggregate and the new multimodal object.

We first provide an overview of the state-of-the-art in hybrid text and image information processing, emphasizing work on trans-media relevance feedback (section 1.1). Then we introduce our monomodal similarity measures for images (section 2.1) and for text (section 2.2). We then focus on our two main contributions (section 3). An important part of our article is dedicated to show how these techniques can be applied to a number of scenarios such as unsupervised text illustration (section 4.3.2); image annotation or captioning (section 4.3.1); ranking and retrieval (section 4.1); cross-media clustering and improved visualisation (section 4.4). Finally, we share some details of a system we developed for assisting in the creation of travel blogs (TBAS) in section 5.1. This system illustrates the way many of these techniques can interact in a real world content creation example. Similarly, we present another tool for browsing multimedia corpora, such as Wikipedia in section 5.2.

1.1 Prior Art

The scientific challenge is to understand the nature of the interaction between text and images: how can text be associated with an image (and reciprocally an illustrative image with a text)? how can we organize and access text and image repositories in a better way than naive late fusion techniques? The main difficulty is to overcome the *semantic gap* and, especially, the fact that visual and textual features are expressed at different semantic levels.

Naive techniques combine the scores from a text retrieval system and from an image retrieval system into a single relevance score: this is the late fusion approach. Departing from the classical late fusion strategy, recent approaches have considered fusion at the feature level (early fusion), estimating correspondences or joint distributions between components across the image and text modes from training data. One of the first approaches in this family is the co-occurrence Model by Mori et al. [23] where keywords are assigned to patches based on the co-occurence of clustered image features and textual keywords in a labeled training data-set. Instead of looking at the co-occurrences, Vinokourov et al [35] propose to find correlations between images and attached texts using the Kernel Canonical Correlation Analysis (KCCA). As far as the "pure" visual mode is concerned, features associated to images have become more and more complex: there were attempts of developing more expressive visual vocabularies, potentially hierarchical, relying on latent semantic extraction techniques such as Probabilistic Semantic Analysis [2,22] or Latent Dirichlet Allocation [3]. Duygulu et al experimented with machine translation models where, a lexicon links a set of discrete objects (words in annotations) onto another set of discrete objects (image regions) [10, 15].

Another group of works uses graph models to represent the structure of an image through a graph, e.g. with a Markov network [5] or a concept graph [24]. For example, Carbonetto et al [5], allow interactions between blobs through a Markov random field (MRF). Hence, the probability of an image blob being aligned to a particular word depends on the word assignments of its neighbouring blobs. Similarly, the ALIP system of Li and Wang [18] models the interaction between blocks using Markov Models. Simple block-based features are extracted from each training image at several resolutions and a two-dimensional multiresolution hidden Markov model (2D MHMM) is used to describe statistical properties of the feature vectors and their spatial dependences.

The use of pseudo-relevance feedback or any related query expansion mechanisms has been used widely in image retrieval. The basic challenge is how to control that the objects retrieved by the first retrieval step are indeed good representatives of the user needs and do not constitute noise.

The approaches proposed in this paper are mostly aligned with a family of models, inspired by Cross-Lingual Information Retrieval. In cross-lingual systems, a user generates her query in one language (e.g. English) and the system retrieves documents in another language (e.g. French). The analogy here is to consider the visual feature space as a language constitued of blobs or patches, simply called *visual words*.

Several works were proposed in this direction by Lavrenko et al [16,17,11]. They are mainly based on pseudo-feedback methods (known also as query expansions) where the initial query is automatically enriched with new words based on a first retrieval step. In [16] they extended the cross-lingual relevance models to cross-media relevance models. These models were further generalized to continuous features in [17] with non-parametric kernels, while in [11] the distribution of words was modelled with Bernoulli distributions.

Cross-media relevance models can also be considered as the ancestors of our methods [7] as well as previous methods used in ImageCLEF challenges [6], [21]. Indeed, the latter methods do not rely on relevance models but act in the same spirit. For example, from a query image, a visual similarity is first computed and an initial set of (assumed) relevant objects is retrieved. As the objects are multimodal, each image has also a text, and this text can feed any 'text' feedback method (other than relevance models). In other words, the modality of data is switched , from image to text or text to image, during the (pseudo) feedback process. These last models can be called *intermedia feedback*, or *transmedia feedback* techniques. In that sense, transmedia techniques generalize the pseudo-feedback idea present in cross-media relevance models, but are freed from the particular textual and/or visual models proposed by cross-media relevance models.

These latter techniques, based on intermedia feedback, go beyond simple late fusion and also simplify the difficulties often met by early fusion techniques, especially the famous "semantic gap" between the visual and textual features. This approach can be seen as an "intermediate level" fusion as the media fusion is after a first mono-media retrieval step based on mono-modal similarities (see 2.1 and 2.2).

Using pseudo-relevance feedback to automatically enrich the representation in one mode with the other one is an approach that has been exploited in several works. Most of the time, the Web itself is used as the multimedia repository and search engines such as Google Image are often usefully exploited to this aim.

For instance, to automatically annotate images – or, in other words, to associate with an image a synthetic textual representation that can be further used for indexing and retrieval – Li [20] and Wang [36] combined traditional Web-CBIR (content-based image retrieval) and Text Mining techniques (namely text clustering and term extraction) on the textual data (title, caption, surrounding text) associated with retrieved images. By a careful analysis of these textual data, they were able to find a reliable annotation that captures multiple aspects (regions) of the image. In the same vein, but restricted to purely textual information, Dowman [9] proposed a method to automatically annotate audio-video news streams, by a first retrieval step on the BBC Web site based on the textual transcripts of the news. After checking the relevance of the first page retrieved, they extracted a summary from this error-free Web page (title, category tags, named entitites) that was then used to annotate the original (noisy) transcript. Our work differs mainly from these approaches by the final objective (multimedia retrieval instead of image annotation), by the repository that is used and by the way the textual representation of an image is built from a set of pseudo-relevant documents.

Automatic illustration of a textual keyword or concept or, in a nearly equivalent view, automatic gathering of huge amount of images representative of a concept constitutes another famililly of methods and applications that can use the same principle of pseudo-relevance feedback in an hybrid (textual+visual) way. The ultimate goal is often to build a categorization model able to recognize a given concept (or keyword) in an image. Starting from a concept as a textual keyword, Yanai [37] proposed to leverage the vast resource of images on the Internet, by using search engine such as Google Image (that does nothing else than a purely textual retrieval based on the textual metadata that can be associated with images); the retrieved images were then used to build an initial visual model of the concept, following the principle of pseudo-relevance feedback; this model was then used to re-score candidate images and updated iteratively following an EM-like algorithm; by segmenting the candidate images into regions, he was able to offer more accurate visual models of the concept, taking into account only the more relevant regions of the image. Li adopted a very similar approach in [19], but in a purely visual way: there was no cross-media step. All these approaches share with our methodology the fact to be based on some pseudo-relevance feedback and, at some stage, switch from one medium to the other one. Once again, our work differs mainly from these approaches by the final objective (multimedia retrieval instead of automatic illustration), by the repository that is used and by the way the visual representation of a textual entity is built from a set of pseudo-relevant documents.

Iyengar and colleagues proposed in [15] a set of cross-media similarity measures that are not based on pseudo-relevance feedback, but on the “information bottleneck” principle, using an intermediate layer of concepts to make the link between both media. They used an audio-video training dataset (the TRECVID dataset), annotated with about one hundred concepts: these annotations allow them to build a textual model and a visual model for each concept or, equivalently, projection operators from textual and visual features towards a common concept space. As they are able to exploit image regions instead the global image, they used an intermediate step of alignment, in order to assign a unique to each image region. They also proposed to use log-linear models of fusion (instead of linear ones) in order to combine both mono-media and cross-media similarity measures. However, experimental results were not very satisfying.

Very recently, Quattoni [26] proposed a completely different approach to exploit textual meta-data (image caption, surrounding text) in image categorization and dimensionality reduction. Based on the “structural learning approach”, they used large quantities of unlabeled data with associated captions in order to improve the learning performance in future image classification problems. They learned a low-dimensional representation that reflects the semantic content of an image, by solving a set of auxiliary problems, namely to predict the presence/absence of a content word from the image features. The low-dimensional representation they obtained turned out to be efficient for generic visual concept detection tasks, even if these visual concepts were not used previously. This very intersting approach could be considered as “cross-media semi-supervised learning” but is of course completely different from what we are proposing in this paper.

2 Monomodal Similarities

First, we will introduce the monomodal similarities, visual and textual, that will serve as basic components for the trans-media similarities.

2.1 Image Similarity

As image signature (image representation), we use the Fisher Vector as proposed in [25]. This is an extension to the bag-of-visual-words [8] and the main idea is to characterize the image with the gradient vector derived from the generative probability model (visual vocabulary). This representation can then be subsequently fed to a discriminative classifier for categorization, or used to compute the similarities between images for retrieval.

The generative probability model in our case is the Gaussian mixture model (GMM) which approximates the distribution of the low-level features in images. It can be seen as a *Visual Vocabulary* where each Gaussian component $\mathcal{N}(\mu_i, \Sigma_i)$ models a visual word.

If we denote the set of parameters of the GMM by $\Phi = \{w_i, \mu_i, \Sigma_i, \, i = 1...N\}$ (w_i, being the mixture's weight), we can compute the gradient vector of the likelihood that the image was generated by the model Φ:

$$\nabla_{\Phi} \log p(I|\Phi) \ . \tag{1}$$

This gradient of the log-likelihood describes the direction in which parameters should be modified to best fit the data (image features). One of its advantages is that it transforms a variable length sample (number of local patches in the image) into a fixed length representation (which we will call Fisher Vector) whose size is only dependent on the number of parameters in the model ($|\Phi|$).

Before feeding these vectors to a classifier or computing similarities between images, each vector is first normalized using the Fisher Information matrix F_Φ, as suggested in [25] (see the paper for the computational details):

$$\mathbf{f}_I = F_\Phi^{-1/2} \nabla_\Phi \log p(I|\Phi) \tag{2}$$

with

$$F_\Phi = E_X \left[\nabla_\Phi \log p(I|\Phi) \nabla_\Phi \log p(I|\Phi)^T \right] \ .$$

and then each vector is re-normalized to have an L1-norm equal to 1.

The similarity measure between two images is then defined as the L1-norm of the difference between the normalized Fisher Vectors:

$$\begin{array}{c} sim_V(I,J) = sim_V(\mathbf{f}_I, \mathbf{f}_J) \\ = norm_{MAX} - ||\tilde{\mathbf{f}}_I - \tilde{\mathbf{f}}_J||_1 = norm_{MAX} - \sum_i |\tilde{f}_I^i - \tilde{f}_J^i| \end{array} \tag{3}$$

where $\tilde{f}^i$ are the elements of the re-normalized Fisher Vector $\tilde{\mathbf{f}}$.

These Fisher Vectors are rich image signatures and have shown state-of-the-art performance both in image categorization [25,1] and and content based image retrieval [7,1].

2.2 Text similarity

First the text is pre-processed including tokenization, lemmatization, word decompounding and standard stop-word removal. Then starting from a traditional bag-of-word representation (assuming independence between words), we adopt the language modeling approach to information retrieval. The core idea is to model a document d by a multinomial distribution over the words denoted by the parameter vector θ_d. A simple language model (LM) could be obtained by considering the frequency of words in d ($\#(d,w)$) (corresponding to the Maximum Likelihood estimator:

$$P_{ML}(w|d) = \frac{\#(w,d)}{|d|} \ .$$

The probabilities could be further smoothed by the corpus language model:

$$P_{ML}(w|C) = \frac{\sum_d \#(w,d)}{|C|}$$

using the Jelinek-Mercer interpolation :

$$\theta_{d,w} = \lambda \ P_{ML}(w|d) + (1-\lambda) \ P_{ML}(w|C) \ . \tag{4}$$

Using this language model, we can define the similarity between two documents using the cross-entropy function:

$$sim_T(d_1,d_2) = \sum_w P_{ML}(w|d_1) \log(\theta_{d_2,w})) \tag{5}$$

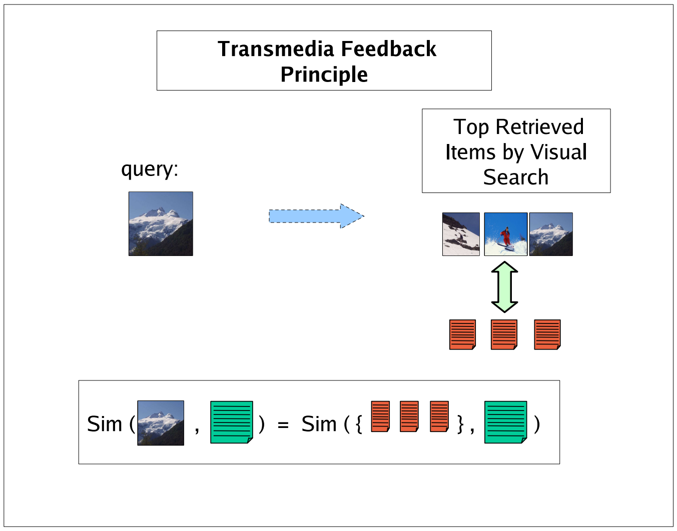

Transmedia Feedback Principle
query:
Top Retrieved Items by Visual Search
Sim (,) = Sim ({ },)

Fig. 1 Trans-media feedback principle

3 Trans-Media Similarities

Given monomodal similarities, we will now define transmedia similarities. We assume that we have a repository that contains a set of multimodal documents: $\mathcal{R} = \{d_1, d_2, \ldots, d_M\}$. We will denote the textual part of d_j by $T(d_j)$ and its visual part $V(d_j)$. Note that we are interested here in multi-modal documents containing text and image, but the techniques described below are applicable to other types of multimodal documents too. Let us now consider a new document d_q, with its corresponding $T(d_q)$ and $V(d_q)$ monomodal contents. Note that one of $T(d_q)$ and $V(d_q)$ can be void. Our concern is now to define how to compute trans-media similarities, i.e. a similarity between $T(d_q)$ and $V(d_j)$ or, reciprocally, a similarity between $V(d_q)$ and $T(d_j)$.

To tackle this task, we rely on the principle of trans-media relevance feedback (see figure 1). The main idea of trans-media relevance feedback is to first use one of the modes to gather relevant documents and then to switch to the other mode. To be more concrete, in our case we gather the N most relevant documents from $\mathcal{R}$ using some purely visual similarity measure (see section 2.1) with respect to the (visual) query $V(d_q)$ or, reciprocally, using a purely textual similarity (see section 2.2) with respect to the (textual) query $T(d_q)$. Let us call these top ranked documents n_{vj} and n_{tj} $(j = 1, \ldots N)$ respectively. Obviously, they are not the same in general.

This results in two "dual" representations of the query d_q : the textual, dual representation of the visual part $V(d_q)$ is an aggregate textual entity denoted by:

$$N_T(V(d_q)) = \{T(n_{v1}), T(n_{v2}), \ldots, T(n_{vN})\} \tag{6}$$

Similarly, the visual, dual representation of the textual part $T(d_q)$ is an aggregate visual entity denoted by:

$$N_V(T(d_q)) = \{V(n_{t1}), V(n_{t2}), \ldots, V(n_{tN})\} \tag{7}$$

We will use these new "dual" representations to compute the trans-media similarity measures. The problem now amounts to define a monomodal similarity measure between some aggregate $N_T(V(d_q))$ (resp. $N_V(T(d_q))$) and $T(d_j)$ (resp. $V(d_j)$). To this aim, we propose two approaches. The first one remains at the feature or representational level (section 3.1) , while the other directly focuses on aggregating the similarities rather than aggregating the features (section 3.2).

3.1 Models for representing aggregated objects

This first approach is more related to textual aggregates, but we can envisage an analog approach for the visual part. The main idea is that we consider the set of aggregated texts in $N_T(V(d_q))$ as a set of relevant texts (they are all related to the visual part of the query document d_q) and try to model this "relevance concept" **F** by a language model (LM) θ_F. Here, a Language Model must be understood as some probability distribution over the words of the vocabulary. This idea has its inspiration from the language modeling approach to information retrieval originally designed to enrich textual queries (see [38] for the basic approach ; more elaborated techniques of feedback for language models can also be envisaged : e.g. [29]).

Let θ_F be a multinomial parameter standing for the distribution of relevant terms in $N_T(V(d_q))$. In other words, θ_F is a probability distribution over words but peaked on relevant terms. It can be easily shown that the likelihood of observing **F** assuming that it is generated by the process characterized by θ_F is:

$$P(\mathbf{F}|\theta_F) = \prod_{n_j \in N_T(V(d_q))} \prod_{w \in n_j} (\lambda P(w|\theta_F) + (1-\lambda) P(w|\theta_{\mathcal{R}}))^{\#(w,n_j)} \tag{8}$$

where $P(w|\theta_{\mathcal{R}})$ is word probability built upon the whole repository (what is called usually the corpus language model), λ is a fixed parameter, which can be understood as a noise parameter for the distribution of terms and $\#(w, n_j)$ is the number of occurence of term w in document n_j.

The language model θ_F is learned by maximum likelihood optimization with an Expectation Maximization algorithm (as described in [7]). Now θ_F is considered as the new representation of d_q and we can use standard similarity measures over probability distributions (like the Kullback-Leibler Divergence or the Cross-entropy) in order to derive a similarity between θ_F and the textual part $T(d_j)$ on any document (in the repository or not), as soon as the latter is represented by a language model too.

Finally, note that we illustrated the approach using $N_T(V(d_q))$ to derive a textual language model of θ_F that can be used in conjunction with the original language model of the textual view of the query $T(d_q)$, using a simple linear mixture. We can also derive a similar scheme using $N_V(T(d_q))$ to derive a new representation (actually some generalized Fisher Vectors) of the "relevance concept", this time relying on Rocchio's method that is more adapted to continuous feature representation ([28]):

$$f_F = \frac{1}{N} \sum_{v \in N_V(T(d_q))} f_v + \frac{1}{|\mathcal{R}|} \sum_{d \in \mathcal{R}} f_{V(d)} \tag{9}$$

3.2 Aggregating Similarity Measures

As an alternative to define a model for the aggregated objects, we can directly aggregate the similarities between the new document and the elements of the aggregated object. For example, if we use the visual query to gather $N_T(V(d_q))$ (see (6)), then the new cross-media similarity measure between d_q and any multi-modal object is defined by :

$$\begin{aligned} sim_{TV}(d,d_q) &= sim_T(T(d), N_T(V(d_q))) \\ &= \textstyle\sum_{t \in N_T(V(d_q))} sim_T(T(d),t) \end{aligned} \tag{10}$$

where sim_T is typically defined by equation (5). Note that we do not use the visual part of d nor the textual part of d_q which suggests that this similarity measure can also be used directly to compute the similarity between a purely visual object and purely textual object.

In analogy we can switch the two modalities and write:

$$\begin{aligned} sim_{VT}(d,d_q) &= sim_V(V(d), N_V(T(d_q))) \\ &= \textstyle\sum_{v \in N_V(T(d_q))} sim_V(V(d),v) \end{aligned} \tag{11}$$

where sim_V is typically defined by equation (3).

In this case, we do not use the textual part of d nor the visual part of d_q.

In the same vein, coming back to the monomodal case, we could define similarity measures in one mode that does not switch mode in the pseudo-relevance feedback phase. This naturally leads to the following equations:

$$\begin{aligned} sim_{TT}(d,d_q) &= sim_T(T(d), N_T(T(d_q))) \\ &= \textstyle\sum_{t \in N_T(T(d_q))} sim_T(T(d),t) \end{aligned} \tag{12}$$

and

$$\begin{aligned} sim_{VV}(d,d_q) &= sim_V(V(d), N_V(V(d_q))) \\ &= \textstyle\sum_{v \in N_V(V(d_q))} sim_V(V(d),v) \end{aligned} \tag{13}$$

where

$$N_T(T(d_q)) = \{T(n_{t1}), T(n_{t2}), \ldots, T(n_{tN})\} \tag{14}$$

is the aggregated text set obtained by pseudo-relevance feedback on the textual query $T(d_q)$ and

$$N_V(V(d_q)) = \{V(n_{v1}), V(n_{v2}), \ldots, V(n_{vN})\} \tag{15}$$

is the aggregated image set obtained by pseudo-relevance feedback on the visual query $V(d_q)$.

Finally, we can combine all the similarities to define a global similarity measure between two multi-modal objects d and d_q using, for instance, a linear combination:

$$\begin{aligned} sim_{glob}(d,d_q) &= \lambda_1 sim_{TT}(d,d_q) + \lambda_2 sim_{TV}(d,d_q) \\ &+ \lambda_3 sim_{VT}(d,d_q) + \lambda_4 sim_{VV}(d,d_q) \end{aligned} \tag{16}$$

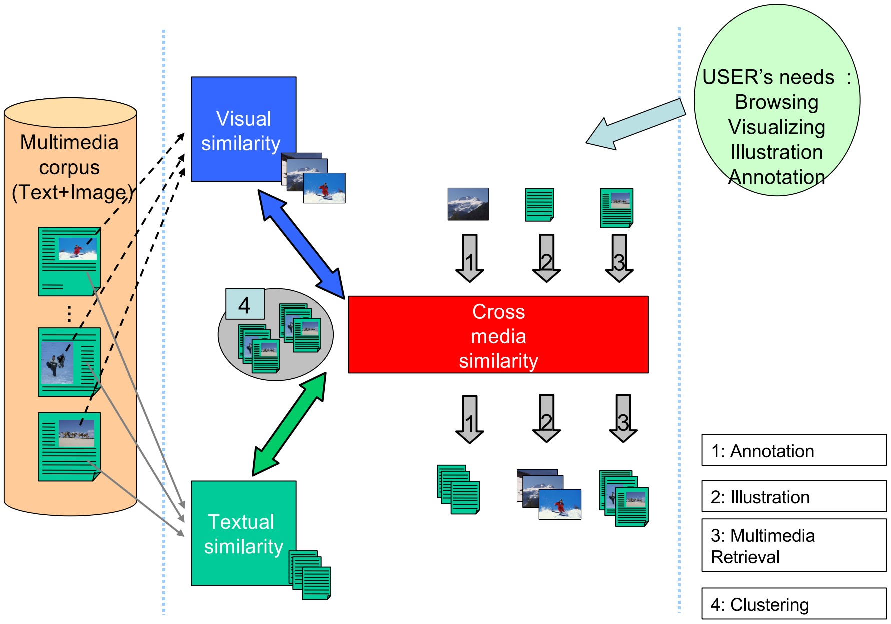


Fig. 2 Applications Scenarios of a trans-media system

4 Application Scenarios

We now describe different scenarios where trans-media similarities are effective. Generally, a user manipulating multimodal data needs clustering and retrieval functionalities. In order to browse a corpus efficiently, a user also need to express her information need in different modes : visual only, textual only or finally in a mixed representation. Moreover, the user also needs to switch her expression mode from a medium to another one. Figures 2 presents some of these scenarios which rely on trans-media similarities. In the following section, we will explain in details these scenarios: retrieval, image annotation, text illustration and clustering.

4.1 Ranking and retrieval

Given a multi-media repository $\mathcal{R}$, our goal is to rank these documents with respect to a new multimedia document (query) d_q based on trans-media relevance feedback. This was already tackled in section 3 for mono-media type of queries ($d_q = T(d_q)$ or $d_q = V(d_q)$) and, in this section we go beyond and show how we can integrate and exploit the hybrid nature of the query to improve the retrieval performance. In accordance with the two approaches of section 3, we have again several options:

1. *Language Model of the Aggregated Text*
   Using the visual part of the query document, we gather the aggregated text:

$$N_T(V(d_q)) = \{T(n_{v1}), T(n_{v2}), \ldots, T(n_{vN})\}$$

and estimate a language model θ_F for this aggregate. But we can also estimate a language model θ_q of the textual part of the query $T(d_q)$. Combining the two language models through interpolation:

$$\theta_{q'} = \alpha\theta_q + (1-\alpha)\theta_F \tag{17}$$

leads to a new query language model $\theta_{q'}$ that can be used with the Cross-Entropy similarity measure (see section 2.2) to perform a new retrieval on the textual part of objects in $\mathcal{R}$. Setting the value of α is done experimentally and adapted to the particular collection. The robustness of the estimation of θ_F has a significant impact on the value of α. Lastly, the value of α can be interpreted as a mixing weight between image and text.

2. *Aggregating Similarity Measures*
In section 3.2 we have defined several trans-media similarity measures. These measures can be used directly to re-rank the documents d_j of $\mathcal{R}$ according to the query d_q. Indeed if $V(d_q)$ is the visual part of the query document d_q and $T(d_j)$ belongs to the initial feedback set $N_T(V(d_q))$, then the rank of the neighbors of $T(d_j)$ in the textual sense will be increased even if they are not so similar from a purely visual viewpoint (even when having no image at all). Similarly, when we use $T(d_q)$ as query we will increase the ranks of the objects whose image part is similar to $V(d_k) \in N_V(T(d_q))$ even if they have no textual similarity (or no text at all) with $T(d_q)$. Combining them as in (16) will take into account both the textual and visual part of the query in the re-ranking process.

The main advantage of this method is that all the measures in (16) involves mainly pair-wise similarities between the documents of $\mathcal{R}$. Therefore assuming that the corpus is of reasonable size, those similarities can be pre-computed, so the step after the first search (relevance feedback) can be very fast and efficient.

These retrieval methods have been evaluated at ImageCLEFphoto 2007 [7] and ImageCLEFphoto 2008 Evaluation Forums [1]. In these challenges, our image and trans-media similarity based retrieval systems demonstrated very good performance (see [14]). We give in table 1 the best results we obtained for ImageCLEFphoto 2008, using mono-media (text or image) or multimedia (text and image, using late fusion or trans-media similarities) approaches [1]. We used MAP (Mean Average Precision) and P@20 (precision at 20) for measuring each run performance.

Media	MAP	P@20
Text only	0.239	0.293
Image only	0.151	0.328
Late fusion	0.348	0.437
Trans-media	0.424	0.554

Table 1 Some results obtained for ImageCLEFphoto 2008

Our runs were ranked among the best ones. As we can see, the results provided by the trans-media approach clearly outperforms both the mono-media approaches and the simple late fusion technique.

4.2 Incorporating diversity into the top list

In ImageCLEFphoto 2008, one of the main problems we had to address was to improve the search results by avoiding the redundancy among the first retrieved elements. In other words, the participants were asked to provide results that are both relevant and diverse.

To address this issue, we opted for a two-step approach. The first step consists in ignoring the question of diversity. In other words, we first try to find the most relevant objects using the material introduced in 3.2 and mentioned in the previous paragraph. Then, in a second step, we re-rank the first relevant elements by taking into account their mutual similarities in order to avoid redundancy and thus to promote diversity.

We investigated two main families of methods: implicit and explicit clustering-based approaches. The first method is commonly known as "Maximal Margin Relevance" (MMR) [4]. It amounts to re-rank the search results so that the element chosen at rank j has to be dissimilar to elements that were already selected at ranks $j' < j$. The second method is based on an explicit clustering of the first k elements, followed by a strategy designed to re-rank the elements so that many different clusters are represented among the first elements of the re-ranked list (see [1] for more details).

In both approaches, we are given two components: $S1$, the similarity vector between a query and the documents of the database and $S2$, the pairwise similarity matrix between documents of the database.

In the experiments we did for ImageCLEFphoto 2008, we used the results given by the trans-media similarities for measuring the relevance of an object with respect to a query. Therefore, $S1$ is typically the trans-media run which, as mentioned in the previous paragraph, was the best run we obtained regarding MAP and P@20.

$S2$ is used for representing the thematic proximity between two documents that will allow us to promote diversity. We tested different similarity matrices. Particularly, in the case of clustering-based approaches, we investigated if either pure textual or multimedia similarities performed better. For the pure textual similarities we used the basic $tfidf$ between the text part of the documents. For the multimedia approach we applied our trans-media similarity measure to the multimedia documents. We report in table 2, the results we obtained. For measuring diversity, the CR@20 (cluster recall at 20) index is used. It is the number of different clusters that is present among the 20 first retrieved elements divided by the total number of clusters that were manually identified by the evaluators.

Media	MAP	P@20	CR@20
Text only	0.360	0.512	0.408
Trans-media	0.369	0.527	0.411

Table 2 Some results obtained for ImageCLEFphoto 2008 with diversity seeking goal

Compared to the pure textual method, the trans-media approach gives better results for both the diversity and the relevance measures. Consequently, for this task, it is also beneficial to combine visual and textual information using our trans-media similarities method in order to improve the results.

4.3 Relating text and image through a repository

In section 3 we defined similarity measures between a new document d and the elements of the aggregated object obtained by trans-media relevance feedback using a multi-modal query d_q. However, neither d nor d_q has to be necessarily multi-modal.

4.3.1 Image Annotation

If we consider simply $d_q = I$ an image, and $d = T$ a pure text not necessarily in $\mathcal{R}$, (11) becomes:

$$\begin{aligned} sim_{TV}(T, I) &= sim_T(T, N_T(I)) \\ &= \textstyle\sum_{T(d_j) \in N_T(I)} sim_T(T, T(d_j)) \end{aligned} \tag{18}$$

a similarity defined between an image and text based on a common multimodal repository.

4.3.2 Text Illustration

By duality, we can use the text T as query $d_q = T$ and the image as the new document. Then (11) becomes:

$$\begin{aligned} sim_{VT}(I, T) &= sim_V(I, N_V(T)) \\ &= \textstyle\sum_{V(d_j) \in N_V(T)} sim_V(I, V(d_j)) \end{aligned} \tag{19}$$

another similarity defined between the image and the text. Furthermore, we can cluster the top retrieved images and illustrate the different views on the given text.

4.4 Trans-media Similarity-based Searching, Clustering and Visualization

Given a set of multimedia documents, our goal is to find the underlying structure of the dataset and also to visualize the proximity relations between the multimedia documents by means of projections in a suitable space[1]. In other words, we want to cluster the datasets, in order to capture the proximity relationships between multimedia documents and then to represent these documents into a space which respects these proximity relationships.

Clustering and visualization processes are of interest for browsing a multimedia corpus such as Wikipedia as we will show in our experiments in section 5.2. Indeed, it allows to seek for information in an exploratory mode by using either the visual or the textual or both the combination of visual and textual aspects of the documents. We can talk about a *multi-view* access of the multimedia corpus.

When dealing with monomedia documents, many methods exist for clustering and visualizing datasets of images or datasets of texts. In our context we are interested in clustering and visualizing multimedia documents where we have to fuse information from two different monomedia sources. However, as we already mentioned previously, we want to combine the visual and the textual information of a same document at an intermediate level which is different from an early fusion approach or a late fusion approach that were already suggested. In other words, we want to better take into

[1] such as graph embedded into a 2D representation for example

account the semantic correlation between the visual view and the textual view of a same document. For doing so, we propose firstly, to apply our trans-media similarities given in section 3 and then, to use the global similarity matrix given by (16) in clustering and projection methods.

Indeed, the global similarity matrix aggregates monomedia similarities with trans-media similarities computed at an intermediate level. As a result, we obtain a global similarity matrix between multimedia documents which aims at overcoming the "semantic gap" between visual and textual representation. Finally, by using this measure, we integrate semantic correlation better than basic early and late fusion methods do.

We will give some examples in section 5 which show that using our global similarity, we are able to retrieve and explore a multimedia corpus in a more efficient way than by using visual similarities or textual similarities only.

Taking in consideration all these different points, we propose a system which allows a user to switch between the different kinds of representation a multimedia document or query can have : monomedia or multimedia representations.

We first compute monomedia similarities based on the methods we have defined in sections 2.1 and 2.2. Given sim_V in (3) and sim_T in (5), we compute trans-media similarities according to section 3 and we finally retain the aggregated global similarity sim_{glob} in (16) as the similarity matrix that we will use for dealing with multimedia representation of documents or queries.

These three similarity matrices are our basic inputs for similarity-based search, clustering and projection processes :

- they are directly used to retrieve similar documents d according to a given document d_q corresponding to a query
- they are the inputs of clustering algorithms which allow to gather in a same cluster documents which are similar
- they are also the inputs of projection algorithms which allow to represent into a 2D space the graph corresponding to the original proximity relationships between documents

As each representation namely visual, textual and trans-media, can be used, the user has a *multi-view acces* of the documents of the corpus. In other words, we have three clustering and projection results that the user could use depending on the type of representation of the documents he is interested in.

5 Multimedia Systems

The next section is devoted to the presentation of two systems, based on the previous scenarios. The first system is a Travel Blog Assistant: its main goal is to help a user generating content for her blog, by linking text and images. The second system is a browsing tool for multimedia corpora, such as Wikipedia. It enables clustering and retrieval in a dataset of composite objects.

5.1 Travel Blog Assistant System

To illustrate some of these techniques working on real examples, we build a so called Travel Blog Assistant System (TBAS). Such a system can be of interest for real travelogue websites such as [31,33,34,13,30,27,32].

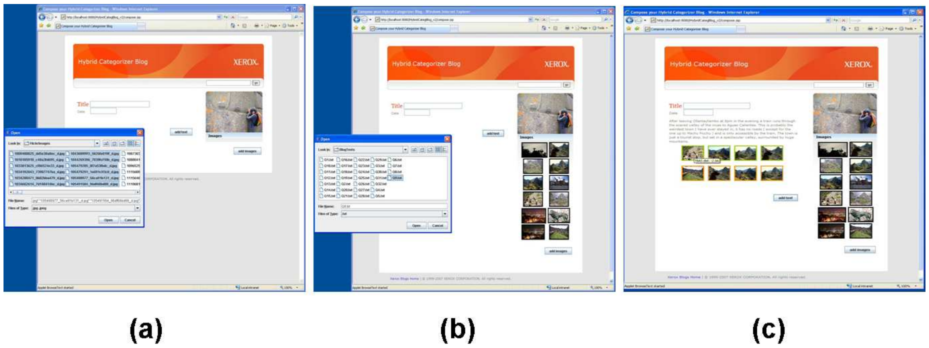


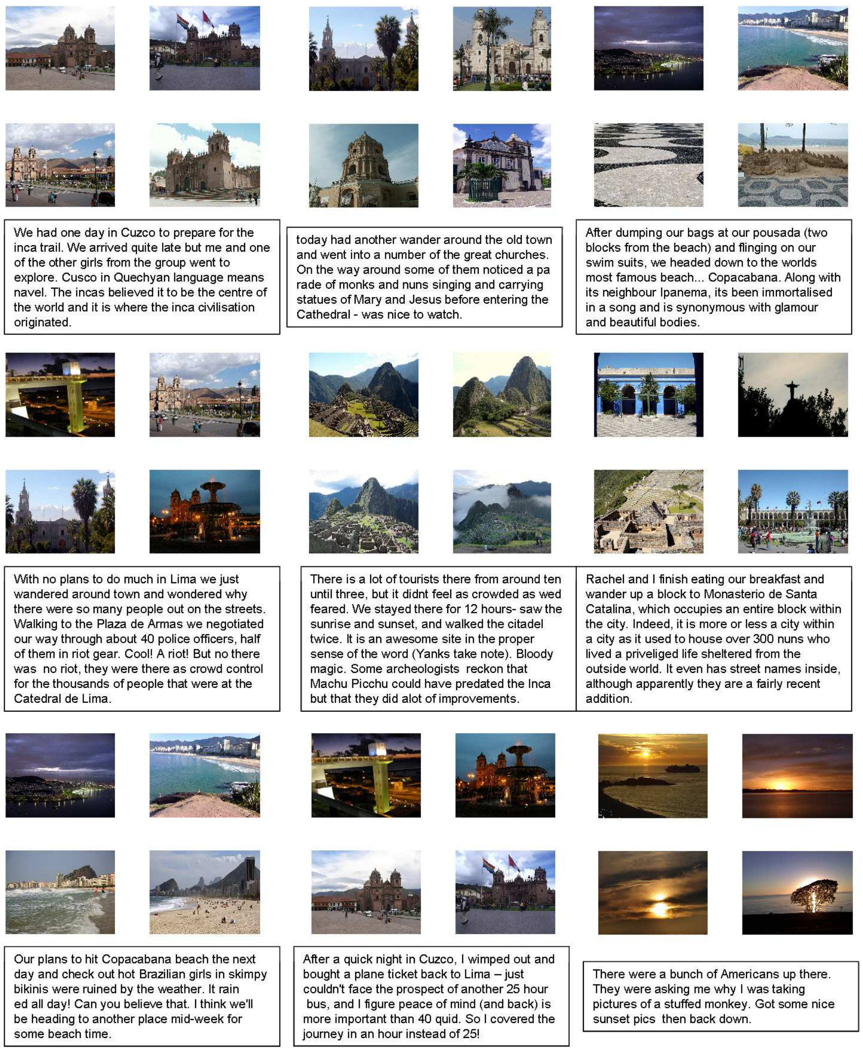
Fig. 3 The prototype interface of our TBAS system.

Figure 3 shows the overall schema of the TBAS system. Its main steps are:

- The user uploads a set of images to be considered. If he edits his blog during his travel he simply plugs his camera and uploads the images captured that day and the previous days (Figure 3(a) and (b)). The images are then pre-processed and image metadata are added to each image. The metadata file can contain different type of information and is typically a structured xml file. First, information such as date, time, location (GPS) can be provided by the camera (exif file). Finally, textual information (tags) are obtained using a trans-media feedback using a repository of multi-media objects (e.g. the database of the travelogue site itself) as described in section 4.3.1.
- As represented in Figure (3(c)), the typed (or uploaded) blog text is pre-processed at a paragraph level and related to the images through the repository (see section 4.3). These similarities allow the system to rank the images according to a given paragraph. The top N ranked images (thumbnails) are shown to the user as illustrative examples of the paragraph The user has naturally the possibility to select between them, or to reject and ask for the next N images.
- When all paragraphs are processed and the illustrative images selected, the system shows a preview of the composed “blog”. The page layout can be computed automatically or selected from a set of template layouts. When the user is satisfied with the result, he can simply publish the blog.

A demo version of the TBAS system was developed and tested on a relatively small database (compared to the data we can get on the web). To simulate the traveler’s image data, we downloaded under Creative Common licences 185 images from the online photo sharing site Flickr [12]. For the travel blog text we collected real blog paragraphs from two travel blog sites, RealT [27] and Tpod [33]. In all cases, in order to ensure the semantic correlation between images and blog texts, we focused on two main destinations, Peru and Brasil, and used a few touristic names (cities, etc.) associated with them as keywords or tags for focusing our search in gathering the images and the blog paragraphs.

In figure 4, we show the most similar images for a given blog paragraph, using our transmedia similarities: we ranked the 185 downloaded Flickr images for each travel blog text and selected the top 4 images for each blog.



We had one day in Cuzco to prepare for the inca trail. We arrived quite late but me and one of the other girls from the group went to explore. Cusco in Quechyan language means navel. The incas believed it to be the centre of the world and it is where the inca civilisation originated.

today had another wander around the old town and went into a number of the great churches. On the way around some of them noticed a pa rade of monks and nuns singing and carrying statues of Mary and Jesus before entering the Cathedral - was nice to watch.

After dumping our bags at our pousada (two blocks from the beach) and flinging on our swim suits, we headed down to the worlds most famous beach... Copacabana. Along with its neighbour Ipanema, its been immortalised in a song and is synonymous with glamour and beautiful bodies.

With no plans to do much in Lima we just wandered around town and wondered why there were so many people out on the streets. Walking to the Plaza de Armas we negotiated our way through about 40 police officers, half of them in riot gear. Cool! A riot! But no there was no riot, they were there as crowd control for the thousands of people that were at the Catedral de Lima.

There is a lot of tourists there from around ten until three, but it didnt feel as crowded as wed feared. We stayed there for 12 hours- saw the sunrise and sunset, and walked the citadel twice. It is an awesome site in the proper sense of the word (Yanks take note). Bloody magic. Some archeologists reckon that Machu Picchu could have predated the Inca but that they did alot of improvements.

Rachel and I finish eating our breakfast and wander up a block to Monasterio de Santa Catalina, which occupies an entire block within the city. Indeed, it is more or less a city within a city as it used to house over 300 nuns who lived a priveliged life sheltered from the outside world. It even has street names inside, although apparently they are a fairly recent addition.

Our plans to hit Copacabana beach the next day and check out hot Brazilian girls in skimpy bikinis were ruined by the weather. It rain ed all day! Can you believe that. I think we'll be heading to another place mid-week for some beach time.

After a quick night in Cuzco, I wimped out and bought a plane ticket back to Lima – just couldn't face the prospect of another 25 hour bus, and I figure peace of mind (and back) is more important than 40 quid. So I covered the journey in an hour instead of 25!

There were a bunch of Americans up there. They were asking me why I was taking pictures of a stuffed monkey. Got some nice sunset pics then back down.

Fig. 4 Top 4 images to be associated with a blog's paragraph.

5.2 Trans-media Browsing of Wikipedia Pages

In this section, we present a recent work done in the context of Cap Digital - Infomagic, which is a french project on digital content creation and knowledge management. In a joint work with INA, we have adressed a task of this project about "browsing a multimedia corpus".

The scenario is the following one : a user has at his disposal a multimedia corpus constituted of documents which have both a visual and a textual representation; he is seeking for information within this corpus but he does not know exactly how to express his query; so, he wants to browse the multimedia corpus in an exploratory way in order to progressively refine the expression of his information needs. Given this context, our goal is to provide the user with a system which has the following properties :

- when seeking information in such a multimedia corpus, the user can be interested in either the visual feature or the textual feature or the combination of both visual and textual features, of the documents. His interest, moreover, can change at any time : suppose, in a first time, that the user entries a textual query, then a list of multimedia objects is retrieved according to the textual similarities. Looking at this list, the user, in a second time, is more interested in a particular multimedia document because its associated image has attracted his attention. Finally, he would like to look for similar multimedia documents by taking into account both the image and the text associated to this relevant multimedia document. In that context, we want the system to allow the user to search using trans-media similarities. In a general manner, we want the system to allow the user to choose the kind of queries, monomodal or transmodal, which reflects the most what he is searching for. As a result, the system has to integrate as many kind of similarities as types of queries the user can express.
- when seeking information in an exploratory mode, an efficient system is the one which assists the user by providing him with a visualization tool which allows to represent proximity relationships between the multimedia documents. In that case, as we have described in section 4.4, clustering and projection methods for information retrieval are relevant processes. Firstly, clustering the set of multimedia documents allows to structure the dataset of multimedia documents into a meaningful partition based on similarities between documents. Using this partition and given a query, the system can suggest the user a list of clusters of relevant documents rather than a simple list of relevant documents. Secondly, projection methods allow to represent the multimedia documents, by means of a graph embedded into a 2D space for example, such that similar documents are close to each other. Using a visualization tool, the user can then explore and browse more intuitively the corpus. In general terms, we want the system to allow similarity-based search where the user can “navigate” in an efficient way within the corpus using the projection and the visualizing tool.

In collaboration with INA, we developed a system which main characteristics are described in section 4.4 and adressed most of the scenarios presented in figure 2. In the context of the previous scenario, this system allows the user to be assisted in his information search needs. We experiment such a system using a subset of the french Wikipedia corpus. We give some details and some screenshots of this systems below.

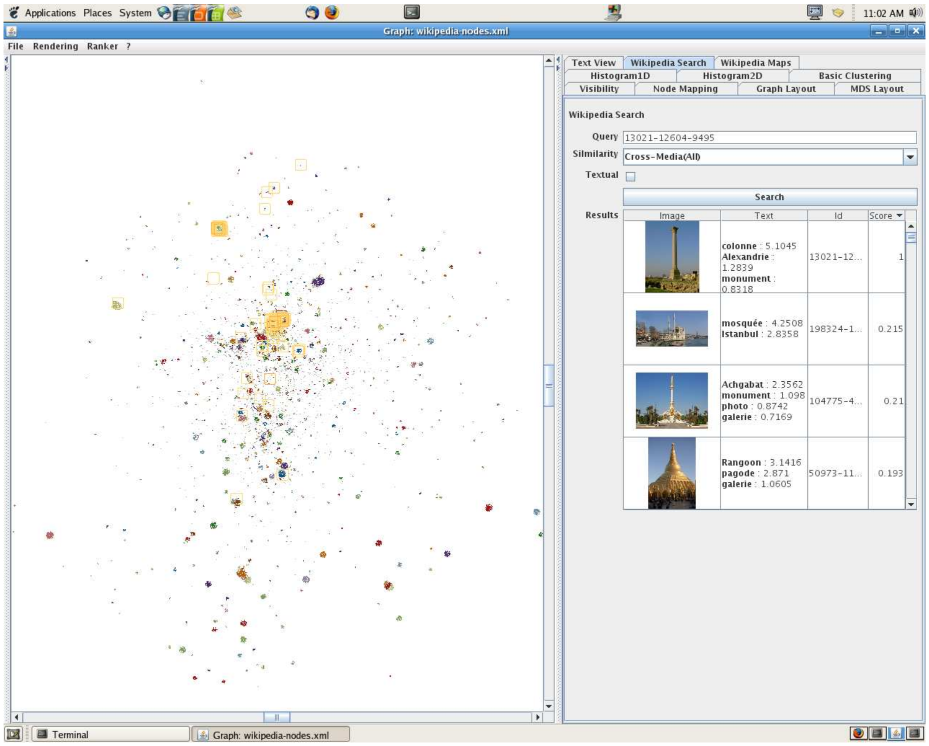
5.2.1 Extracting multimedia documents from Wikipedia corpus

In this paragraph, we briefly introduce the multimedia corpus on which we applied our system. This corpus concerns around 8,500 pages taken from the french Wikipedia corpus. We extracted these pages from the xml dump done in September 2007 and provided by the Wikipedia Foundation[2].

We selected pages according to their categories. In that perspective, we used the category tree[3] of the Wikipedia pages and we selected any pages which has at least one category that has the category “Geography” or “Tourism” as root.

[2] http://download.wikimedia.org/

[3] http://stats.wikimedia.org/EN/CategoryOverviewIndex.htm



Wikipedia Search

Query 13021-12604-9495

Silmilarity Cross-Media(All)

Textual □

Results

Image	Text	Id	Score
	colonne : 5.1045 **Alexandrie** : 1.2839 **monument** : 0.8318	13021-12...	1
	mosquée : 4.2508 **Istanbul** : 2.8358	198324-1...	0.215
	Achgabat : 2.3562 **monument** : 1.098 **photo** : 0.8742 **galerie** : 0.7169	104775-4...	0.21
	Rangoon : 3.1416 **pagode** : 2.871 **galerie** : 1.0605	50973-11...	0.193

Fig. 5 General screenshot of INA's tool

For these pages we looked at their images and we extracted the following textual descriptions :

- the path of titles, from the page title to the paragraph title where an image is encountered
- the free-text description of an image which corresponds to the comments we can see below most of images in Wikipedia
- the whole paragraph where an image is encountered

A multimedia document is an image associated to the text given by the concatenation of these three textual desciptions. We obtained more than 23,500 multimedia objects associated to more than 19,000 images and more than 14,000 texts (an image can leads to more than one multimedia objects as it could be associated to several texts and vice versa).

5.2.2 Browsing the Wikipedia corpus

We have explained previously, in which kind of scenarios we place ourselves. In this paragraph, we give some screenshots of the visualization tool developed by INA for illustrating how our monomedia and transmedia similarities methods can be used for browsing efficiently a multimedia corpus.

In figure 5, we show a general view of the browsing tool :

- on the left part, we can see the representation in a 2D space of the multimedia documents (in this figure the graph represented is the one based on the textual representation only). We can clearly observe "round packets" of points. These last are clusters of similar documents and the different colors corespond to different clusters. The user can zoom in and zoom out this graph. The point are multimedia documents which can be represented either by their associated images or the most salient words of the associated text. Furthermore, the user can acces directly to the Wikipedia page (or paragraph) corresponding to a selected multimedia object.
- on the top right part, there are several tabs which allow the user to customize different parameters such as the layout or the format. But, as we describe it below, the most relevant options that concern this paper are the tabs which allow the user to search into the corpus using different similarities and to change the type of graph depending on the representation of the multimedia documents the user want to use. By choosing one particular tab, there is a particular UI which appears below the tabs region.

  The most relevant tabs with respect to the present paper are the following ones :

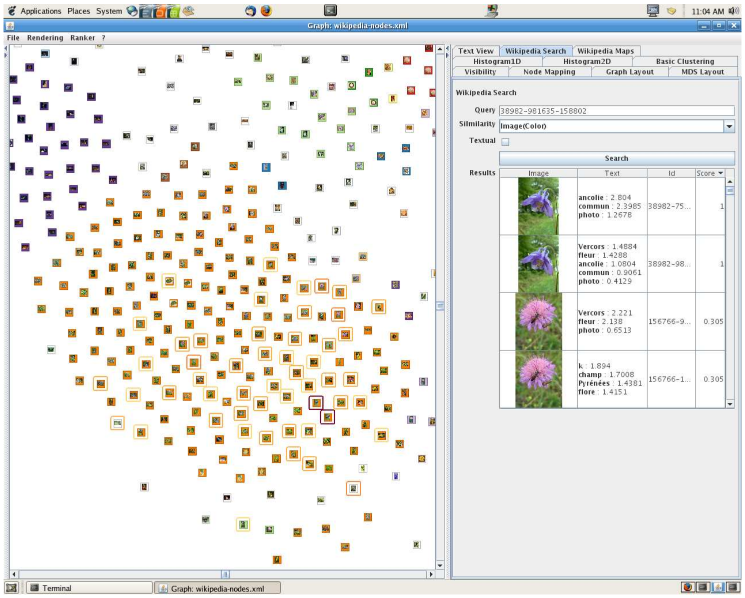
- the "Wikipedia Search" tab which allows the user to search in the corpus by using different kinds of similarity : in figure 5 for example, we selected a multimedia document, then a particular similarity matrix (trans-media in that case) and this gives as output (on the right) the list of the 100 multimedia documents (the associated image and the most relevant textual features) which are the most similar to the query based on the chosen similarity matrix. In the graph, the documents which belong to the list are surrounded with a square
- the "Wikipedia Maps" tab which allows to change the graph according to a representation and its associated similarity matrix. When switching from one graph to the other, there is an animation which allows the user to visualize how the position of the surrounded documents moves.

5.2.3 Combining visual and textual information can be relevant in information retrieval and seeking

In this paragraph, we illustrate with an example, how trans-media similarity based search, using our global similarities given by (16), can improve information retrieval tasks. Suppose in a first time, that we are interested in an image query (a flower which corresponds to the first flower of the list shown in figure 6). Using our tool, we first search for the most similar documents (considering their visual part) by using the visual similarity matrix and by visualizing these documents in the corresponding graph.

In figure 6, we show the region of the graph where most of the retrieved documents are. We can observe that these relevant documents are in one big cluster that is constituted, from the visual point of view, of photos of flowers (the other images of flowers are not surrounded mainly because the retrieved list is limited to the 100 most relevant documents).

In order to refine the results obtained, we can represent the same list of documents in another graph namely the one based on the global similarity matrix. In other words, we want to take into account the combination of the visual and the textual part of the documents. In figure 7, we show the most dense region of relevant documents in this global graph.



Wikipedia Search

Query 38982-981635-158802

Silmilarity Image(Color)

Textual □

Results

Image	Text	Id	Score
	ancolie : 2.804 commun : 2.3985 photo : 1.2678	38982-75...	1
	Vercors : 1.4884 fleur : 1.4288 ancolie : 1.0804 commun : 0.9061 photo : 0.4129	38982-98...	1
	Vercors : 2.221 fleur : 2.138 photo : 0.6513	156766-9...	0.305
	k : 1.894 champ : 1.7008 Pyrénées : 1.4381 flore : 1.4151	156766-1...	0.305

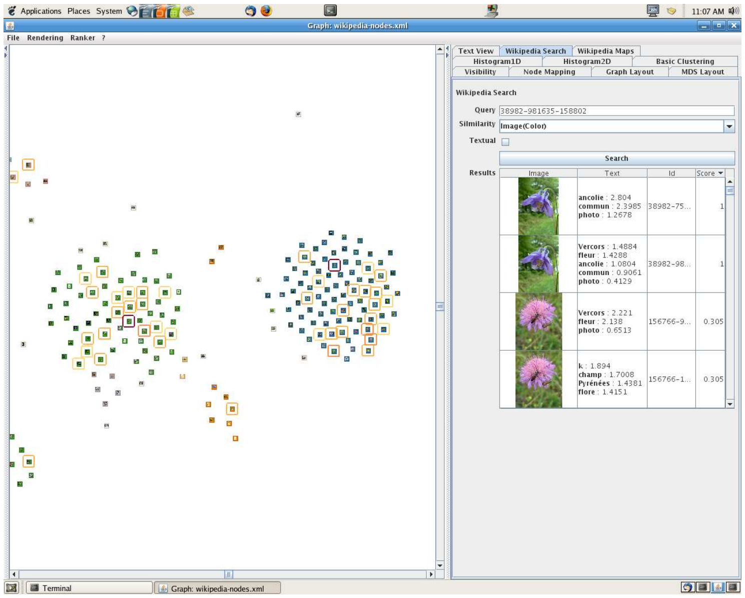
Fig. 6 Retrieved documents represented in the graph based on visual similarities

Unlike the graph based on the visual similarities, we can mainly observe in figure 7, two different clusters. From the visual point of view, it is hard to distinguish these two clusters as there are flowers in both of them. But, by considering the fact that in this representation, we use the global similarities which integrate trans-media similarities, we can deduce that this separation is mainly due to the textual features. This is actually the point. Indeed, by looking at the textual description of those two clusters in more details, we observe that the cluster on the right concerns flowers from the “Vercors” which is a montain region in the south-east of France whereas, the cluster on the left, concerns flowers which are from the “Pyrénées” which is another mountain region but in the south-west of France. In this example, we show that despite high visual similarities, the global similarities, by taking into account in an efficient way the textual information of the multimedia documents, allows to refine information retrieval.

The combination of visual and textual features has allowed to know that the image query is a flower which is encountered in two different regions of France. Furthermore, in an exploratory mode, it also allows, while seeking deeper into the clusters, to discover other flowers that are in those two different regions.

6 Conclusion

We have presented a framework for accessing multimodal data. First of all, the theoretical contribution is the extension of the principle of trans-media feedback, into a metric view: the definition of trans-media similarities. As it was shown, these new similarity measures of cross-content enables to find illustrative images for a text, to annotate an image, cluster or retrieve multi-modal objects. Moreover, the trans-media similarities are not specific to image and text: they can be applied to any mixture of media (speech, video, text) or views of an object. Most importantly, we have shown how



Wikipedia Search

Query 38982-981635-158802

Silmilarity Image(Color)

Textual □

Results

Image	Text	Id	Score
	ancolie : 2.804 **commun** : 2.3985 **photo** : 1.2678	38982-75...	1
	Vercors : 1.4884 **fleur** : 1.4288 **ancolie** : 1.0804 **commun** : 0.9061 **photo** : 0.4129	38982-98...	1
	Vercors : 2.221 **fleur** : 2.138 **photo** : 0.6513	156766-9...	0.305
	k : 1.894 **champ** : 1.7008 **Pyrénées** : 1.4381 **flore** : 1.4151	156766-1...	0.305

Fig. 7 Retrieved documents represented in the graph based on global similarities

these techniques can be used in two use cases: the travel blog assistant system and the multimedia browsing tool. These two applications stress the necessity of cross-media systems, where no monomedia systems can solve the user's problem, nor adress all the different user's need at the same time.

Acknowledgements The authors want to thank particularly INA for their contributions in our work and Florent Perronin for his greatly appreciated help in applying some of the Generic Visual Categorizer (GVC) components.

We would like also to acknowledge the following Flickr users whose photographs we reproduced here under Creative Common licences:

Tatiana Sapateiro	http://www.flickr.com/photos/tatianasapateiro
Pedro Paulo Silva de Souza	http://www.flickr.com/photos/pedrop
Leonardo Pallotta	http://www.flickr.com/photos/groundzero
Laszlo Ilyes	http://www.flickr.com/photos/laszlo-photo
Jorge Wagner	http://www.flickr.com/photos/jorgewagner
UminDaGuma	http://www.flickr.com/photos/umindaguma
Scott Robinson	http://www.flickr.com/photos/clearlyambiguous
Gabriel Flores Romero	http://www.flickr.com/photos/gabofr
Jenny Mealing	http://www.flickr.com/photos/jennifrog
Roney	http://www.flickr.com/photos/roney
David Katarina	http://www.flickr.com/photos/davidkatarina
T. Chu	http://www.flickr.com/photos/spyderball
Bill Wilcox	http://www.flickr.com/photos/billwilcox
S2RD2	http://www.flickr.com/photos/stuardo
Fred Hsu	http://www.flickr.com/photos/fhsu
Abel Pardo López	http://www.flickr.com/photos/sancho_panza
Cat	http://www.flickr.com/photos/clspeace
Thowra_uk	http://www.flickr.com/photos/thowra
Elena Heredero	http://www.flickr.com/photos/elenaheredero
Rick McCharles	http://www.flickr.com/photos/rickmccharles
Marília Almeida	http://www.flickr.com/photos/68306118@N00
Gustavo Madico	http://www.flickr.com/photos/desdegus
Douglas Fernandes	http://www.flickr.com/photos/thejourney1972
James Preston	http://www.flickr.com/photos/jamespreston
Rodrigo Della Fávera	http://www.flickr.com/photos/rodrigofavera
Dinesh Rao	http://www.flickr.com/photos/dinrao
Marina Campos Vinhal	http://www.flickr.com/photos/marinacvinhal
Jorge Gobbi	http://www.flickr.com/photos/morrissey
Steve Taylor	http://www.flickr.com/photostheboywiththethorninhisside/

Finally would like also to acknowledge the users who wrote the blog pharagraphs were used and reproduced here. These texts can be found at the folloing adresses:

http://realtravel.com/cuzco-journals-j1879736.html
http://realtravel.com/machu_picchu-journals-j5181463.html
http://realtravel.com/rio-journals-j4669810.html
http://www.travelpod.com/travel-blog-entries/sarah_s_america/south_america/1140114720/tpod.html
http://www.travelpod.com/travel-blog-entries/rachel_john/roundtheworld/1146006300/tpod.html
http://www.travelpod.com/travel-blog-entries/eatdessertfirst/world_tour_05/1160411340/tpod.html
http://www.travelpod.com/travel-blog-entries/idarich/rtw_2005/1140476400/tpod.html
http://www.travelpod.com/travel-blog-entries/twittg/rtw/1132765860/tpod.html
http://www.travelpod.com/travel-blog-entries/emanddave/worldtrip2006/1155492420/tpod.html

References

1. J. Ah-Pine, C. Cifarelli, S. Clinchant, G.Csurka, and J. Renders. Xrce's participation to imageclefphoto 2008. In *Working Notes of the 2008 CLEF Workshop*, 2008.
2. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3, 2003.
3. D. Blei, Michael, and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003.
4. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.
5. P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.

6. Y.-C. Chang and H.-H. Chen. Approaches of using a word-image ontology and an annotated image corpus as intermedia for cross-language image retrieval. In *CLEF 2006 Working Notes*, 2006.
7. S. Clinchant, J. Renders, and G. Csurka. Xrce's participation to imageclefphoto 2007. In *Working Notes of the 2007 CLEF Workshop*, 2007. http://clef.isti.cnr.it/2007/working_notes/CLEF2007WN-Contents.html.
8. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.
9. M. Dowman, V. Tablan, H. Cunningham, and B. Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In *Proc. of the 14th International World Wide Web Conference*, 2005.
10. P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation :learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.
11. S. Feng, V. Lavrenko, and R. Manmatha. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.
12. Flickr. http://www.flickr.com, 2007.
13. Footstops. http://footstops.com/, 2007.
14. M. Grubinger, P. Clough, A. Hanbury, and H. Müller. Overview of the ImageCLEFphoto 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, 2007. http://www.clef-campaign.org/2007/working_notes/CLEF2007WN-Contents.html.
15. G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. Khudanpur, D. Klakow, M. Krause, R. Manmatha, h. Nock, D. Petkova, B. Pytlik, and P. Virga. Joint visual-text modeling for automatic retrieval of multimedia documents. In *Proceedings of ACM Multimedia*, 2005.
16. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, 2003.
17. V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.
18. J. Li and J. Z. Wang. Alip: The automatic linguistic indexing of pictures system. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 1208–1209, Washington, DC, USA, 2005. IEEE Computer Society.
19. L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic object picture collection via incremental model learning. In *CVPR*, 2007.
20. X. Li, L. Chen, L. Zhang, F. Lin, and W. ying Ma. Image annotation by large-scale content-based image retrieval. In *Proc. of the 14th Annual ACM international Conference on Multimedia (MM06)*, 2006.
21. N. Maillot, J.-P. Chevallet, V. Valea, and J. H. Lim. Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval. In *CLEF 2006 Working Notes*, 2006.
22. F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: Constraining the latent space. In *ACM MM*, 2004.
23. Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *In MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
24. J. Pan, H. Yang, C. Faloutsos, and P. Duygulu. Gcap: Graph-based automatic image captioning. In *CVPR Workshop on Multimedia Data and Document Engineering*, 2004.
25. F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
26. A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *CVPR*, 2007.
27. Realtravel. http://realtravel.com/, 2007.
28. J. J. Rocchio. Relevance feedback in information retrieval. *The SMART Retrieval System - Experiments in Automatic Document Processing, Salton Gerard*, 1971.
29. T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006.
30. Travbuddy. http://www.travbuddy.com/, 2007.
31. Travelblog. http://www.travelblog.org/, 2007.
32. Travellerspoint. http://www.travellerspoint.com, 2007.
33. Travelpod. http://www.travelpod.com/, 2007.
34. Trippert. http://trippert.com/, 2007.

35. A. Vinokourov, D. R. Hardoon, and J. Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. In *Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, 2003.
36. X. Wang, L. Zhang, and W.-Y. M. F. Jing. Annosearch: Image auto-annotation by search. In *CVPR*, 2006.
37. K. Yanai and K. Barnard. Probabilistic web image gathering. In *Proc. of ACM Multimedia Workshop on Multimedia Information Retrieval (MIR05)*, 2005.
38. C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, 2001.