A Multimodal Approach to Exploit Similarity in Documents

Matteo Cristani and Claudio Tomazzoli

University of Verona
{matteo.cristani,claudio.tomazzoli}@univr.it

Abstract. Automated document classification process extracts information with a systematic analysis of the content of documents.

This is an active research field of growing importance due to the large amount of electronic documents produced in the world wide web and available thanks to diffused technologies including mobile ones.

Several application areas benefit from automated document classification, including document archiving, invoice processing in business environments, press releases and research engines.

Current tools classify or "tag" either text or images separately.In this paper we show how, by linking image and text-based contents together, a technology improves fundamental document management tasks like retrieving information from a database or automated documents.

We present an investigation of a model of conceptual spaces for investigation using joint information sources from the text and the images forming complex documents. We present a formal model and the computable algorithms and the dataset from which we took a subset to make experiments and relative tests and results.

1 Introduction

Nowadays the wide availability of electronic documents through the Internet or private business networks has changed the way people search for information. We deal with a huge quantity of knowledge which has to be organized and searchable to be utilized. Also for this reason in information technology research community there is an always growing interest in the field of automatic document classification. Although several innovative studies are produced every year, some topics are still to be deeply investigated. Among these, the problem of efficient classification and retrieval of documents containing both text and images has been treated in a non multidisciplinary approach.

There are several publications of efficient information retrieval from text. There are also publications about information extraction from images and even text contained in images [1], but the joint analysis of text and image information from a complex document still lacks a well documented solution. For example, if a brochure from an isolated hotel in the Dolomites describes the hotel's features and includes maps and pictures of mountainous surroundings, the categorizer

A. Moonis et al. (Eds.): IEA/AIE 2014, Part I, LNAI 8481, pp. 490–499, 2014.

will automatically discover the content and link the text and the images together. Then someone searching for an isolated mountain lodge within a certain price range would retrieve the brochure even if "isolated lodge in the mountains" were never mentioned in the actual text.The paper is organized as follows: Section 2 presents the model and the approach of extracting joint textual and image information; Section 3 presents the framework and the computable algorithms; Section 4 shows experiments and related results and finally in Section 5 we make some conclusions.

2 The Model

We found the model described in [2] as a valuable starting point for our model; we will use accordingly the term "mode" of a document for both text and image and we will use the "bag of word" representation for features set of both modes, but we define those contributes in a more general sense than in [2].Then we apply "noise" as suggested in [3] and we define our general model, which will be used later in the framework. Suppose we have *n documents* $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ and *m tags* $\mathcal{T} = \{t_1, t_2, ..., t_m\}$. Following PLSA[1] approach, The goal is to construct a set of feature vectors $\{X_1, X_2, ..., X_n\}$ in a latent semantic space $\mathbb{R}^k$ to represent these multimedia objects having matrix $A = U\Sigma V^T$.

The Model for Multimodal Documents
We are considering both textual and visual contributions to the meaning of a document. Suppose we are given a matrix Q of content links, where $Q_{i,j}$ can represent the similarity measurement between the ith document and the jth document. Recalling the works in latest literature we have that documents can be described as *multimodal* when made of both text and visual content, each defined as "mode"; a repository that contains a set of multimodal documents is then $D = \{d_1, d_2, ..., d_n\}$. We will use these "dual" representations to compute the trans-media similarity measures as defined in [2] : $sim_{glob}(d_i, d_j) = Q_{i,j}$ using a linear combination where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are appropriate weighting factors:

$$\lambda_1 sim_{TT}(d_i, d_j) + \lambda_2 sim_{TV}(d_i, d_j) + \lambda_3 sim_{VT}(d_i, d_j) + \lambda_4 sim_{VV}(d_i, d_j)$$

so we have that the elements in our matrix Q of similarity of multimodal content of documents can be

$$Q_{i,j} = \lambda_1 sim_{TT}(d_i, d_j) + \lambda_2 sim_{TV}(d_i, d_j) + \lambda_3 sim_{VT}(d_i, d_j) + \lambda_4 sim_{VV}(d_i, d_j) \tag{1}$$

We assume that the documents with stronger links ought to be closer to each other in the latent semantic space.

[1] Probabilistic Latent Semantic Analisys

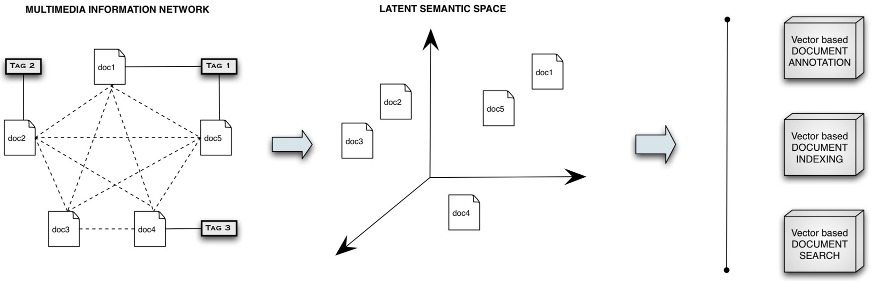
MULTIMEDIA INFORMATION NETWORK
Tag 2
doc1
Tag 1
doc2
doc5
doc3
doc4
Tag 3
LATENT SEMANTIC SPACE
doc1
doc2
doc5
doc3
doc4
Vector based DOCUMENT ANNOTATION
Vector based DOCUMENT INDEXING
Vector based DOCUMENT SEARCH

Fig. 1. Model for Multimodal Documents

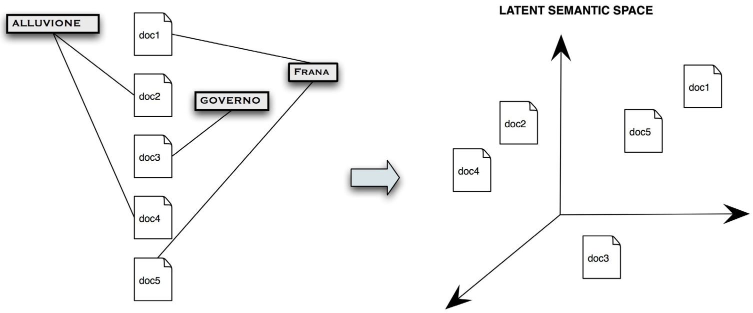

Fig. 2. documents with stronger links will be closer

Based on this assumption, we introduce the quantity Ω to measure the smoothness of documents in the underlying latent space, as in [3].

$$\Omega(X) = \frac{1}{2} \sum_{i,j=1}^{n} Q_{i,j}(X_i - X_j)(X_i - X_j)^T$$

where, $\|M\|_2^2$ is the l_2 norm of matrix M , and X_i and X_j are the ith and jth row of X. It is easy to see that by minimizing the above regularization term, a pair of documents with larger $Q_{i,j}$ will have closer feature vectors X_i and X_j in the latent space. Given D as the diagonal matrix with its elements as the sum of each row of Q and $L = D - Q$, with some matrix operations we obtain $\Omega(X) = trace(X^T LX)$ using the factorization $X = U\Sigma^k$, defining H as $H = U\Sigma_k V^T = XV^T$ and knowing that $VV^T = I$ we have

$$\Omega(X) = trace(H^T LH) \tag{2}$$

2.1 The Noisy Model

Due to the fact that we consider both textual and visual contribution to the meaning of a document, we have to consider the existence of noise in process so a noise term ε exist on the matrix Q such that

$$Q = H + \varepsilon$$

where H is the matrix which denotes the noise-free tag links, after the noise ε has been removed. The goal is to make a correctly representative H of "minimal rank". Let H_i, $1 \leq i \leq n$ denote the row vectors of H, which is the associated noise-free tag vector for the ith document.

Each tag vector represents the occurrence of the corresponding tag in the document corpus. Tag vectors of synonyms should be the same (or within a positive multiplier of one another), such as the tag vector for the synonym terms "person" and "human". Moreover, many tags probably do not independently occur in the corpus since they are semantically correlated. For example, the tag 'animal' often correlates with its subclasses such as "cat" and "tiger". This indicates from the viewpoint of linear algebra, that the tag vector of "animal" could be located in a latent subspace spanned by those of its subclasses. Since the rank of matrix H is the maximum number of independent row vectors, it follows from the above dependency among tags, that H ought to have a low rank structure. The topic vectors that represent occurrences of the associated topics in the document corpus span a latent semantic space, which contains most of tag vectors. Therefore, the rank of H should be no more than the maximum number of independent topic vectors in the latent space. Hence we can impose a low-rank prior to estimate the noise-free H from the observed noisy Q.

The *nuclear norm* of a matrix M is computed as the sum of all the singular values of the matrix. Let $\|M\|_*$ denote the nuclear norm of matrix M, then

$$\|M\|_* = \sum_i \sigma_i(M)$$

where $\sigma_i(M)$ are singular values in M.The *nuclear norm* can be used to solve the optimization problem of determining the lowest rank approximation.

So our problem can be described as finding

$$min\|Q - H\|_F + \gamma\|H\|_* \tag{3}$$

where $\|M\|_F$ is the squared summation of all elements in a matrix M (the Frobenius norm) and γ is a balancing parameter.

This problem has a unique analytical solution: recalling from above that $Q = U\Sigma V^T$ we define an operation $(x)_+ = max(0, x)$ and the matrix Σ_+^γ which is the diagonal matrix from Σ where the singular values are defines as $(\sigma - \frac{\gamma}{2})_+$. with values σ from matrix Σ.

$$\Sigma_+^\gamma = diag((\sigma - \frac{\gamma}{2})_+)$$

The solution to the problem can be described as

$$min\|Q - H\|_F + \gamma\|H\|_* = H_\gamma = U\Sigma_+^\gamma V^T \tag{4}$$

The difference with normal *Latent Semantic Indexing* is that it directly selects the largest k singular values of A where this Formulation subtracts something $(\frac{\gamma}{2})$ from each singular value and thresholds them by 0. Suppose the resulting noise free matrix H is of rank k, then the Support Vector Machine of H has form as $H = U\Sigma_k V^T$ where Σ_k is a $k \times k$ diagonal matrix. Similar with *Latent Semantic Indexing*, the row vectors of $X = U\Sigma_k$ can be used as the latent vector representations of documents in latent space. It is also worth noting that minimizing the rank of H gives a smaller k so that the obtained latent vector space can have lower dimensionality, and then the storage and computation in this space could be more efficient.

2.2 The Global Model for Multimodal Documents

Considering both contribution to the model we can make use of both Equation 2 and 3 so our problem can be completely described as finding

$$min\|Q - H\|_F + \gamma\|H\|_* + \lambda trace(H^T LH) \tag{5}$$

Here λ is another balancing parameter. In contrast to Formulation (3), Formulation (5) does not have an closed-form solution. Fortunately, this problem can be solved by the *Proximal Gradient method* known from literature which uses a sequence of quadratic approximations of the objective function in order to derive the optimal solution.

3 The Framework

3.1 The Matrix Q of Similarity for Multimodal Content

We are considering both textual and visual contributions to the meaning of a document.

We define matrix Q_t of content links, where $Q_t(i,j)$ can represent the similarity measurement between the text of the ith document and the the text of the jth document.

We define matrix Q_v of content links, where $Q_v(i,j)$ can represent the similarity measurement between the image of the ith document and the the image of the jth document.

Following PLSA approach as above specified we have $Q_t \cong U_t\Sigma_t^k V_t^T$ for the textual mode and $Q_v \cong U_v\Sigma_v^k V_v^T$ for the visual mode.

We have $S_T = U_t\Sigma_t^k$; similarly, the visual, dual representation of the textual part is: $S_V = U_v\Sigma_v^k$ We will denote the textual part of d_j by $S_T(d_j)$ and its visual part $S_V(d_j)$ which are the jth rows of matrix S_T and S_V. Recalling that in *Basic Model for Multimodal Documents* in equation 1 we had

$$Q_{i,j} = \lambda_1 sim_{TT}(d_i, d_j) + \lambda_2 sim_{TV}(d_i, d_j) + \lambda_3 sim_{VT}(d_i, d_j) + \lambda_4 sim_{VV}(d_i, d_j)$$

And *assuming that the both text and image part of a document shall define the same meaning for the document in the meaning space* we will use these partial latent semantic representations to define the single components of the equation above

$$sim_{TV}(d_i, d_j) = \|S_T(d_i) - S_V(d_j)\|_F \tag{6}$$

$$sim_{VT}(d_i, d_j) = \|S_V(d_i) - S_T(d_j)\|_F \tag{7}$$

$$sim_{TT}(d_i, d_j) = \|S_T(d_i) - S_T(d_j)\|_F \tag{8}$$

$$sim_{VV}(d_i, d_j) = \|S_V(d_i) - S_V(d_j)\|_F \tag{9}$$

This model benefits from two major aspects: it is simple to understand and it is simple to implement, both because it involves only measure of distance in a vector space.

The main assumption is that there is *one* meaning space so that features in text and features in images all refers to a set of concepts or meanings which are the same but are expressed with words and with images.

When these meanings are expressed with words the dimensionality of the feature space is different than the dimensionality of the feature space coming from the images, but using a dimensionality reduction algorithm we can reduce these different dimensions to be the same, so that we can than compute a distance.

3.2 Algorithms

First algorithm is to be used to evaluate the vector representation of text features, which are extracted from the words in text.

Algorithm 1. S_t: finds the vectorial representation of the features on texts

```
Input: A finite set $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ of texts
Output: The Matrix $S_t$ and the integer $m$, size of the textual vocabulary TW
$Q_t \leftarrow \emptyset$ ; $TW \leftarrow \emptyset$ ;
for $i \leftarrow 1$ to $n$ do
    for $j \leftarrow 1$ to $countword(t_i)$ do
        $TW \leftarrow TW \cup \{w_j\}$ ;
for $i \leftarrow 1$ to $n$ do
    $Q_t(i) \leftarrow \frac{1}{countword(t_i)}\{numword_{1,i}, \ldots numword_{m,i}\}$ ;
Compute $U_t$, $\Sigma_t^k$, $V_t$ as $Q_t \cong U_t \Sigma_t^k V_t^T$;
$S_t \leftarrow U_t \Sigma_t^k$;
return $S_t$,$m$;
```

Second algorithm is to be used to evaluate the vector representation of visual features, using a bag of visual word model generated computing SIFT on images.

Algorithm 2. S_v: finds the vectorial representation of the features on images

```
Input: A finite set $V = \{v_1, v_2, \dots, v_n\}$ of images and the integer $m$, size
       of the visual vocabulary
Output: The Matrix $S_v$
$Q_v \leftarrow \emptyset$; $FV \leftarrow \emptyset$ ;
for $i \leftarrow 1$ to $n$ do
    $FV(i) \leftarrow$ extract SIFT features from $v_i$ ;
Compute $Q_v$ as a Bag of Visual Words Model using FV as a feature
vector and GMM with $m$ Visual Terms;
Compute $U_v$, $\Sigma_v^k$, $V_v$ as $Q_v \cong U_v \Sigma_v^k V_v^T$;
$S_v \leftarrow U_v \Sigma_v^k$;
return $S_v$;
```

Third algorithm deals with the problem of combining the two sources of information a single matrix.

Last algorithm is the classical algorithm for *Proximal Gradient Approximation* applied to our matrix problem.

Algorithm 3. Q: finds the Matrix Q of similarity of multimodal content

```
Input: A finite set $D = \{d_1, d_2, \dots, d_n\}$ of documents and numeric weight
       factors $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$
Output: The Matrix Q as in equation ??
$V \leftarrow \emptyset$ ; $T \leftarrow \emptyset$ ; $S_v \leftarrow \emptyset$ ; $S_t \leftarrow \emptyset$;
for $i \leftarrow 1$ to $n$ do
    $T(i) \leftarrow$ extract the text in $d_i$;
    $V(i) \leftarrow$ extract the image in $d_i$;
Compute $S_t$ and $m$ using algorithm 1 with input $T$;
Compute $S_v$ using algorithm 2 with input $V$ and $m$;
Compute $sim_{TV}(d_i, d_j)$ using equation 6;
Compute $sim_{VT}(d_i, d_j)$ using equation 7;
Compute $sim_{TT}(d_i, d_j)$ using equation 8;
Compute $sim_{VV}(d_i, d_j)$ using equation 9;
for $j \leftarrow 1$ to $n$ do
    for $i \leftarrow 1$ to $n$ do
        $Q(i, j) \leftarrow$
        $\lambda_1 sim_{TT}(d_i, d_j) + \lambda_2 sim_{TV}(d_i, d_j) + \lambda_3 sim_{VT}(d_i, d_j) + \lambda_4 sim_{VV}(d_i, d_j)$
return $Q$;
```

Algorithm 4. Proximal Gradient for minimizing equation 5

```
Input: The Matrix Q of similarity of multimodal content and numeric weight
       factors $\lambda$, $\gamma$, $\epsilon$, $maxIter$
Output: The Matrix $H$ denoised matrix of $Q$
$H_0 \leftarrow 0$ ; $\tau \leftarrow 1$
Compute $\sigma_{max}(I + \lambda L^T)$ as the largest singular value of $(I + \lambda L^T)$
$\alpha \leftarrow 2\sigma_{max}(I + \lambda L^T)$
repeat
    $K(H_{\tau-1}) \leftarrow \|Q - H_{\tau-1}\|_F + \lambda trace(H^T_{\tau-1} L H_{\tau-1})$
    $G_\tau \leftarrow H_{\tau-1} - \frac{1}{\alpha}\nabla K(H_{\tau-1}) = H_{\tau-1} - \frac{2}{\alpha}(H_{\tau-1} - Q + \lambda L^T H_{\tau-1})$
    Compute $diag(\sigma)$ as singular value decomposition $G_\tau = Udiag(\sigma)V^T$ and
    remember $(x)_+ = max(0, x)$
    $H_\tau \leftarrow Udiag((\sigma - \frac{\gamma}{\alpha})_+)V^T$
    $\tau \leftarrow \tau + 1$
until
Convergence $(rank(H_\tau) - rank(H_{\tau-1}) \le \epsilon)$ or maximum iteration number
($maxIter$) achieved
return $H_\tau$
```

4 Experimental Results

The test dataset is made of almost 10 years of daily issues of a local newspaper; actually under a non disclosure agreement. It is made of almost 10 years of daily issues of a local newspaper, each having 64 pages with an average of 4 articles per page.

This sums to around 800.000 documents most but not all of which have both text and image contributes. We decided to concentrate for the first experiments on a subset of articles of which we knew the tag "author" as well the others such as "topic" because in the dataset this tag was the one with less occurrences, so that we had a subset of maximum size with all tags.

In local newspapers articles are organized in a way so that page groups and hi level topics can be considered synonymous; for instance "economy" is found at pages 6,7 and "news story" at pages 9,10.

This organization is almost stable throughout the time span considered, almost 10 years of daily editions.

Authorship Attribution

We decided to consider "signed" articles and try to apply our algorithm to the problem of *authorship attribution* having around 40.000 documents and 300 Authors.

We pre-elaborated the articles so that they can be made object to a Matlab set of codes.After a few unsuccessful attempts we decided to pre-define the final dimension of the SVD dimensionality reduction algorithm.

As the resources consumed by the extraction of features from images are much greater than the ones used for text elaboration, we decided that the leading factor should be the dimension of the image features matrix.

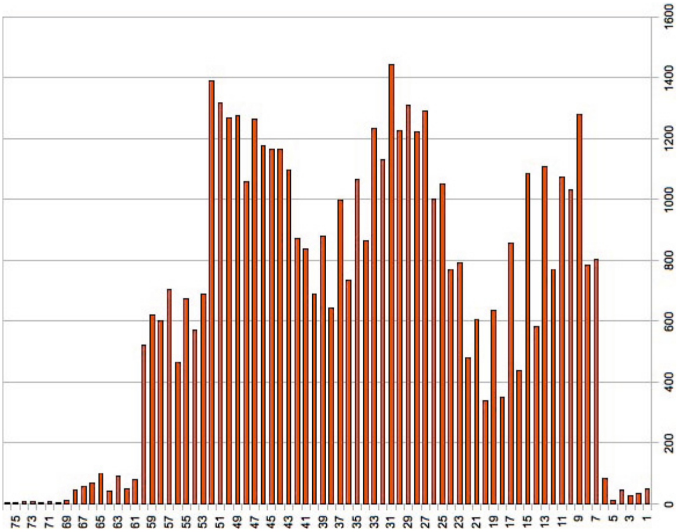

Fig. 3. *Signed* Article Distribution per page

Given that, we applied the code of algorithms 1, 2, 3, 4 obtaining results as in schema below

Information Source	*Accuracy*
Text	0,18
Images	0,24
Text and Images	0,32

Classification
We then decided to try our algorithm in guessing the number of the page the article was in, roughly indicating the topic.

Also in this case we found that a better result can be achieved, in fact *results are better than chance (0.0014)* and *accuracy* and *precision* are close to the ones of text only, but *recall* is higher than the one for text only, while compared to images only *accuracy* is higher and *precision* is close but *recall* is lower.

Information Source	*Accuracy*	*F1*
Text	0.0907	0.0165
Images	0.0768	0.0725
Text and Images	0.0881	0.0428

5 Conclusions

The first part of this work was dedicated to point out the overview of the research and the problems and choices we got through during the path of this study. Then we focused on the model we would use to determine different contribution to classification of the text and image information of a document; we've given the details of the definition of a meaning space using Probabilistic Latent Semantics for multimodal documents including consideration and modeling of the possible noise that shall be considered in this process and how to deal with it. Then we focused on the definitions of the distances in the meaning space and we've given the definition of computable algorithm for our model. Part of this work was then dedicated to the description of the test dataset and its possible use and organization to be useful in our research. We used it in experiments to validate the model and we presented these experiments and related results.

References

1. Ye, Q., Huang, Q., Gao, W., Zhao, D.: Fast and robust text detection in images and video frames. Image and Vision Computing 23, 565–576 (2005)
2. 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 and Applications 42, 31–56 (2009)
3. Qi, C., Aggarwal, G., Tian, Q., Ji, H., Huang, T.: Exploring context and content links in social media: A latent space method. IEEE Transactions Pattern Analysis and Machine Intelligence (August 2011)
4. Kesorn, S., Poslad, K.: An enhanced bag of visual word vector space model to represent visual content in athletics images. IEEE Transactions on Multimedia (October 2011)
5. Denoyer, L., Gallinari, P.: Bayesian network model for semi-structured document classification. Information Processing and Management 40, 807–827 (2004)
6. Bouguila, N., ElGuebaly, W.: Discrete data clustering using finite mixture models. Pattern Recognition 42, 33–42 (2009)
7. Mikhailov, D.V., Emelyanov, G.M.: Semantic clustering and affinity measure of subject-oriented language texts. Pattern Recognition and Image Analysis 20, 376–385 (2010)
8. Yang, L., Geng, Y., Cai, B., Hanjalic, A.: Object retrieval using visual query context. IEEE Transactions on Multimedia (July 2011)
9. Qin, J., Yung, N.H.C.: Scene categorization via contextual visual words. Pattern Recognition 43, 1874–1888 (2010)
10. Park, G., Baek, Y., Lee, H.-K.: Web image retrieval using majority-based ranking approach. Multimedia Tools and Applications 31, 195–219 (2006)
11. Chan, W., Coghill, G.: Text analysis using local energy. Pattern Recognition 34, 2523–2532 (2001)
12. Aronovich, L., Spiegler, I.: Cm-tree: A dynamic clustered index for similarity search in metric databases. Data & Knowledge Engineering 63, 919–946 (2007)
13. Sable, C.L., Hatzivassiloglou, V.: Text-based approaches for non-topical image categorization. International Journal on Digital Libraries 3, 261–275 (2000)