UNED @ Retrieving Diverse Social Images Task

A.Castellanos
NLP & IR Group, UNED
C/ Juan del Rosal,16
Madrid, Spain
acastellanos@lsi.uned.es

A. García-Serrano
NLP & IR Group, UNED
C/ Juan del Rosal,16
Madrid, Spain
agarcia@lsi.uned.es

J.Cigarrán
NLP & IR Group, UNED
C/ Juan del Rosal,16
Madrid, Spain
juanci@lsi.uned.es

ABSTRACT

This paper summarizes the participation of UNED at the 2014 Retrieving Diverse Social Images Task [3]. We propose a novel approach based on Formal Concept Analysis (FCA) to detect the latent topics related to the images and a later Hierarchical Agglomerative Clustering (HAC) to put together the images according to these latent topics. The diversification will be based on offering images from the different topics detected. In order to detect these latent topics, two kinds of data have been tested: only information related to the description of the images and all the textual information related to the images. The results show that our proposal is suitable for search result diversification, achieving similar results to those in the state of the art.

1. INTRODUCTION

Diversification, in the sense proposed by this task, refers to the creation of a diverse result list given an user query. The rationale is that the users are not only interested in accurate results but also in results covering different topics or situations. From the IR-based point of view, a good definition of the problem and a review of the state of the art can be found at [1]. To address this task we propose a novel approach based on an image representation using the concept/s covered by the image. For that, we use the information related to the images to create a conceptual-based data representation. The creation of this conceptual representation is tackled by means of the application of Formal Concept Analysis, a data organization technique. We expect this representation will make explicit the latent concepts in the images as well as the relationships between them. Based on this FCA representation, a HAC algorithm is proposed for grouping similar images, according to the latent concepts detected. Our hypothesis is that the resultant groups can represent the different topic addressed in the images. So, the image diversification will be done by taking one image from the different HAC-based clusters/topics.

2. WORK PROPOSAL

Our proposal is based on taking the information related to the images so as to create a conceptual-based representation. This representation intends to discover latent concepts in the data and the relationships between them. For that, we propose the application of Formal Concept Analysis as a modelling technique. By means of FCA, the set of images to be diversified will be organized according to their latent concepts.

After the FCA application, a hierarchy organizing the images in *formal concepts* according to their shared features will be obtained. However, it still remains the diversification of the images according to this representation. For that, we proposed the application of a HAC algorithm to group together the *formal concepts* that could be considered as similar (belonging to the same topic).

After this grouping, each HAC cluster can be considered as an image set covering a similar topic. Then, for each cluster the best image in the group (the one with a higher ranking according to the ranking provided by the task, taken from an IR system) will be taken and offered as result. The final result list will be a ranked list of images, ordered according to the provided ranking. As our methodology does not provide any score associated to the results and it was required by the organization, a dummy score has been set (1 for the 1st result in the ranking, 0.99 for the second, and so on).

2.1 Formal Concept Analysis

Formal Concept Analysis (FCA) is a mathematical theory of concept formation [6, 7] derived from lattice and ordered set theories that provide a theoretical model to organize *formal contexts*. A *formal context* is a set structure $\mathbb{K} := (G, M, I)$, where G is a set of objects, M a set of attributes and I a binary relationship between G and M, ($I \subseteq G \times M$), denoted by gIm, which is read as: the object g has the attribute m. An example of *formal context* is shown in the Figure 1a. From the point of view of a content representation system, the *formal context* can be seen as the set of contents (images) to be represented (G) is the set of items to be represented, the set of textual (or other kind of) features (M), representing the items, and the binary relationship I can be read as item g has the feature m.

From the information in the *formal context*, a set of *formal concepts* can be inferred. A *formal concept* is a pair (A, B) of objects and attributes, which has the following properties: If an object a in A is tagged with an attribute b, then b must is included in B ($B = A^I$ includes all the attributes shared by the objects). Conversely, if an object a is tagged with all the attributes in B, then a must be included in A (i.e. $A = B^I$: includes all those objects filtered out by the attributes). Finally, the whole sets of *formal concepts* can be

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

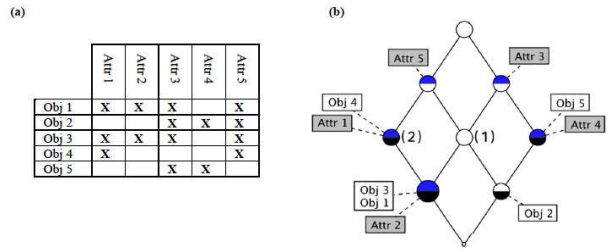

Figure 1: (a) A sample *formal context* and (b) the concept lattice associated with the *formal context*

organized according to an order relationship, from the most generic to the most specific. This order can be proven to be a lattice and, consequently, to be represented as a Hasse Diagram, as the one in the Figure 1b.

2.2 FCA-based Modelling

Following the FCA rationale, the information of each of the 123 locations included in the test set is modelled. After this modelling, a set of *formal concepts*, grouping together the images sharing a same set of features, will be obtained. To select the features to model the images we propose two different kinds of data: 1) use only the description information of the images (description, title and tags), and 2) use all the textual information related to the images (description, title, tags, user information and date information).

In order to select only those most-representative features, we applied Kullback-Leibler Divergence (KLD) [4] on the textual contents related to the images. Basically, KLD compares the textual content related to a given image to the textual contents of the rest of the images. In this way, KLD allows representing each image by the textual contents that better differentiates a given image from the other ones. Hopefully, it will allow the improvement of the diversity of the images selected.

2.3 HAC-based grouping

From the FCA-based modelling, a hierarchy organizing the obtained set of *formal concepts* is provided. It remains the creation of a set of diverse image groups based on this organization. For that, we propose the application of a HAC algorithm [5]. Specifically, we propose a Single Linking based hierarchical clustering that groups together similar *formal concepts*. In order to set the "similarity" of two *formal concepts*, we have applied the Zero-Induces index [2]. This index measures the similarity of two *formal concepts*, based on the number of features that they share.

3. RESULTS

Table 1 shows the results obtained by our approaches for the official runs. In general the performance of our approaches are in the same level than other approaches in the state of the art, according the results of last year for the same task. The inclusion of all the textual information related to an image seems to slightly improve the performances, both for precision and also for diversity based results. It is reasonable to think that including different kinds of data will improve the diversity of the results. However, as it can be seen in the results, also the precision based results

Table 1: Official Metrics for Retrieving Diverse Social Images Task

Approach	P@20	CR@20	F-measure
RUN-2: Description Info	0,7581	0,4325	0,5429
RUN-5: All Textual Info	0,7772	0,4343	0,5502

are favoured by the inclusion of more information than only the related to the description. However, this improvement is not enough to definitely conclude that this information can represent a valuable indicator for image diversity.

4. CONCLUSIONS

In this paper we presented a Conceptual-based modelling (based on FCA) for improving diversity in the retrieving of social images. With FCA we intend to infer the latent topics addressed in the information related to the images. Once the latent concepts has been detected, in order to put together the most similar ones, we have applied an HAC approach.

Our hypothesis is that this concept modelling could help in the diversification of the retrieved images. For that, given a query the system will retrieve images trying to cover all the identified topics.

In this work we have also experimented with different kinds of information to describe the images. The obtained results proved our proposal as suitable to offer accurate and also diverse results. Related to the kind of information to use, the inclusion of the most information possible about the images seems to be the best choice. For both, precision and diversity based results, its performance is better, although the improvement in the diversity is not very large.

5. ACKNOWLEDGMENTS

This work has been partially supported by VOXPOPULI (TIN2013-47090-C3-1-P) Spanish project.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[2] F. Alqadah and R. Bhatnagar. Similarity measures in formal concept analysis. *Annals of Mathematics and Artificial Intelligence*, 61(3):245–256, 2011.

[3] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînsca, and H. Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *Proceedings of MediaEval Benchmarking Initiative for Multimedia Evaluation.*, 2014.

[4] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[5] C. D. Manning, P. Raghavan, and H. Schütze. Hierarchical clustering. pages 377–403. 2008.

[6] R. Wille. Concept lattices and conceptual knowledge systems. *Computers & mathematics with applications*, 23(6):493–515, 1992.

[7] R. Wille. *Restructuring Lattice Theory: An Approach Based On Hierarchies Of Concepts*, volume 5548 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009.