

Unsupervised Cleaning for Web Image Retrieval

Chen Cao, Yang Peng, Mustaqim Moulavi, Harsh Bhavsar

Department of Computer & Information Science & Engineering, University of Florida

INTRODUCTION

• Image retrieval cleaning aims at improving the accuracy of web image search and retrieval. It first removes the irrelevant images and then re-orders the rest of the images using the confidence score to be relevant images. A practicable and efficient cleaning method should be highly accurate without manual labeling and pre-processing.

• In this project, a novel algorithm for unsupervised image retrieval cleaning is proposed. We delved into the data distribution and devised three stages noise robust approaches to solve the problem.

FRAMEWORK

Many categories of images are retrieved by different textual-keywords via search engines and image sharing sites. First, graphs are built on pairwise images to explore the in-category data distribution. Isolated nodes are regarded as irrelevant images. Then, classifiers are trained on pairwise categories. Images that classified to other categories are also treated as irrelevant images. After we remove all irrelevant images, for each category, images in dominant clusters, which are most likely to be relevant images, are used as manifold learning pseudo queries for re-ordering.

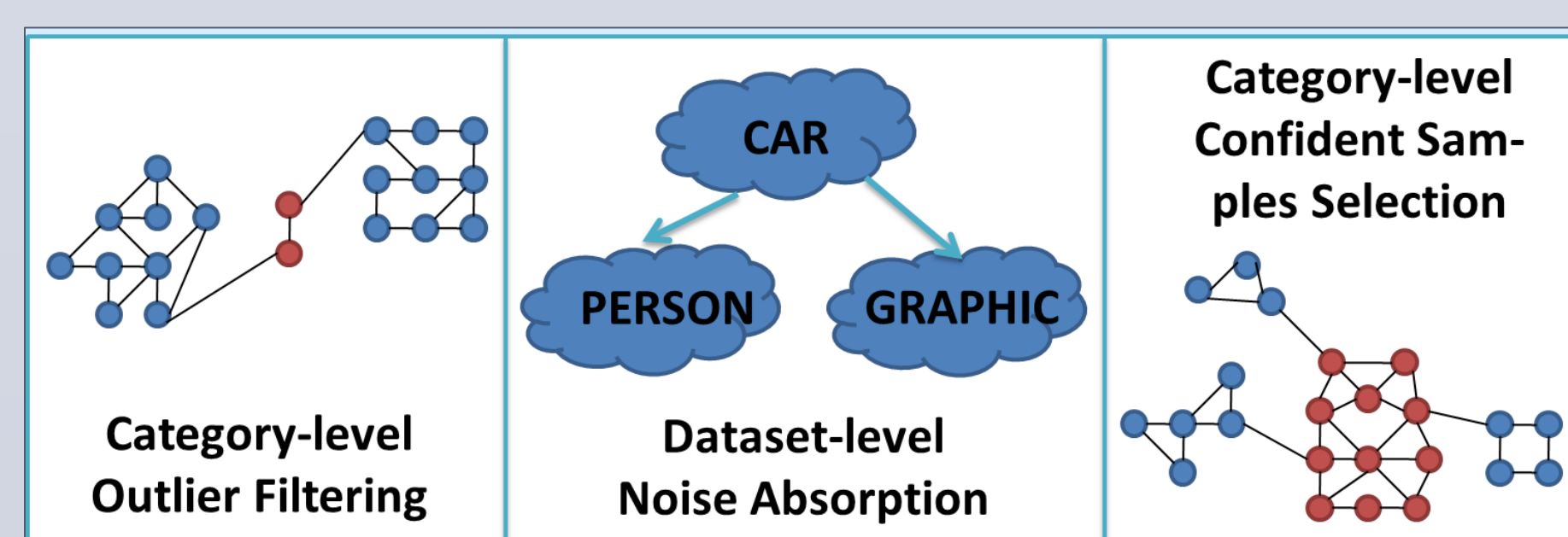


Figure: The flowchart of our three stages algorithm.

ALGORITHMS

• Graph Building

$X=\{x_1, \dots, x_n\}$ represents the features of retrieved images, the edge matrix W is built as $w_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)$ if $i \neq j$ and $w_{ii}=0$. W is normalized as $S=D^{-1/2}WD^{-1/2}$, and D is diagonal with $d_{ii}=\sum_{j=1:n} w_{ij}$.

• Category-level outlier filtering

X is mapped to a new feature space as $g(x_i)=F^*(i, \cdot)^T$, where $F^*=(I-\alpha S)^{-1}[y_1, \dots, y_i, \dots, y_n]$ $=(I-\alpha S)^{-1}y_i$, y_i is a column vector with $y_i=1$ and $y_{others}=0$. x_i is warped as the i -th row of matrix F^* , and a dominant score is computed as to sum all dimensions of this row.

Spectral clustering is implemented on $g(X)$ for k clusters. Clusters with the lowest mean dominant score are considered as noise.

• Dataset-level noise absorption

There are total t keywords-queries denoted as $Q=\{q_1, \dots, q_i, \dots, q_t\}$ and the retrieval dataset $X^{(Q)}=\{X^{(1)}, \dots, X^{(i)}, \dots, X^{(t)}\}$. Linear kernel SVM classifier is trained for every two categories q_i and q_j . The confidence score $C_i(x_a^{(i)})$ of an image $x_a^{(i)}$ to be the relevant image in its category is $C_i(x_a^{(i)}) = \min_{j \neq i} D_{i,j}(x_a^{(i)})$. If the score is low, $x_a^{(i)}$ is regarded as noise to be filtered and absorbed by q_j .

• Category-level confident sample selection

For each category, we want to select the data points with high density score, i.e. confidence, and discard the data points with low density score. In order to calculate the density score, we use elastic net SVM regression to estimate the density of data points. The density score of x is a function of $w^T x$. We use a modified version of SMO (Sequential Minimization Optimization) to train the model.

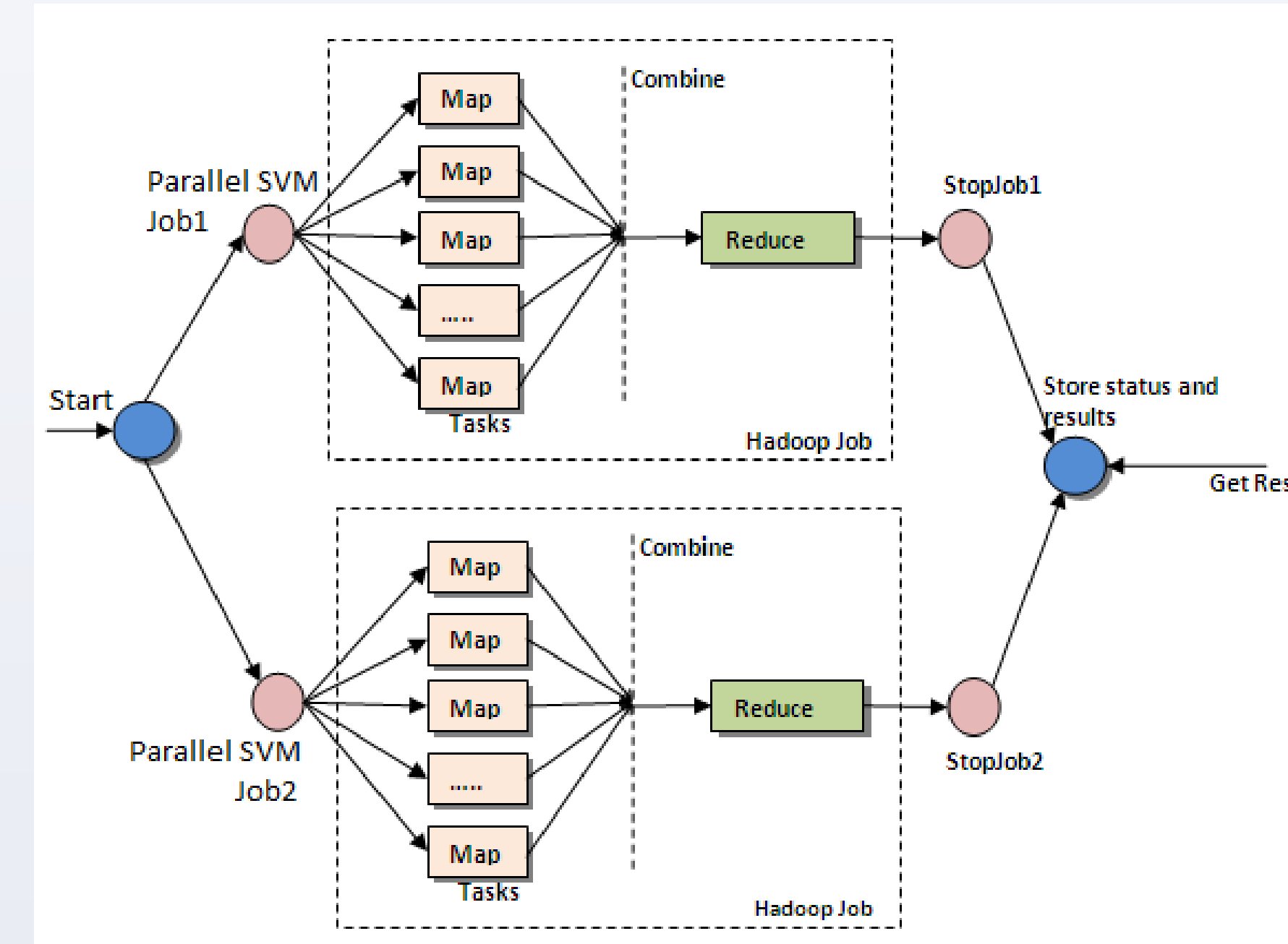


Figure: Parallel MapReduce based SVM on Amazon EC2 clusters.

RESULTS

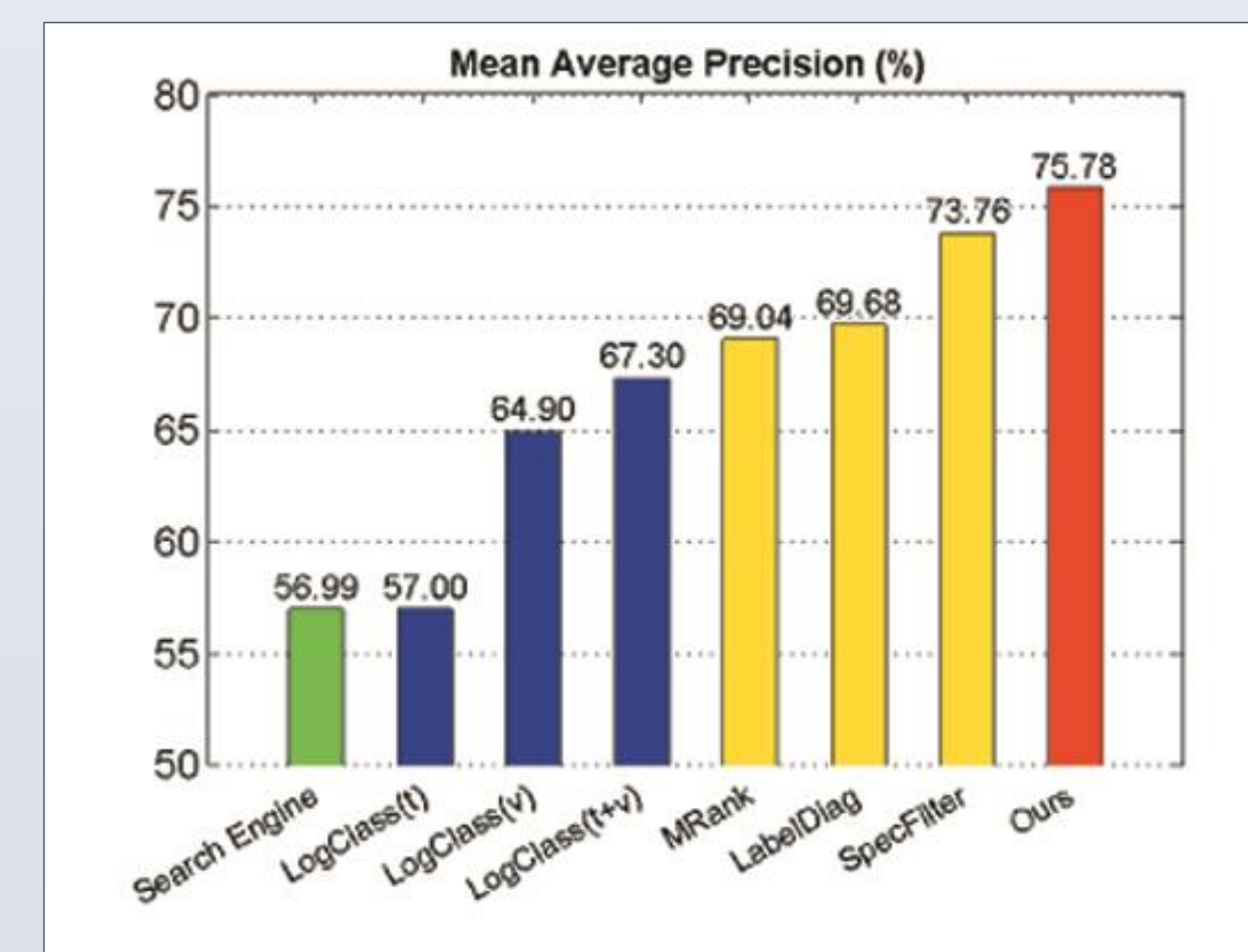


Figure: INRIA dataset: MAP of ranking order results over 353 categories.

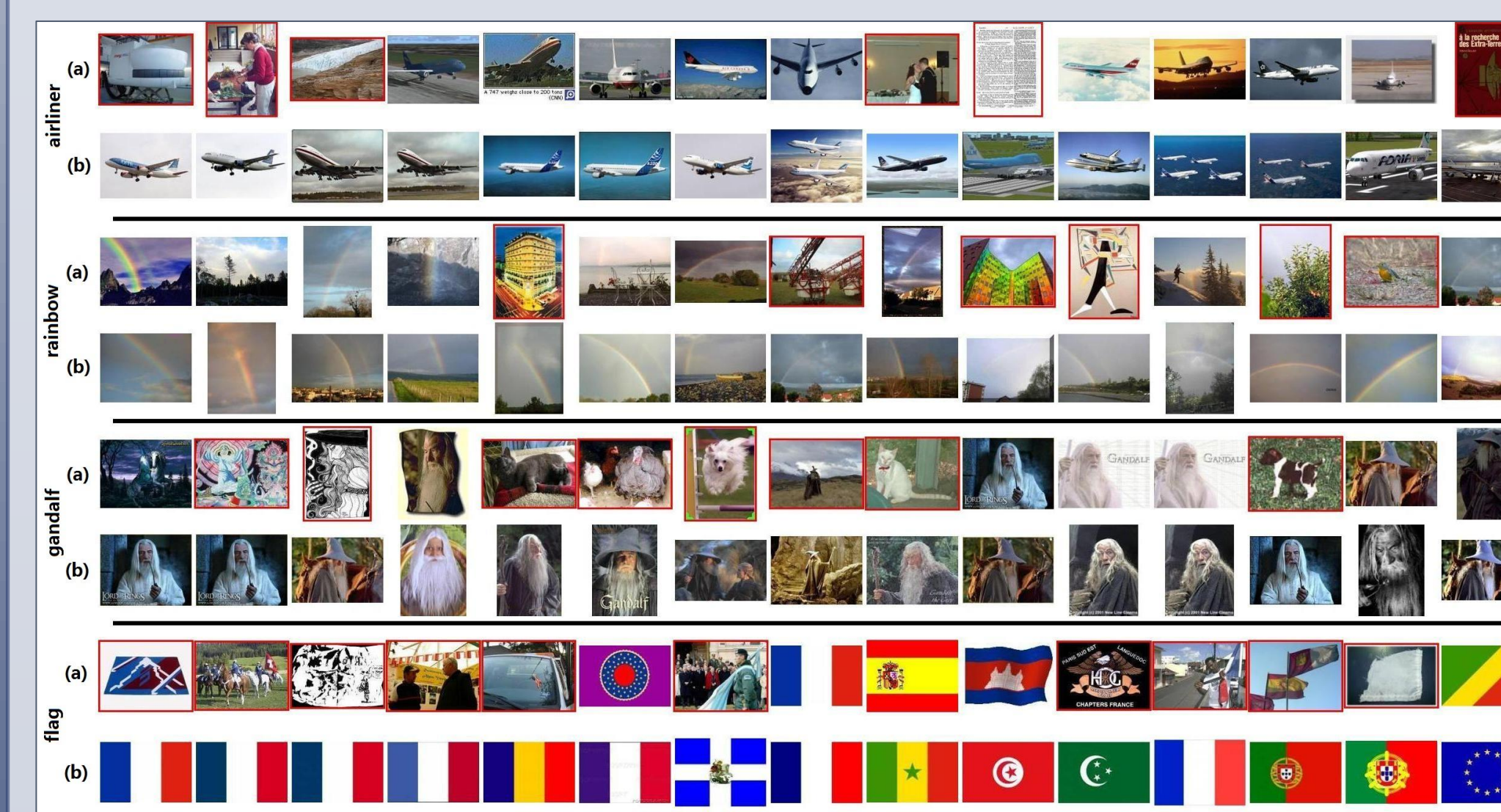


Figure: Examples in INRIA dataset: Top-15 ranked images using (a) the search engine and (b) our approach.

CONCLUSION

In this project, we propose a new unsupervised cleaning algorithm to improve the accuracy of web image retrieval, by three level irrelevant images filtering and relevant samples selection. We warp data into a noise resistant spectral space to cluster and remove isolated noise. Then a specially designed cross category noise absorption algorithm is proposed to make the dataset cleaner. Finally, confident samples used for re-ordering are selected by a regression formulation with linear kernel and sparsity constraints. Our approach demonstrates the best performance among all the recent methods in experimental comparison on INRIA dataset.

REFERENCES

[Spectral Clustering] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In NIPS, 2002.
[Elastic Net SVM] G.-B. Ye, Y. Chen, and X. Xie. Efficient variable selection in support vector machines via the alternating direction method of multipliers. In ICAIS, 2011.
[SMO] J. Platt, and et al. Sequential minimal optimization: A fast algorithm for training support vector machines. In technical report Microsoft Research, 1998.
[INRIA dataset and LogClass] J. Krapac, M. Allan, J. Verbeek, and F. Juried. Improving web image search results using query-relative classifiers. In CVPR, 2010.
[MRank] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In NIPS, 2004.
[LabelDiag] J. Wang, Y. Jiang, and S. Chang. Label diagnosis through self tuning for web image search. In CVPR, 2009.
[SpecFilter] W. Liu, Y. Jiang, J. Luo, and S. Chang. Noise resistant graph ranking for improved web image search. In CVPR, 2011.

ACKNOWLEDGEMENTS

Great thanks to Dr. Daisy Wang for providing us with an opportunity to work on this project and present this poster. Thanks to Amazon for providing the AWS credits. These credits were utilized for training and testing the SVM classifiers on Amazon EC2 clusters. We also thank Yottamine Analytics whose services were used for implementing SVM classifier on Amazon EC2 clusters.