BENCHMARKING RESULT DIVERSIFICATION IN SOCIAL IMAGE RETRIEVAL

Bogdan Ionescu[1], *Adrian Popescu*[2], *Henning Müller*[3], *María Menéndez*[4], *Anca-Livia Radu*[1,4]

[1]LAPI, University Politehnica of Bucharest, Romania
[2]CEA-LIST, France
[3]University of Applied Sciences Western Switzerland, Sierre
[4]DISI, University of Trento, Italy
bionescu@imag.pub.ro, adrian.popescu@cea.fr, henning.mueller@hevs.ch

ABSTRACT

This article addresses the issue of retrieval result diversification in the context of social image retrieval and discusses the results achieved during the MediaEval 2013 benchmarking. 12 systems and their results are described and analyzed in this text. A comparison of the use of expert vs. crowdsourcing annotations shows that crowdsourcing results are slightly different and have higher inter observer differences but results are comparable at lower cost. Multimodal approaches have best results in terms of cluster recall. Manual approaches can lead to high precision but often lower diversity. With this detailed results analysis we give future insights on this matter.

Index Terms— social photo retrieval, result diversification, image content description, re-ranking, crowdsourcing.

1. INTRODUCTION

An efficient image retrieval system should be able to present results that are both relevant and that are covering different aspects (diversity) of a query, e.g., providing different representations rather than duplicates. Relevance was more thoroughly studied in existing literature than diversification [1, 2, 3] and even though a considerable amount of diversification literature exists, the topic remains an important one, especially in social media [4, 5, 6, 7, 8]. Due do the subjectivity of this task, a critical point are the evaluation tools and especially ground truth annotation which tends to be restrained, not enough attention being payed to its statistical significance.

Benchmarking activities provide a framework for evaluating systems on a shared dataset and using a set of common rules. The results obtained are thus comparable and a wider community can benefit from it. In this paper we analyze the contributions to the community of the MediaEval 2013 Retrieving Diverse Social Images Task [21] and its dataset [22] (Div400), which focus on fostering new technology for improving both relevance and diversification of search results with explicit emphasis on the actual social media context. These two characteristics of retrieval results are antinomic, i.e., the improvement of one of them usually results in a degradation of the other; this requires a deeper analysis.

The retrieving diverse social images task builds on the useful experience accumulated during ImageCLEF [5, 26] and goes beyond the state-of-the-art in the following directions: (1) the proposed evaluation framework [22] focuses on improving the current technology by using Flickr's relevance system as reference[1] (which is one of the most popular platforms) and addresses in particular the social dimension reflected in the nature of data and methods devised to retrieve it; (2) while smaller in size than ImageCLEF collections [5, 26], Div400 contains images that are already associated to topics by Flickr. This design choice ensures that there are many relevant images for all topics and pushes diversification in priority; (3) unlike ImageCLEF, which worked with generic ad-hoc retrieval scenarios, a real-world usage scenario is set up (i.e., tourism) to disambiguate the diversification need; (4) finally, a comparison of expert and crowdsourced ground truth production is performed to assess the potential differences between lab and real life evaluations.

The main focus of this work is to provide a comparative analysis of the state-of-the-art systems submitted to the task that addressed a broad category of approaches varying from single modal to multimodal, from using graph representations, re-ranking, optimization approaches, clustering, heuristic to hybrid approaches that included humans in the loop. This analysis is helpful in that it evidences strong and weak points of current diversification technology and can be used to guide further work in the area. The remaining of the paper is organized as follows: Section 2 describes the evaluation framework and the dataset, Section 3 overviews MediaEval 2013 participant systems, Section 4 discusses the experimental results while Section5 concludes the paper.

2. EXPERIMENT AND DATA DESCRIPTION

To benchmark retrieval diversification techniques, the following task was designed and validated within the 2013

This work was supported by CUbRIK FP7 n287704, PROMISE FP7 n258191 and MUCKE CHIST-ERA.

MediaEval benchmark [9]. The task builds on current state-of-the-art retrieval technology, e.g., using the Flickr media platform[1], with the objective of fostering approaches that will push forward the advances in the field. Given the important social role of geographic queries and their spatio-temporal invariance, experimentation with the retrieval of photos with landmark locations was considered. For 396 locations, up to 150 photos (with Creative Commons redistributable licenses) and associated metadata are retrieved from Flickr and ranked with Flickr's default "relevance" algorithm. To compare different retrieval mechanisms, data was collected with both textual (i.e., location name — *keywords*) and GPS queries (*keywordsGPS*). Location metadata consists of Wikipedia links to location webpages and GPS information and photo metadata includes social data, e.g., author title and description, user tags, geotagging information, time/date of the photo, owner's name, the number of times the photo has been displayed, number of posted comments, rank, etc. Apart from these data, to support contributions from different communities, some general purpose content descriptors are provided for the photos — visual descriptors, e.g., histogram of oriented Gradients, Local Binary Patterns, MPEG-7 related features, etc; and textual models, e.g., probabilistic models, Term Frequency-Inverse Document Frequency (TF-IDF) weighting and social TF-IDF weighting (an adaptation to the social space) [23]. The dataset provides 43,418 photos and is divided into a *devset* of 50 locations (5,118 photos, in average 102.4/location) intended for training and a *testset* of 346 locations (38,300 photos, in average 110.7/location) for evaluation. The dataset was made publicly available [21, 22].

Data are annotated for both relevance and diversity of the photos. The following definitions were adopted: *relevance* — a photo is relevant if it is a common photo representation of the location, e.g., different views at different times of the day/year and under different weather conditions, inside views, creative views, etc, which contain partially or entirely the target location (bad quality photos are considered irrelevant) — photos are tagged as relevant, non-relevant or with "don't know"); *diversity* — a set of photos is considered to be diverse if it depicts complementary visual characteristics of the target location (e.g., most of the perceived visual information is different — relevant photos are clustered into visually similar groups). Annotations were carried out mainly by experts with advanced knowledge of location characteristics. To explore differences between experts and non-experts annotations, a subset of 50 locations from the *testset* was annotated using crowd-workers (via the CrowdFlower[2] platform). In all cases, visual tools were employed to facilitate the process.

Annotations were carried out by several annotators and final ground truth was determined after a lenient majority voting scheme. Table 1 presents the number of distinct annotations (the number of annotators is depicted in the brackets), Kappa inter-annotator agrement (*devset* reports weighted Kappa [25], *testset* reports Free-Marginal Multirater Fleiss' Kappa [24] as different parts of the data are annotated by different annotators) and cluster statistics. Expert annotations achieved a good agreement as average Kappa is above 0.6 and up to 0.8 (values above 0.6 are considered adequate and above 0.8 are considered almost perfect[25]). Only 0.04% of the photos achieved "don't know" answers. The diversity annotations lead to an average of around 12 clusters/location and 5-6 images/cluster. For the crowd annotations, the agreement is significantly lower, namely 0.36, and up to 1% of the photos achieved "don't know" answers, which reflects the variable backgrounds of the crowd (on average it leads to 4.7 clusters/location and 32.5 images/cluster).

Table 1: Expert and crowd annotation statistics.

devset (expert)	*testset* (expert)	*testset* (crowd)
relevance (annotations - avg.Kappa - % relevant img.)		
6(6) - 0.64 - 73	3(7) - 0.8 - 65	3(175) - 0.36 - 69
diversity (annotations - avg.clusters/location - avg.img./cluster)		
1(3) - 11.6 - 6.4	1(4) - 13.1 - 5	3(33) - 4.7 - 32.5

Given the dataset above, the task required developing techniques that allow the refinement of the initial Flickr retrieval results by selecting a ranked list of up to 50 photos that are equally relevant and diverse representations of the query.

3. SYSTEM DESCRIPTIONS

In total, 24 teams from 18 countries registered to the 2013 Retrieving Diverse Social Images Task and 11 submitted results. The key contributions are presented in the following:

- SOTON-WAIS (*re-ranking, Greedy optimization; multimodal*): use a pre-filtering to remove images unlikely to be relevant (e.g., with faces in focus, blurred, predominantly text, not viewed on Flickr, etc.) followed by re-ranking with a proximity search (use of Lucene[3]) to improve precision and a Greedy Min-Max diversifier [10];

- SocSens (*Greedy optimization, clustering; multimodal, human*): the visual approach involves Greedy optimization of a utility function that weights both relevance and diversity scores. First, the text-based approach involves Hierarchical Clustering with image ranking using random forests. Diversification is achieved by stepping through the clusters iteratively and selecting the most relevant images. A second text-based approach uses camera Exif information and weather data with k-means clustering to diversify images based on view angle, distance, indoor/outdoor and meteo conditions. The multimodal approach involves late fusion of the outputs of the previous schemes. The human approach required assessors to manually refine results provided with the text-based approach (refinement limited to the first 15 images) [15];

[1] http://www.flickr.com/services/api/

[2] http://crowdflower.com/

[3] http://lucene.apache.org/

- CEA (*re-ranking; multimodal*): uses a re-ranking approach focusing mainly on the utility of social cues for diversification. Query results are diversified by images from different users or that were taken by the same user on different days. Textual and visual re-ranking that selects images in these spaces are proposed [19];
- UPMC (*re-ranking, clustering; multimodal*): use re-ranking to improve relevance, Agglomerative Hierarchical Clustering with cluster and cluster image sorting according to a priority criterion and a final re-ranking that alternates images from different clusters to account for diversity [11];
- MMLab (*clustering, Greedy optimization; multimodal*): visual approach using a variant of LAPI's approach [13]. The text-based approach uses a Greedy optimization of the Average Diverse Precision metric. The multimodal approach uses Hierarchical Clustering [16];
- BMEMTM (*heuristic, clustering, re-ranking; multimodal, human*): visual diversification was achieved by downranking images containing faces and using a hierarchic clustering to diversify (images selected from different clusters). Text-based diversification consists mainly of re-ranking based on weights in the text models. For multi-modality, the visual approach was run after text. For the human approach, annotators were asked to manually cluster/tag images and then a relevant image is selected from each cluster and results ordered according to the initial Flickr ranking [14];
- MUCKE (*re-ranking, clustering; multimodal*): uses a k-Nearest Neighbor inspired re-ranking algorithm as a preliminary filter followed by k-means++ clustering. Clusters are ranked by social relevance and final diversification was achieved by retaining one image from each cluster by descending similarity to the cluster centroid [17];
- TIA-INAOE (*functional optimization; multimodal*): transposed the diversification problem into the optimization of an objective function that combines relevance and diversity estimates. Optimization was achieved with a multi-objective evolutionary algorithm that simultaneously maximize image diversity in consecutive positions while minimizing divergence from the original ranking [12];
- LAPI (*re-ranking, clustering; multimodal*): uses re-ranking according to the similarity against the "most visually common" image in the set for precision and a clustering diversification mechanism that retains only the best ranked representative images in each cluster [13];
- UEC (*web inspired ranking; multimodal*): uses VisualRank for precision (rank values are estimated as the steady state distribution of a random-walk Markov model) followed by Ranking with Sink Points for diversification [18];
- ARTEMIS (*graph representation; visual*): uses a matching graph representation through quantized interest point similarities (whereas groups of similar instances become connected components). Diversification is achieved by selecting from each cluster the images with the highest similarity scores cumulated over its matches [20].

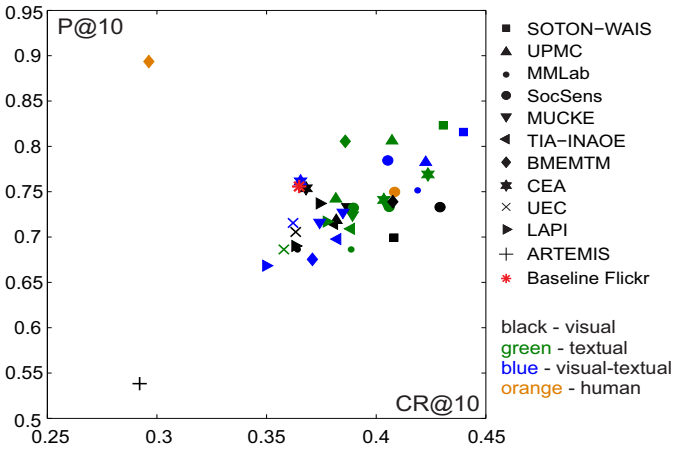

Fig. 1: Precision vs. cluster recall averages at 10 images.

4. EXPERIMENTAL RESULTS

System performance was assessed on the *testset* using Cluster Recall at X (CR@X) — a measure assessing how many clusters from the ground truth are represented among the top X results (only relevant images considered) [5] and Precision at X (P@X) — measures the number of relevant photos among the top X results. Systems were optimized to CR@10 as official metric (highest possible value is 0.77 as on average there are 13 clusters per location, see Table 1).

Retrieval with GPS information yields more accurate results than using solely keywords, e.g., for the initial Flickr results, P@10 with keywords is 0.7045 compared to 0.7881 using GPS data in addition. Diversity is however slightly higher for keywords, CR@10 is 0.3985 compared to 0.3437 using GPS as results are sparse. In the following we focus on presenting the average overall results.

Figure 1 plots overall precision against recall averages for all participant runs at a cutoff at 10. For *visual approaches*, highest diversification is achieved with a Greedy optimization of VLAD+SURF descriptors, CR@10=0.4291 — SocSens run1 [15], while lowest diversification is provided by a matching graph approach also with feature point information (RootSIFT and Hessian), CR@10=0.2921 — ARTEMIS [20]. Using simple color information (histograms and detection of faces) still achieves high recall, e.g., CR@10=0.4076 — BMEMTM run1 [14], which proves that there is no superiority of classes of descriptors, the difference in performance being mainly related to the method. Compared to visual, *text information* tends to provide better results (see green points). Highest diversification is achieved using a re-ranking with Lucene and Greedy Min-Max optimization, CR@10=0.4306 — SOTON-WAIS run2 [10] where data are represented with time-related information. On the other end, bag-of-words of TF-IDF data and web inspired ranking leads to CR@10=0.3579 — UEC run2 [18]. Surprisingly, *human*

Table 2: Precision and cluster recall averages for best team runs.

team best run	P@10	P@20	P@30	P@40	P@50	**CR@10**	CR@20	CR@30	CR@40	CR@50
SOTON-WAIS run3 [10]	**0.8158**	**0.7788**	0.7414	0.7059	0.6662	**0.4398**	0.6197	0.7216	0.7844	0.8243
SocSens run1 [15]	0.733	0.7487	**0.7603**	0.7145	0.5915	0.4291	**0.6314**	0.7228	0.7473	0.7484
CEA run2 [19]	0.769	0.7639	0.7565	**0.7409**	**0.7153**	0.4236	0.6249	0.7346	0.8148	**0.8668**
UPMC run3 [11]	0.7825	0.73	0.7254	0.7099	0.6891	0.4226	0.6268	0.747	0.8154	0.854
MMLab run3 [16]	0.7515	0.7404	0.7335	0.7185	0.697	0.4189	0.6236	**0.7492**	**0.8205**	0.8653
BMEMTM run1 [14]	0.7389	0.7164	0.7182	0.7115	0.6927	0.4076	0.6139	0.7184	0.7935	0.844
MUCKE run2 [17]	0.7243	0.7228	0.7183	0.708	0.6884	0.3892	0.5749	0.6877	0.7684	0.8306
TIA-INAOE run2 [12]	0.7091	0.7136	0.7146	0.7045	0.6851	0.3885	0.5732	0.6897	0.7719	0.8228
LAPI run2 [13]	0.717	0.7111	0.6896	0.6477	0.5795	0.3774	0.5734	0.682	0.7472	0.7722
baseline Flickr	0.7558	0.7289	0.7194	0.708	0.6877	0.3649	0.5346	0.6558	0.7411	0.7988
UEC run1 [18]	0.7056	0.7092	0.7076	0.6948	0.6752	0.3633	0.5448	0.6743	0.7572	0.8154
ARTEMIS run1 [20]	0.5383	0.3379	0.2269	0.1702	0.1361	0.2921	0.3306	0.331	0.331	0.331

approaches were less effective than the automatic ones as users tend to maximize precision at the cost of diversity, e.g., BMEMTM run4 [14] achieves P@10=0.8936 but CR@10 is only 0.2963. However, human-machine integration improves also the diversity, e.g., CR@10=0.4048 — SocSens run4 [15]. Overall, the best performing approach is *multi-modal*, namely CR@10=0.4398 — SOTON-WAIS run3 [10], it improves diversification of the state-of-the-art Flickr initial results with at least one additional image class.

Table 2 presents the official ranking of the best team approaches for various cutoff points (highest values are in bold). In addition to the information from Figure 1, Table 2 shows that the precision tends to decrease with the increase of the number of images as it is more likely to obtain non-relevant pictures. On the other hand, cluster recall increases with the number of pictures as it is more likely to get pictures from additional classes.

To determine the statistical significance of the results and thus to examine the relevance of the dataset, a stability test was run [26]. Stability is examined by varying the number of topics which is used to compute performance. Stability tests are run with different topic subset sizes, which are compared to the results obtained with the full testset (346 topics). For each topic subset, 100 random topic samplings are performed to obtain stable averages. Spearman's rank correlation coefficient [27] is used to compare the obtained CR@10 rankings and the obtained values are 0.61, 0.86, 0.93, 0.96, 0.97, 0.98, 0.99 for subsets which contain 10, 50, 100, 150, 200, 250 and 300 topics. These results show that there is little change in the ranking when at least 100 topics are used. The size of the testset is clearly sufficient to ensure statistical significance of the ranking and therefore of the results.

Performance assessment depends on the subjectivity of the ground truth, especially for the diversification part. The final experiment consisted of comparing both results achieved with expert and crowd annotations. Table 3 presents the four best team runs (highest results are depicted in bold; results on a selection of 50 locations from testset, see Section 2). Although precision remains more or less similar in both cases, cluster recall is significantly higher for the crowd annotations.

Table 3: Expert vs. crowd annotations — precision and cluster recall averages for team best runs.

team best run	*expert GT*		*crowd GT*	
	P@10	CR@10	P@10	CR@10
SOTON-WAIS run3 [10]	**0.8755**	0.4129	**0.7714**	0.745
SocSens run1 [15]	0.7959	0.4139	0.7286	0.7636
CEA run2 [19]	0.8265	0.4081	0.7082	0.7287
UPMC run3 [11]	0.8408	**0.4151**	0.749	**0.788**
baseline Flickr	0.7980	0.3345	0.6816	0.6643

This is due to the fact that workers tend to under-cluster the images for time reasons. Nevertheless, what is interesting is the fact that regardless of the ground truth, the improvement of the baseline is basically the same: 0.0784 for experts compared to 0.0807 for the crowd, which shows that results are simply translated but the relevance is still the same. Crowd annotations are an attractive alternative to expert annotations, being fast — order of hours compared to expert ones that require weeks — while the performance is similar.

5. CONCLUSIONS AND OUTLOOK

This article describes the MediaEval task on retrieving diverse social images. The strong participation in a first year shows the interest in the topic. Several groups increased specific aspects of the results on the strong Flickr baseline, particularly linked to diversity. Approaches combining a large variety of modalities from manual re-ranking, GPS to visual and text attributes have the potential to increase results. Detecting objects such as faces was used and via the analysis of the clusters of relevant images several categories can likely be deduced and used in connection with detectors for these aspects to optimize results.

For a continuation it seems important to look into criteria that can stronger discriminate the runs, so making the task harder. More clusters are an option, or a hierarchy of clusters. A larger collection is an option but creating diversity ground truth for large collections is tedious and expensive. Crowdsourcing could be a valid approach as the experiments show.

6. REFERENCES

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based image retrieval at the end of the early years", IEEE Transaction on Pattern Analysis and Machine Intelligence, 22(12), pp. 1349 - 1380, 2000.

[2] R. Datta, D. Joshi, J. Li, J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Comput. Surv., 40(2), pp. 1-60, 2008.

[3] R. Priyatharshini, S. Chitrakala, "Association Based Image Retrieval: A Survey", Mobile Communication and Power Engineering, Springer Communications in Computer and Information Science, 296, pp 17-26, 2013.

[4] R.H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, "Visual Diversification of Image Search Results", ACM World Wide Web, pp. 341-350, 2009.

[5] M.L. Paramita, M. Sanderson, P. Clough, "Diversity in Photo Retrieval: Overview of the ImageCLEF Photo Task 2009", ImageCLEF 2009.

[6] B. Taneva, M. Kacimi, G. Weikum, "Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity", ACM Web Search and Data Mining, pp. 431-440, 2010.

[7] S. Rudinac, A. Hanjalic, M.A. Larson, "Generating Visual Summaries of Geographic Areas Using Community-Contributed Images", IEEE Transactions on Multimedia, 15(4), pp. 921-932, 2013.

[8] A.-L. Radu, B. Ionescu, M. Menéndez, J. Stöttinger, F. Giunchiglia, A. De Angeli, "A Hybrid Machine-Crowd Approach to Photo Retrieval Result Diversification", Multimedia Modeling, Ireland, LNCS 8325, pp. 25-36, 2014.

[9] MediaEval 2013 Workshop, Eds. M. Larson, X. Anguera, T. Reuter, G.J.F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, M. Soleymani, co-located with ACM Multimedia, Barcelona, Spain, October 18-19, CEUR-WS.org, ISSN 1613-0073, Vol. 1043, `http://ceur-ws.org/Vol-1043/`, 2013.

[10] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, P. Lewis, "Experiments in Diversifying Flickr Result Sets", Working Notes Proceedings [9], 2013.

[11] C. Kuoman, S. Tollari, M. Detyniecki, "UPMC at MediaEval 2013: Relevance by Text and Diversity by Visual Clustering", Working Notes Proceedings [9], 2013.

[12] H.J. Escalante, A. Morales-Reyes, "TIA-INAOE's Approach for the 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings [9], 2013.

[13] A-L. Radu, B. Boteanu, O. Pleş, B. Ionescu, "LAPI @ Retrieving Diverse Social Images Task 2013: Qualitative Photo Retrieval using Multimedia Content", Working Notes Proceedings [9], 2013.

[14] G. Szűcs, Z. Paróczi, D.M. Vincz, "BMEMTM at MediaEval 2013 Retrieving Diverse Social Images Task: Analysis of Text and Visual Information", Working Notes Proceedings [9], 2013.

[15] D. Corney, C. Martin, A. Göker, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, L. Aiello, B. Thomee, "SocialSensor: Finding Diverse Images at MediaEval 2013", Working Notes Proceedings [9], 2013.

[16] B. Vandersmissen, A. Tomar, F. Godin, W. De Neve, R. Van de Walle, "Ghent University-iMinds at MediaEval 2013 Diverse Images: Relevance-Based Hierarchical Clustering", Working Notes Proceedings [9], 2013.

[17] A. Armagan, A. Popescu, P. Duygulu, "MUCKE Participation at Retrieving Diverse Social Images Task of MediaEval 2013", Working Notes Proceedings [9], 2013.

[18] K. Yanai, D.H. Nga, "UEC, Tokyo at MediaEval 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings [9], 2013.

[19] A. Popescu, "CEA LISTs Participation at the MediaEval 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings [9], 2013.

[20] A. Bursuc, T. Zaharia, "ARTEMIS @ MediaEval 2013: A Content-Based Image Clustering Method for Public Image Repositories", Working Notes Proceedings [9], 2013.

[21] B. Ionescu, M. Menéndez, H. Müller, A. Popescu, "Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation", Working Notes Proceedings [9], 2013.

[22] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, "Div400: A Social Image Retrieval Result Diversification Dataset", ACM Multimedia Systems, 19-21 March, Singapore, 2014.

[23] A. Popescu, G. Grefenstette, "Social Media Driven Image Retrieval", ACM ICMR, April 17-20, Trento, Italy, 2011.

[24] J.J. Randolph, "Free-Marginal Multirater Kappa (multirater κfree): an Alternative to Fleiss Fixed-Marginal Multirater Kappa", Joensuu Learning and Instruction Symposium, 2005.

[25] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit", Psychological Bulletin, Vol. 70(4), pp. 213-220, 1968.

[26] T. Tsikrika, J. Kludas, A. Popescu, "Building Reliable and Reusable Test Collections for Image Retrieval: The Wikipedia Task at ImageCLEF", IEEE Multimedia, 19(3), pp. 24-33, 2012.

[27] A. Lehman, "Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide", Cary, NC: SAS Press. p. 123. ISBN 1-59047-576-3, 2005.