Bioingenium at ImageClefmed 2010: A Latent Semantic Approach

Jose G Moreno, Juan C Caicedo, and Fabio A González

National University of Colombia,
Bioingenium Research Group
{jgmorenofr,jccaicedoru,fagonza}@unal.edu.co
http://www.bioingenium.unal.edu.co

Abstract. This paper describes the participation of the Bioingenium Research Group in the ad hoc Medical Image Retrieval task for the 2010 ImageCLEF forum. The work aimed to explore semantic relationships in textual information and transfer into visual information by building a unified image searching index. The proposed strategy is based on the use of Non-Negative Matrix Factorization to decompose matrix data and build latent semantic spaces.

Keywords: Non-Negative Matrix Factorization, Content-Based Information Retrieval

1 Introduction

The Medical Image Retrieval task of ImageCLEF aims to evaluate computational methods to access and retrieve visual contents for medical applications. For the 2010 forum, the challenge consists of proposing computational alternatives for retrieving images from a database containing 77,000 images extracted from medical papers and 16 query topics [1]. This paper describes the approach followed by the Bioingenium Research Group at the National University of Colombia, which focused on combining visual features extracted from images together with text captions taken from the manuscript, to build a unified index for searching medical images. Latent Semantic Indexing (LSI) strategies are used at the core of our approach to model implicit relationships between visual and textual data.

Since our image indexing method integrates visual and textual information, we are able to map different types of queries to the latent semantic space so as to search for images. Be it a text query, an image example or a mixed query, the system uses the same index to retrieve relevant images. To achieve this, we use Non-Negative Matrix Factorization algorithms to build latent semantic spaces using multimodal information. We paid special attention to the case in which a visual example is provided as query to search the multimodal index. When only visual features are used to match contents in the collection, the multimodal index still has information taken from textual data, which can influence the search results.

These algorithms can be computationally expensive and therefore it might take too long for them to process large image collections. For this reason, they had to be suited for running experiments in the ImageCLEFmed 2010 collection by using a structured initialization strategy based on Singular Value Decomposition, which allowed the algorithms to converge faster to the desired factorization. We run experiments that involved different configurations of our model to evaluate their performance.

The contents on this paper are organized as follow: Section 2 summarizes the methods used to process visual and textual data separately. Section 3 presents the main methods of our retrieval framework. Section 4 presents the experimental setup, the results and discussions. Finally, the paper ends with concluding remarks in Section 5.

2 Data Processing

In our approach, textual data and visual data are first processed in an independent manner. The purpose of this preprocessing step is to build two matrices that represent the content of each modality for all images in the collection by using a certain set of features.

2.1 Text Processing

We used the paper title and image caption as semantic context for each image. The Natural Language Toolkit (NLTK) [8] was used to build a vector space representation of textual information. Common text processing techniques, such as stop words removal and word stemming, were applied to the corpus and a TF-IDF weighted scheme was used as final text representation. Some terms were removed from the vector space to generate a more compact representation by pruning terms with too low or too high frequency within the corpus.

2.2 Visual Processing

We used a Bag of Features strategy in which each image is represented as a histogram of frequencies of predefined visual patterns [9]. These visual patterns are organized in a codebook of DCT features [10] that is quantized using the k-means algorithm. Blocks of 8x8 pixels are first taken from a regular grid in each image in the collection and processed using the Discrete Cosine Transform in the three RGB color channels. The coefficients of these transform are used as features to construct the codebook and to match visual patterns when building the histogram representation. The size of the codebook was set to 2,000 and 5,000 features in our experiments.

3 Retrieval Framework

Latent semantic strategies have been shown to be a powerful set of methods to find latent relationships between features in document collections. Using a term-document representation, it is possible to find latent patterns such as highly correlated terms through matrix decompositions [6]. A Singular Value Decomposition (SVD) of the term-document matrix is used in information retrieval to construct latent semantic spaces so that documents that refer to certain topics can be highly scored even when there is not an explicit occurrence of the query terms. This is possible thanks to the implicit relationships existing between terms that are found during the indexing process. This technique usually shows improved retrieval performance compared to a document retrieval engine that only uses the term frequencies to score documents.

We use Non-negative Matrix Factorization as a flexible algorithm to model a latent semantic space, which correlates visual features and text terms in a multimodal collection.

3.1 Input data

The term-document matrix X_t of size mxn is used to represent text data in the collection, where m is the number of images and n is the number of terms. The values in each cell are the frequency of the corresponding term for an image. Similarly, the feature-document matrix X_v is a matrix of size mxp, with m being the number of images and p the number of visual features.

3.2 Non-negative Matrix Factorization (NMF)

In document clustering as well as in document retrieval and document classification, an important task is to recognize the semantic relationships existing between terms in a collection. NMF is a novel technique proposed to address problems of this nature and has been actively used for the analysis of text documents [12] as well as image collections [4][5][7]. This method decomposes a non-negative matrix into two lower rank non-negative matrices; the first one encoding the basis of the latent space and the second one encoding the coefficients of the document representation in that latent space. NMF is modeled as an optimization problem with non-negative restrictions on both matrix factors for which different objective functions can be used. We used the Divergence objective function to find the factorization [13]:

$$\sum_{i,j} X_{ij} \log \frac{X_{ij}}{WH_{ij}} - X_{ij} + WH_{ij} \tag{1}$$

3.3 Singular Value Decomposition (SVD) Initialization

NMF is a powerful matrix decomposition strategy but it has an important problem to consider: its convergence performance is slow. In order to make it usable

for a large-scale image retrieval application, we applied an initialization algorithm based on SVD operations [11].

SVD is a very common matrix decomposition strategy, in which three matrices are obtained: two orthonormal matrices and a diagonal matrix. The orthonormal matrices have the particularity of being linearly independent, but the can still have negative values. Based on an interesting result about the increase of the rank unit matrix when the negative values are changed to zero, a method that allows getting a non-negative SVD decomposition is used to build W and H initial values. This strategy is known as Double Non-Negative Singular Value Decomposition [11] and its use speeds up the indexing process for about 3 times.

3.4 Latent Representation

We model the factorization problem using an asymmetric algorithm as it is described in [2]. In this method, the text matrix X_t is first exploited to build the basis of the latent semantic space and then the matrix X_v is used to adapt the visual representation to such latent semantic space. The algorithm follows two basic steps:

1. Learning the semantic basis: NMF is applied to solve $X_t = W_t H_t$, where X_t is the text matrix.
2. Adapting a visual basis: A modified version of NMF is applied to find $X_v = W_v H_t$ in order to learn only a W_v that spans the same latent space obtained from the text analysis, but using visual features.

Since the basis of the latent space is formed using textual data, it is expected that this representation helps reduce the semantic gap in the CBIR system. This approach is meant to use the same semantic representation found in text decomposition to calculate the corresponding basis that satisfies the NMF decomposition of the visual data. With the new W_v basis, images lacking text can be mapped into the semantic space in the same way as text documents are, but notice that only visual features are required.

Indexing texts or projecting text queries to the semantic space is straightforward because a W_t basis has been also learned. When queries have both modalities available, an automatic strategy to combine both is used. The W_t and W_v basis are simply concatenated to allow a mixed projection of multimodal data.

4 Experiments and Results

We participated with three different approaches that are classified according to the information used in the query: only text terms, only visual features and mixed text-visual information. To tune the model parameters, we used the queries from the 2009 ImageCLEFmed challenge and tried solving them on the 2010 collection. Our goal was to maximize the Mean Average Precision (MAP) according to the ground truth for the 2009 queries and then use the same configuration to solve the 2010 topics.

The main parameter in our model is the size of the semantic space, which determines the number of latent concepts. We calculated various semantic spaces using different sizes to evaluate the performance with different configurations. In order to explore the search space before submitting our runs, we used a logarithmic scale to set the latent space size parameter.

Below is a description of our 6 submitted runs:

- $Text - k = 2^{11}$: only text information was used in the queries as well as in the collection. In this experiment we used 2,048 as the size of the latent semantic space.
- $AsymmetricMixed - k = 2^{11}$: both visual and text information is used to process the queries. The size of the semantic space was the same as in the previous experiment.both visual and text information is used to process the queries. The size of the semantic space was the same as in the previous experiment.

The following runs only used visual information to query the system:

- $AsymmetricDCT2000 - k = 2^5$: a codebook of 2,000 visual patterns was used to represent image features. The size of the latent semantic space was set to $2^5 = 32$.
- $AsymmetricDCT5000 - k = 2^5$: a codebook with 5,000 elements and $2^5 = 32$ latent semantic dimensions.
- $AsymmetricDCT5000 - k = 2^7$: a codebook with 5,000 elements and $2^7 = 128$ latent semantic dimensions.
- $AsymmetricDCT5000 - k = 2^{7.5}$: a codebook with 5,000 elements and $2^{7.5} = 181$ latent semantic dimensions.

We provide a comparison of the results obtained with the 2009 and 2010 queries, even though the ground truths were obtained by analyzing different versions of the same collection. In particular, the number of images in the collection when the ground truth of 2009 was built was about 10% smaller than the 2010 version. However, this may be considered as a good estimation of the expected performance.

We observed that when the size of the latent space is increased, a better performance is obtained using text queries. Table 1 shows the MAP and precision in 1000 (P1000) obtained using the 2009 and 2010 queries, respectively, with the corresponding configurations for each run. Using only visual information on the 2009 queries, the configuration described for $AsymDCT2000 - k = 2^5$ (2,000 visual features and 2^5 latent semantic dimensions) obtained the best performance. This result is even a very good score compared to the results obtained by participants in the 2009challenge when using only visual information [3]. However, in the 2010 challenge the results were not as encouraging.

The performance results shown in Table 1 show consistency between the results obtained for the 2009 and 2010 queries, both for MAP and Precision at 1,000 results (P1000). The search strategy based only on text information outperformed both the visual and mixed strategies. However, it is well known that only visual information lacks of semantic information to search for images and this gap still remains an open problem for image retrieval. Our method is

Table 1. MAP and P1000 values for 1 Text, 1 Mixed and 4 Visual queries.

		Text	*AsymMixed*	*AsymDCT*			
		2^{11}	$5000 - 2^{11}$	$2000 - 2^{5}$	$5000 - 2^{5}$	$5000 - 2^{7}$	$5000 - 2^{7.5}$
2009	MAP	0.1426	0.0655	0.0127	0.0057	0.0039	0.0018
	P1000	0.0593	0.0437	0.0150	0.0082	0.0070	0.0058
2010	MAP	0.1005	0.0395	0.0015	0.0014	0.0018	0.0000
	P1000	0.0367	0.0203	0.0036	0.0033	0.0025	0.0001

an attempt to bridge the gap by representing visual information together with text data in order to bring some semantic structure to the representation space. Our results did not diverge too much from those obtained by other competitors in the challenge.

The mixed approach also showed a better performance compare to the sole use of visual features, but it is still not as good as using only text, even though more information is provided to the search engine with this approach. These results suggest that visual features are not contributing to finding more relevant images, but are instead reducing the number of retrieved images.

5 Concluding Remarks

This paper summarizes an exploratory work in semantic representations by which the Bioingenium Research Group participates in ImageCLEF2010. The main goal of this strategy was to build a unified indexing method to allow representing simultaneously visual information as well as textual information using NMF algorithms. The modeling of latent semantic spaces was followed to approximate hidden relationships between visual features and text terms.

To make this strategy feasible for real world image collections, we employed a structured initialization of the NMF algorithm is based on non-negative decompositions of large matrix data to help speed up the indexing time. This allowed us to process the ImageCLEF2010 collection of 77,000 images with their corresponding textual data and conduct experiments using the ground truth of 2009 and submit runs for the 2010 topics.

References

1. Henning Mller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Charles E. Kahn Jr., and William Hersh.: Overview of the CLEF 2010 medical image retrieval track. In the Working Notes of CLEF 2010, Padova, Italy, 2010.
2. González, Fabio A. and Caicedo, Juan C. and Nasraoui, Olfa and Ben-Abdallah, Jaafar: NMF-based multimodal image indexing for querying by visual example. CIVR '10: Proceedings of the ACM International Conference on Image and Video Retrieval, 2010.

3. Henning Mller , Jayashree KalpathyCramer , Ivan Eggel , Steven Bedrick , Sad Radhouani , Brian Bakke , Charles E. Kahn Jr. , William Hersh: Overview of the CLEF 2009 medical image retrieval track. In the Working Notes of CLEF 2009.
4. Lee, Daniel D. and Seung, H. Sebastian: Learning the parts of objects by non-negative matrix factorization. Nature, 1999.
5. Cooper, M. and Foote, J.: Summarizing video using non-negative similarity matrix factorization. IEEE Workshop on Multimedia Signal Processing, 2002.
6. Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2, 2008.
7. W. Liu, N. Zheng, and X. Lu. Non-negative matrix factorization for visual coding. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2003.
8. Bird, S. 2006. NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on interactive Presentation Sessions (Sydney, Australia, July 17 - 18, 2006). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 69-72.
9. E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. European Conference on Computer Vision. pages 490–503. 2006.
10. J. S. Hare, S. Samangooei, P. H. Lewis, and M. S. Nixon. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval, pages 359–368, New York, NY, USA, 2008. ACM.
11. C. Boutsidis, E. Gallopoulos, SVD based initialization: A head start for nonnegative matrix factorization, Pattern Recognition, Volume 41, Issue 4, April 2008, Pages 1350-1362, ISSN 0031-3203.
12. Paul Pauca, Farial Shahnaz, Michael Berry and Robert Plemmons: Text Mining using Nonnegative Matrix Factorizations, Procceedings SIAM Inter. Conference on Data Mining, Orlando, April 2004.
13. D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. Advances in Neural Information Processing Systems, 13:556–562, 2001.