Performance Measures for Multilabel Evaluation

A Case Study in the Area of Image Classification

Stefanie Nowak
Fraunhofer IDMT
Ehrenbergstr. 31
98693 Ilmenau, Germany
nwk@idmt.fraunhofer.de

Hanna Lukashevich
Fraunhofer IDMT
Ehrenbergstr. 31
98693 Ilmenau, Germany
lkh@idmt.fraunhofer.de

Peter Dunker
Gracenote, Inc.
2000 Powell Street
Emeryville, CA 94608
pdunker@gracenote.com

Stefan Rüger
Knowledge Media Institute
The Open University
Milton Keynes, UK
s.rueger@open.ac.uk

ABSTRACT

With the steadily increasing amount of multimedia documents on the web and at home, the need for reliable semantic indexing methods that assign multiple keywords to a document grows. The performance of existing approaches is often measured with standard evaluation measures of the information retrieval community. In a case study on image annotation, we show the behaviour of 13 different evaluation measures and point out their strengths and weaknesses. For the analysis, data from 19 research groups that participated in the ImageCLEF Photo Annotation Task are utilized together with several configurations based on random numbers. A recently proposed ontology-based measure was investigated that incorporates structure information, relationships from the ontology and the agreement between annotators for a concept and compared to a hierarchical variant. The results for the hierarchical measure are not competitive. The ontology-based results assign good scores to the systems that got also good ranks in the other measures like the example-based F-measure. For concept-based evaluation, stable results could be obtained for MAP concerning random numbers and the number of annotated labels. The AUC measure shows good evaluation characteristics in case all annotations contain confidence values.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Experimentation, Performance, Measurement

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

1. INTRODUCTION

Recent trends in multimedia research go from the explicit categorization of media items into several classes to the annotation of media items with a variable number of concepts. For the categorization approach the evaluation per category is sufficient as each media item exclusively belongs to one class. In evaluation of media item annotation, often the same evaluation measures are applied as in the evaluation of categorization approaches. The prediction is evaluated for each concept in isolation. This allows using the well-known evaluation measures like Precision, Recall, F-measure or Accuracy. E.g., Fan et al. [9] use the Accuracy to determine the quality of the classifier for each concept.

The opposite way is to start with the media document and evaluate if all concepts are assigned correctly on a per item basis. Then instead of comparing a single predicted label to a single ground-truth label, one needs to compare two sets of labels. As a result, the predicted labels can be *fully correct* (label sets are identical), *fully wrong* (the intersection of the sets is empty), or *partly correct* (the sets have common labels, but are not fully identical).

In this paper, we compare the behaviour of different evaluation measures on the results of the ImageCLEF **L**arge-**S**cale **V**isual **C**oncept **D**etection and **A**nnotation **T**ask (LS-VCDT) 2009. It is structured as follows: In Sec. 2 a review of different performance measures for multilabel annotation evaluation including related work is given. In Sec. 3 the LS-VCDT of ImageCLEF 2009 is explained in more detail and the submissions of the participating groups are briefly outlined. In Sec. 4 the results of the case study on multilabel evaluation measures are presented and discussed. Finally we summarize our findings and future work in Sec. 5.

2. PERFORMANCE MEASURES

In multilabel classification evaluation, two paradigms for evaluation exist as briefly pointed out in Sec. 1. Tsoumakas et al. [29] name them *example-based evaluation* and *label-based evaluation*. The example-based evaluation generates a score for each media item (example) and then averages over all items. The label-based evaluation subdivides the annotations in a single evaluation per concept and averages later over all concepts. We would like to call the latter *concept-*

based evaluation throughout the paper to distinguish from the word *label* that we use for the concrete annotations for one media item (see Sec. 2.5).

2.1 General notations

In multilabel classification each data example (in our case each image) is associated with a set of labels. Let X be a dataset consisting of examples X_n, $n = \overline{1,N}$, where N is a total number of examples in the dataset. We denote the class membership for the data set X as $Y = \{y_{nc}\}$, $c = \overline{1,C}$, where C is a total number of concepts, and

$$y_{nc} = \begin{cases} 1, & \text{if example } n \text{ belongs to concept } c, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

In other words, the annotation matrix $\{y_{nc}\}$ is a binary matrix, where the rows y_n correspond to examples and the columns y_c correspond to concepts. In each matrix row y_n each non-zero element y_{nc} indicates that the example n is associated with a concept c. In our case of multilabel classification, each row of Y can have multiple non-zero values. Likewise, the non-zero elements of the column y_c indicate the examples belonging to the concept c. Alternatively, a set of labels for the example n is denoted as $\mathcal{Y}_n$.

If it is not stated otherwise, the ground-truth annotations are denoted as Y, $\mathcal{Y}_n$, y_c or y_{nc}, while the estimated annotations (suggested by the system) are denoted as Z, $\mathcal{Z}_n$, z_c or z_{nc}, respectively.

The important characteristics of a multi-labelled dataset are label *cardinality* (LC) and label *density* (LD). LC shows how many labels have been assigned to a dataset example in average:

$$LC(X) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{nc}, \quad (2)$$

and the LD is defined as fraction of the average number of the used labels to the total number of available labels:

$$LD(X) = \frac{1}{N \cdot C}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{nc}. \quad (3)$$

2.2 Concept-based evaluation measures

Concept-based evaluation measures judge the quality of annotation systems for each concept. In the following different concept-based evaluation measures are introduced.

2.2.1 Precision, Recall, F-Measure

The traditional information retrieval evaluation measures, namely Precision, Recall and F-measure are initially calculated for each concept independently. For the concept c two binary vectors z_c and y_c are compared. The number of true positive ($TP(c)$), false positive ($FP(c)$) and false negative ($FN(c)$) examples is calculated for the concept c. If the elements of the binary vectors are treated as logical values, then $TP(c)$, $FP(c)$ and $FN(c)$ can be written as:

$$TP(c) = \sum_{n=1}^{N}(z_{nc} \vee y_{nc}), \quad (4)$$

$$FP(c) = \sum_{n=1}^{N}(\neg z_{nc} \vee y_{nc}), \quad (5)$$

$$FN(c) = \sum_{n=1}^{N}(z_{nc} \vee \neg y_{nc}). \quad (6)$$

Then the average concept-based Precision, Recall and F-measure are defined as follows:

$$\text{Precision}_{cb}(Z,Y) = \frac{1}{C}\sum_{c=1}^{C}\frac{TP(c)}{TP(c)+FP(c)}, \quad (7)$$

$$\text{Recall}_{cb}(Z,Y) = \frac{1}{C}\sum_{c=1}^{C}\frac{TP(c)}{TP(c)+FN(c)}, \quad (8)$$

$$\text{F-measure}_{cb}(Z,Y) = \frac{1}{C}\sum_{c=1}^{C}\frac{2\cdot TP(c)}{2\cdot TP(c)+FN(c)+FP(c)}. \quad (9)$$

2.2.2 AUC, EER - ROC curve measures

The concept-based measures Area-Under-Curve (AUC) and Equal Error Rate (EER) can be calculated from the Receiver Operating Characteristics (ROC) curve and are common measures for different recognition tasks, e.g. [16]. A ROC curve can be used to graphically plot true-positive-rates and false-positive-rates of a binary classifier, according to different threshold values. The EER is defined as the point where both values are equal. The measure AUC describes the overall quality of a classification system independent from an individual threshold configuration, with the specific trade-off between true-positive and false-positive. It is calculated by integration of the ROC curve, whereas an AUC value of 1 equals a perfect system with no false positives and an AUC value of 0.5 equals a random system. For the evaluation per concept, the EER and the AUC of the ROC curves summarize the performance of the individual runs by taking the average values of all concepts.

2.2.3 MAP - Precision-Recall Curve measure

The measure Mean Average Precision (MAP) is a concept-based measure that approximates the area under the uninterpolated Precision-Recall Curve averaged over several information needs. In other words, it is the average of the Precision values calculated after each relevant document is retrieved for a single query and later averaged over queries. MAP is often used as a single-value measure that summarizes the quality across recall levels of ranked retrieval results and is e.g. utilized as standard evaluation measure in the TREC community. For a detailed explanation see [20].

2.3 Example-based evaluation measures

Example-based evaluation measures judge the quality of annotation systems for each media item. In the following different example-based evaluation measures are introduced.

2.3.1 Precision, Recall, F-Measure, Accuracy

In contrast to the concept-based variants of Precision, Recall and F-measure, the scores are firstly calculated for each example and then averaged over all examples. In this work, we use example-based Precision, Recall, F-measure and Accuracy as suggested in [29]:

$$\text{Precision}_{eb}(Z,Y) = \frac{1}{N}\sum_{n=1}^{N}\frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Z}_n|}, \quad (10)$$

$$\text{Recall}_{eb}(Z,Y) = \frac{1}{N}\sum_{n=1}^{N}\frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n|}, \quad (11)$$

$$\text{F-measure}_{eb}(Z,Y) = \frac{1}{N}\sum_{n=1}^{N}\frac{2 * |\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n| + |\mathcal{Z}_n|}, \quad (12)$$

$$\text{Accuracy}_{eb}(Z,Y) = \frac{1}{N}\sum_{n=1}^{N}\frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n \cup \mathcal{Z}_n|}. \quad (13)$$

2.3.2 Alpha Evaluation

Shen et al. [28] propose an α-evaluation and multilabel class Recall and Precision. α-evaluation generates a score while taking the ground-truth, predicted labels, missed labels and false positive labels into account. Moreover, false positives and missed labels can be penalized differently if it is more suitable for the particular application. The parameter α introduces the so-called *forgiveness rate* as a trade-off between the fully correct and partly correct prediction. In Eq. 14, the α-evaluation formula for equally treated false positives and missed labels is depicted.

$$\text{score}(\mathcal{Z}_n, \mathcal{Y}_n) = \left(\frac{|\mathcal{Y}_n \cap \mathcal{Z}_n|}{|\mathcal{Y}_n \cup \mathcal{Z}_n|}\right)^{\alpha} \quad \alpha \geq 0 \quad (14)$$

$$\text{Accuracy}_{\alpha}(Z,Y) = \frac{1}{N}\sum_{n=1}^{N}\text{score}(\mathcal{Z}_n, \mathcal{Y}_n) \quad (15)$$

If $\alpha = 1$, Eq. 15 is equal to the example-based Accuracy measure as depicted in Eq. 13.

2.4 Hierarchical Evaluation Measures

Hierarchical evaluation measures stick to the paradigm of example-based evaluation. Next to the ground truth annotations of the media items (examples), a hierarchical structure of the concepts to be detected is required. Different hierarchical measures for unilabel classification are summarized in [11]. Intuitively, the concepts, that are located near in a hierarchy are more similar than the ones that are located far. The idea is to judge an annotation from the predictor that does not match exactly to the ground truth by their distance in the hierarchy.

The most important measures are the *depth independent distance-based misclassification costs* and the *depth dependent distance-based misclassification costs*. In the former case, the predicted concept is compared to the correct one and the number of edges of the shortest path in the hierarchy between both are counted. In the latter case, an additional weight is assigned to each edge in the hierarchy. So, misclassifications in deeper levels of the hierarchy get lower costs assigned than at an upper level.

Blockeel et al. [3] propose an evaluation measure for hierarchical multilabel classification evaluation. They extrapolate distances between individual labels to distances between sets of labels by mapping the feature vectors of the sets into Euclidean space where the individual labels form the base vectors. In [4] a hierarchical loss function is proposed that considers classification into a hierarchy with multiple and partial paths. The first wrongly classified node is regarded as mistake and adds to the loss. The mistakes that are made in classification below the first wrongly classified node are not considered as loss anymore. Underlying is the assumption that for each classification a path from root to leaf or from root to an internal node is present. They compare their work to the zero-one loss and symmetric-difference loss.

2.5 Ontology-based Evaluation Measures

The challenge in example-based multilabel classification evaluation is the way how to deal with partly incorrect label sets. In Sec. 2.3 the score is derived by counting the numbers of labels in both sets and calculating a fraction of different sets. In Sec. 2.4 the measures rely on a path in an associated taxonomy and base their score on the length of the path between concepts. Further, there are approaches that do not use the length of the path between concepts, but calculate a semantic similarity between concepts.

One early approach to calculate semantic similarity between concepts in a *is-a* taxonomy was proposed by Resnik [26]. He uses the information content to derive a similarity and compares the semantic similarity to the hierarchical depth-independent misclassification costs. In [1] a comparison of different ontology based similarity measures was conducted. They compare ontology-based distance approaches with information-theoretic approaches, vector space methods and algorithms based on Levensthein distance. Lord et al. [18] compare semantic similarity measures and their application to ontological annotations on the example of the Gene Ontology.

However, except for the measures explained in Sec. 2.2, the difficulty is still to find the matching labels in both sets between which the similarity (or distance) should be computed. In [25], we propose an ontology-based score (OS) for the evaluation of multilabel annotations incorporating ontology information. The OS considers three different characteristics:

1. Depth-dependent distance-based misclassification costs between concepts

   We extended the depth-dependent distance-based misclassification costs calculation from the unilabel case to the multilabel case. As the overlap of the system output and the ground truth are in most cases partly correct, we defined a matching procedure that maps labels from the system annotations to the ground truth and vice versa. The distance between mapped labels is calculated and a cost dependent on the depth in the hierarchy is assigned. The calculation of misclassification costs favours systems that annotate a photo with concepts close to the correct ones more than systems that annotate concepts that are far away in the hierarchy from the correct concepts.

2. Ontology-based penalty

   Next to the *is-a* relationship in the ontology, different relationships between concepts are defined. The ontology restricts e.g., that for a certain sub-node only one concept can be assigned at a time (*disjointness*) or that concepts postulate other concepts. If relationships in Z are violated, not the depth-dependent costs are used, but a penalty is assigned.

3. Annotator Agreements

   The annotator agreements serve as scaling factor for the costs. Due to the difficulty to judge some concepts

objectively, the OS considers the agreements between annotators for rescaling the costs. In a user study 11 annotators annotated all concepts of the ontology in a small set of photos. The annotation of the majority of annotators was regarded as correct and the percentage of annotators that annotated correctly equals the agreement factor (see [23]). The outcome of the user study is an agreement-map. The factor 1 denotes total agreement on a concept over the whole photo set. If a system label is mapped to a concept of the ground truth with low agreement, the costs are scaled down.

Formalized, the matching procedure for each example n between the predicted set of labels $\mathcal{Z}_n$ and the ground-truth set of labels $\mathcal{Y}_n$ is defined as follows:

Each set contains labels l_i respectively l_j that are assigned to a multimedia document X_n. First, the false positive labels $\mathcal{Z'}_n = \mathcal{Z}_n \setminus (\mathcal{Z}_n \cap \mathcal{Y}_n)$ and the missed labels $\mathcal{Y'}_n = \mathcal{Y}_n \setminus (\mathcal{Z}_n \cap \mathcal{Y}_n)$ are computed. Please note that $|\mathcal{Z'}_n| + |\mathcal{Y'}_n| \leq |\mathcal{Z}_n \cup \mathcal{Y}_n|$ is always valid, because the number of false positive and missed labels can never be greater than the number of the union of labels in both sets. A crosscheck on the predicted label set Z_n is performed. If labels in $\mathcal{Z}_n$ violate relationships from the ontology, these labels get the maximum costs of 1 as penalty assigned and are removed from $\mathcal{Z'}_n$, $\mathcal{Y}_n$ and $\mathcal{Y'}_n$ if contained. This ensures that the measure does not assign costs two times. Next, for each label l_i from $\mathcal{Z'}_n$ a match to a label l_j from $\mathcal{Y}_n$ is calculated and for each label l_j from $\mathcal{Y'}_n$ a mapping to a label l_i from $\mathcal{Z}_n$ is performed in an optimization procedure (see Eq. 16). The costs between two labels l_i and l_j depend on the shortest path in the hierarchy between both concepts. Each link is associated with a cost that is cut in halves for each deeper level of the tree and is maximal equal to 1 for a path between two leaf nodes of the deepest level. The costs for a link at level l of the hierarchy are calculated as $\text{cost link}_l = \frac{2^{(l-1)}}{2^{(L+1)}-2}$ with L as the number of links from the leaf to the root. If $Z_n = \emptyset$, the matching costs for all labels l_j of $\mathcal{Y'}_n = \mathcal{Y}_n$ are set to the maximum. The matching costs are computed as follows:

$$\begin{aligned} \text{match}(\mathcal{Z}_n, \mathcal{Y}_n) = &\sum_{l_i \in \mathcal{Z'}} \left((\min_{l_j \in \mathcal{Y}} \text{cost}(l_i, l_j)) \cdot a({l_j}^*) \right) \\ &+ \sum_{l_j \in \mathcal{Y'}} \left((\min_{l_i \in Z} \text{cost}(l_i, l_j)) \cdot a(l_j) \right), \end{aligned} \tag{16}$$

with ${l_j}^* = \text{argmin}_{l_j \in Z}(\text{cost}(l_i, l_j))$.
$a(l_j)$ determines the *annotation agreement factor* for a concept l_j and ranges between $[0, 1]$. The final score for each multimedia document X_n is based on the matching costs between $\mathcal{Z}_n$ and $\mathcal{Y}_n$ divided by the number of different concepts in both label sets (see Eq. 17). The score is 1 if all concepts are correctly annotated and goes to 0 if no concept was found. Additionally, Shens α-factor, ($\alpha \geq 0$), introduced in Sec. 2.3.2, can be incorporated to weight the strictness of the score regarding fully and partly correct annotations, depending on the application demands.

$$OS(Z, Y) = \frac{1}{N} \sum_{n=1}^{N} \left(1 - \frac{\text{match}(\mathcal{Z}_n, \mathcal{Y}_n)}{|\mathcal{Z}_n \cup \mathcal{Y}_n|} \right)^{\alpha} \tag{17}$$

In the experiments reported in this paper α is set to 1. The measure is called Hierarchical Score (HS) if the crosscheck on the system annotations $\mathcal{Z}_n$ is not performed. Then the measure only includes the structure information of the ontology and the annotator agreement factors.

3. LS-VCDT IN IMAGECLEF 2009

ImageCLEF is an evaluation track that belongs to the **C**ross **L**anguage **E**valuation **F**orum (CLEF) and poses yearly benchmarks in the area of image retrieval and annotation. In our case study we refer to the results of the LS-VCDT of ImageCLEF 2009 (see [24]). The task of the participants in the LS-VCDT was to annotate a set of 13.000 photos from the MIR Flickr 25.000 image dataset [14] with 53 visual concepts. For the training of the algorithms, a set of 5.000 photos with annotations and a Photo Tagging Ontology was provided. The frequencies of the concept occurrence in training and test set is depicted in Fig. 1. The Photo Tagging Ontology could be used to solve the annotation task by e.g. taking advantage of the relations between concepts and their hierarchical ordering. For more information about the ontology and the visual concepts see [23]. As evaluation measures, the two paradigms of concept-based and example-based annotation evaluation were followed in the official ImageCLEF LS-VCDT results. For the concept-based evaluation the EER and AUC were calculated from ROC curves as explained in Sec. 2.2.2. The evaluation on example basis was performed with the OS as illustrated in Sec. 2.5. We had submissions from 19 research groups with altogether 73 run configurations.

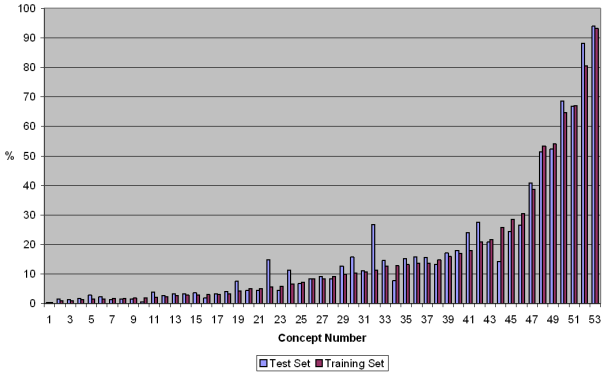

Figure 1: The figure shows the percentage in which the concepts occurred in training and test set of the LS-VCDT.

3.1 Choice of configurations for Case Study

In the case study about the characteristics of evaluation measures in image annotation, not always all submitted configurations are used due to a more intuitive visualization of the results. In some plots just one configuration per group is utilized. For that not the best submitted configuration of every group, but the configuration with the largest variance between the result ranks for both measures is chosen. Each group was allowed to submit up to five configurations. In Table 1, part of the official results for the EER and the OS measure are displayed. The 19 configurations and one random configuration were chosen out of the 73 configurations because of their variance in the ranking of the results

System	Rank	EER	Rank	OS	Rank Diff
System 9	1	0.234	21	0.740	-20
System 11	5	0.250	23	0.731	-18
System 5	11	0.256	2	0.810	+9
System 19	14	0.267	1	0.811	+13
System 6	15	0.272	9	0.793	+6
System 3	17	0.292	40	0.613	-23
System 13	24	0.331	53	0.482	-29
System 7	26	0.342	67	0.368	-41
System 14	31	0.357	66	0.376	-35
System 12	35	0.384	72	0.261	-37
System 15	40	0.440	28	0.716	+12
System 18	46	0.452	11	0.779	+35
System 2	48	0.454	62	0.396	-14
System 10	53	0.467	10	0.786	+43
System 17	54	0.479	32	0.690	+22
System 1	56	0.483	17	0.756	+39
System 8	59	0.486	20	0.744	+39
Random 0	-	0.500	-	0.384	-
System 4	70	0.502	29	0.709	+41
System 16	73	0.527	65	0.385	+8

Table 1: The table shows the configurations with the largest variances in the ranking of the official LS-VCDT results for the measures EER and OS. One configuration was selected from each participating group for the case study.

of both measures. As we would like to analyse the differences and weaknesses of the evaluation measures, we believe that these are the most interesting configurations for the case study. We exemplarily confirm that the characteristics of evaluation measures are valid for all submitted runs.

3.2 Image annotation approaches

In the following, a brief summary of the technologies submitted by the participants is provided. The systems of the participants are numbered from 1 to 19 in alphabetic order [12, 5, 22, 2, 13, 7, 30, 27, 6, 8, 31, 17, 21, 10, 15].

Most systems follow this processing workflow: Feature extraction, feature reduction, classification and post processing. The majority of the groups considers the hierarchy and ontology information in the post processing step and applies specific rules to fulfil the requirements for disjoint concepts or branches in the hierarchy. Some systems integrate the hierarchy information already in the classification and feature selection process, whereas others do not consider the ontology requirements at all.

Within the feature extraction, several systems with leading ranks have in common that they make extensive use of keypoint based local edge direction histograms, e.g. SIFT features [19]. These features can describe precisely a specific structure of a local pattern. Most approaches extend the local features with global color or edge histograms or Gist of Scene features to gather information about the overall visual impression of the photos. Additionally, e.g. System 5, applies an individual region finder for each concept to determine the most relevant image region for the specific concept. The local and global features are then calculated from the selected region and from the complete image.

A selection or reduction process is applied most of the times due to the high dimension of these feature combinations with partly more than 1.000 dimensions. Certain groups utilize a codebook or visual words approach to cluster the high dimensional feature space, e.g. with a k-means algorithm. Others apply a feature selection algorithm, individually trained for each ontology concept. Such selection process prefers, e.g. global features for the concepts "summer" or "night", while SIFT features are preferred for concepts like "trees".

For classification, most groups conduct a SVM classifier, which was individually trained for each concept. The individual SVM parameters were estimated within a cross validation using the provided training data. The classification is partly conducted as one-against-all or multi-class SVM process. Differences between the groups are related to the used SVM details, e.g. System 6 utilizes SVM with average kernels, sparse L^1 multiple kernel learning (MKL) and non-sparse L^p MKL. Other groups, e.g. System 19 use sparse logistic regression classifiers.

The biggest differences between the systems can be found in the post processing. This step covers the interpretation of the classification results according to the definitions in the ontology, e.g. selecting one concept in a group of disjoint concepts. A careful applied post processing can lead to benefits in terms of the OS measure as can be derived from the increased rank numbers in Table 1. Exemplarily, System 8 could achieve a 39 ranks better result by a label refinement process, which utilizes a co-occurrence statistics of the disjoint concepts. In contrast, the leading System 9 considering the EER measure, looses 20 ranks because of incorporating no post processing that deals with disjoint concepts. System 5 performs a simple manipulation of the result scores to have only a single concept labelled within a group of disjoint concepts, which leads to an improvement of 9 ranks.

4. CASE STUDY ON EVALUATION MEASURES FOR IMAGE ANNOTATION

This section presents the results of the case study on evaluation measures for image annotation. For all 19 teams from the LS-VCDT challenge, one configuration per team was chosen as explained in Sec. 3.1. All submissions contained confidence values between [0:1] for each concept in each photo. It was agreed upon a threshold of 0.5 for the measures that need a binary decision to judge presence or absence of a concept. Additionally, we investigated several random configurations. The configuration Random0 stands for the results of uniformly distributed pseudo random numbers that varied between [0:1]. All other random runs are denoted by RandomXX where XX stands for the percentage of 1 values in the annotations.

The results refer to the test set of the LS-VCDT. It consists of 13.000 photos and was annotated with 53 concepts by human judges. The occurrence of concepts in the test and training set can be seen in Fig. 1. While in most cases the frequency of a concept in training and test set was stable, the occurrence of concepts over the whole data set varied extremely. The ground truth annotations of the test set show a label cardinality $LC = 9.0554$ and a label density $LD = 0.1709$. This means that in average per photo 9 labels were assigned by the human judges.

4.1 Results

In the following, the results of the case study are presented. We first focus on the results of the chosen run con-

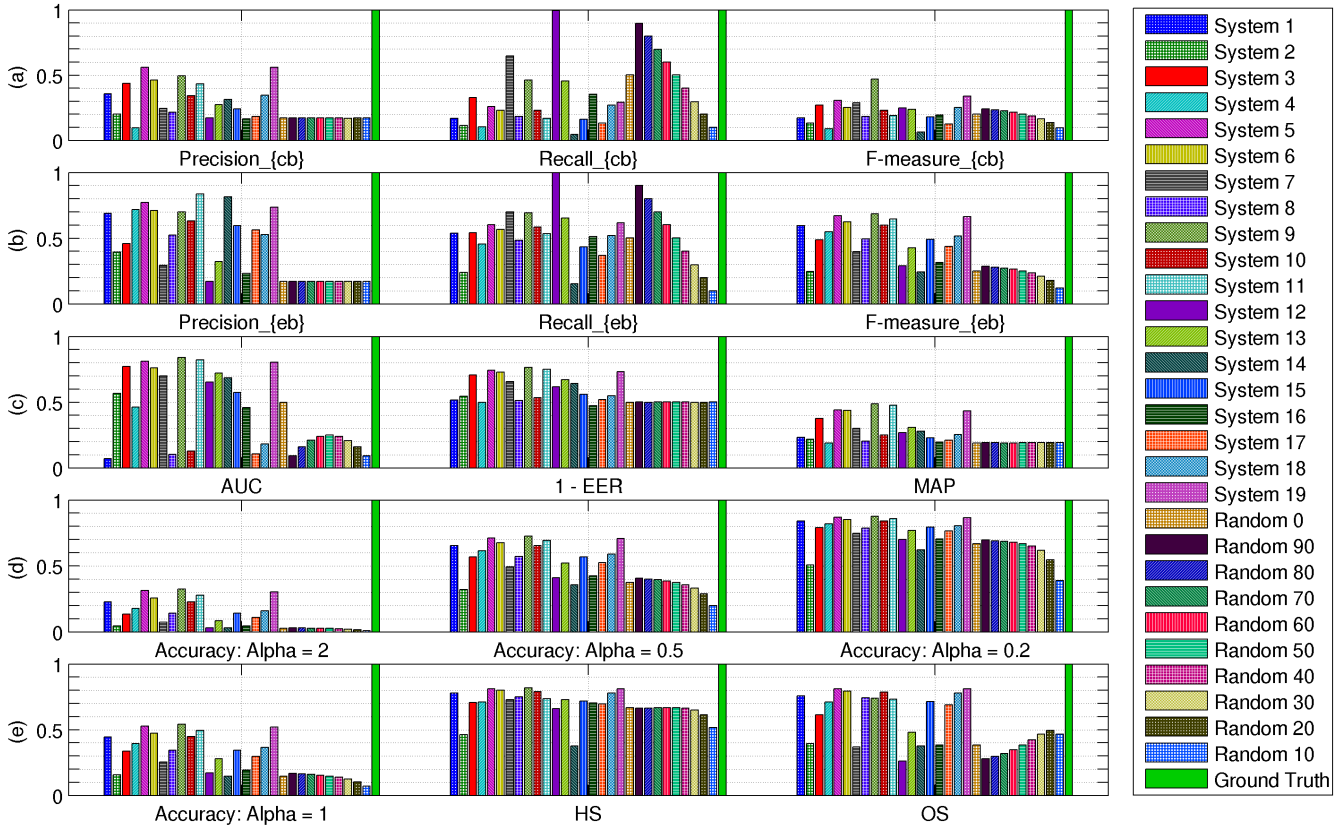

Figure 2: This figure illustrates the results of the concept-based and example-based evaluation measures for the chosen run configurations. The results in the diagrams are ordered ascending to the system numbers, followed by the random runs and the ground truth result for each measure.

figurations and second prove that the characteristics of the evaluation measures can be transferred to all runs.

Fig. 2 illustrates the results of the evaluation of the chosen run configurations. The first row depicts the results for concept-based Precision, Recall and F-measure. Contrasting row (b) shows the results for the example-based variants of Precision, Recall and F-measure. Row (c) depicts the scores of AUC, EER and MAP. For an easier comparison, the results of 1-EER are visualized. Row (d) shows the scores for the α-evaluation measure with different values for α and finally row (e) presents the example-based Accuracy, HS and OS scores. In each bar diagram the same order of runs is utilized, beginning with systems 1-19, followed by the random runs and the ground truth.

In Fig. 2 (a) the results for the concept-based Precision, Recall and F-measure are depicted. The Precision for the submissions varies between 0.1 and 0.6 with an average of 0.3. The random runs achieve a precision of 0.17. In terms of Recall, the submissions score at minimum 0.05 and 0.99 at maximum with a mean of 0.3. System 12 is the system that achieves almost a Recall of 1. This means that nearly all concepts were annotated as present for each image. This fact is also illustrated by the LD which is 0.99 for System 12. The reason for this behaviour can be two-fold. First, the system really annotates all concepts as present or second, the threshold of 0.5 which is used to map the confidence values to a binary decision is not well selected for this system. Depending on the number of annotated labels the random runs get at most a score of 0.9 Recall in case 90% of the annotations are set to 1. The F-measure combines the scores of Precision and Recall. For the systems the F-measure varies between 0.07 (System 14) and 0.47 (System 9) with a mean of 0.22 and the random runs get at most a F-measure of 0.24. These values indicate that with random runs containing a high percentage of annotated concepts a score higher than the average of the submissions can be achieved. This also holds when regarding all submissions to the task. The mean of all submissions in terms of F-measure is 0.20.

In Fig. 2 (b) the results of the example-based Precision, Recall and F-measure are depicted. Compared to the concept-based ones, one can say in general, that the example-based evaluation measures report higher scores, e.g. for the systems 4, 8 and 14. The average values are 0.56 Precision, 0.54 Recall and 0.49 F-measure. For the random configurations one can see little difference between the example-based and concept-based Precision measure. For both evaluation paradigms the random configurations do not achieve better results than 0.18 Precision. Having a look at the plots in row (a) und (b), especially at the example-based and concept-based F-measures, it is interesting, that these measures differ not only in scale, but also in the order of the systems. This finding can be explained by different averaging methods. The concept-based F-measure uses averaging over all concepts and all concepts are weighted equally, so that several badly estimated concepts could drastically lower the average score. Choosing the concept-based F-measure, the

systems and the random runs are mixed all over the result list. For the example-based F-measure, the random runs are grouped at the end of the result list with the lowest scores, except for System 14. Therefore, it can be derived, that using the example-based evaluation, a simple adaption of thresholds of randomly generated scores can not achieve high ranked results, which is in the concept-based F-measure an indicator for manipulation possibilities.

The results for the concept-based measures AUC, EER and MAP are presented in row (c). These measures are calculated based on ranked annotation results and use the confidence values that were provided by the participants. The results show that in terms of AUC, scores between 0.07 and 0.84 were achieved with a mean of 0.54 for the submissions. The random runs get at most a score of 0.5 in case of the run with random numbers in the interval [0:1] and worse (< 0.25) in case of binary runs. In terms of 1-EER, the scores range between 0.47 and 0.77 and the random runs get a score of 0.5. In terms of MAP, the mean score is 0.31, ranging between 0.19 and 0.49. Randomly a score of 0.19 can be achieved at maximum.

Fig. 2 (d) illustrates the results of the α-evaluation measure with parameter values (2, 0.5, 0.2) for α. The results for α=1, which equals the example-based Accuracy, are depicted in Fig. 2 (e). It is apparent that with smaller values of α, the results of the systems get better as the measure becomes more forgiving. Despite of System 14, all lower ranks are occupied from the random configurations. For $\alpha = 0.2$ even with random runs a score of about 0.65 can be achieved. It is also obvious that with the α-evaluation not the whole range of values can be achieved by the systems. E.g., for α=2, the best system (despite the ground truth) achieves a score of 0.32 and the worst of about 0.01. That is a difference of 0.31. With decreasing α the interval grows over 0.47 ($\alpha = 1$) to 0.53 ($\alpha = 0.5$) and decreases to 0.49 ($\alpha = 0.2$).

The results for HS and OS are illustrated in Fig. 2 (e). A classification score of about 0.65 can be achieved with the random configurations in terms of HS. These results show, that the HS can not differentiate between good and bad classification systems. The OS measure reports better results. The system scores vary between 0.26 and 0.81 with an average of 0.63 in terms of OS. The random runs achieve at most a value of 0.49. The OS tends to give good results if the annotations show a density which is comparable to the LD of the ground truth. This can be observed e.g. for System 17, which has a low value in terms of concept-based F-measure (0.12), a bad score in terms of AUC (0.11), but an OS of 0.69 and a LD of 0.11. In comparison, the LD of the ground truth is 0.17. Systems that get low scores in the AUC and F-measure can achieve good results in the OS if they stick to the ontology rules. Good systems in terms of AUC and F-measure remain with good results in the OS.

The scatter plots in Fig. 3 visualize the results of all 73 submitted configurations, the 10 random configurations and the ground truth. Exemplarily pairs of evaluation measures were chosen and plotted against each other. The runs that were utilized in the bar diagrams of Fig. 2 are denoted as big circles and the other submissions as small circles. Further, the crosses represent the random configurations and the star depicts the score of the ground truth. The indices attached to each symbol denote the name of the run.

In Fig. 3 (a), (b) and (c) the example-based Precision, Recall and F-measure are compared with their concept-based counterparts. The analysis of the results as outlined in the discussion of the bar diagrams can be transferred to all runs. In general, the example-based variants assign higher scores. The concept-based Recall is slightly higher for two systems than the example-based one. For many systems the assigned labels are uniformly distributed over the examples while the Recall values for different concepts vary significantly. Due to different averaging methods, this results in a higher example-based Recall, but a lower concept-based Recall. In case of the two systems with higher concept-based Recall than example-based, the Recall values over the concepts are uniformly distributed. The random runs achieve the same values for both variants in terms of Precision and Recall. The Precision score is low, but the Recall goes to 1 with rising number of annotations. The random runs get scores assigned that are in the average of all scores in terms of the concept-based F-measure. In contrast, the scores range at the lower bottom of the result lists for the random runs in terms of example-based F-measure. From the plots (k) and (l) one can derive that the F-measure is not fundamentally influenced by the number of labels annotated. The random runs with a high percentage of annotations just get slightly better scores than the ones with a lower percentage. It is also obvious that the correct percentage of annotations as in the ground truth does not increase the score.

The characteristics of the MAP measure are compared to the concept-based Precision and EER and in the plots (d) and (e), respectively. The scores for the concept-based Precision and MAP are in agreement. The MAP tends to assign stricter scores which is clearly observable compared to the EER. For both measures the random runs get lower scores assigned than the systems. Independent which percentage of annotations are randomly set to 1, the random runs get the same score assigned which is amongst the lowest score of all submissions. The MAP measure is therefore a stable measure which is robust against manipulations from random runs similar as the concept-based and example-based Precision.

Scatter plot 3 (f) shows the measure 1-EER compared with AUC. Two clusters are visible in the plot. The cluster at top contains the results for systems that submitted confidence values for each annotation. For this cluster AUC and 1-EER correlate. In some configurations the annotations were submitted as binary values. The scores for these systems cluster at the bottom of the plot. In terms of EER the scores between the cluster at the bottom and the worst runs with confidence values are similar. Whereas for the AUC all runs with binary decisions get at most a score of 0.25 and the runs with confidence values get at least a score of 0.46. This can e.g. be observed for the run with pseudo random numbers between [0:1] which receives a score of 0.5. Summarizing, the AUC measure disadvantages submissions containing binary decisions.

The plots (g), (h), (i) and (j) illustrate the characteristics of the OS measure. As already outlined in the discussion of the bar diagrams, the HS measure cannot differentiate between random runs and submissions of average performance. The OS measure assigns the lowest scores to the random runs. The behaviour of OS and HS is correlated for the top submissions. Keeping in mind that the difference between both measures are the penalties of the ontology, it is obvious that the top systems are not penalized to a great

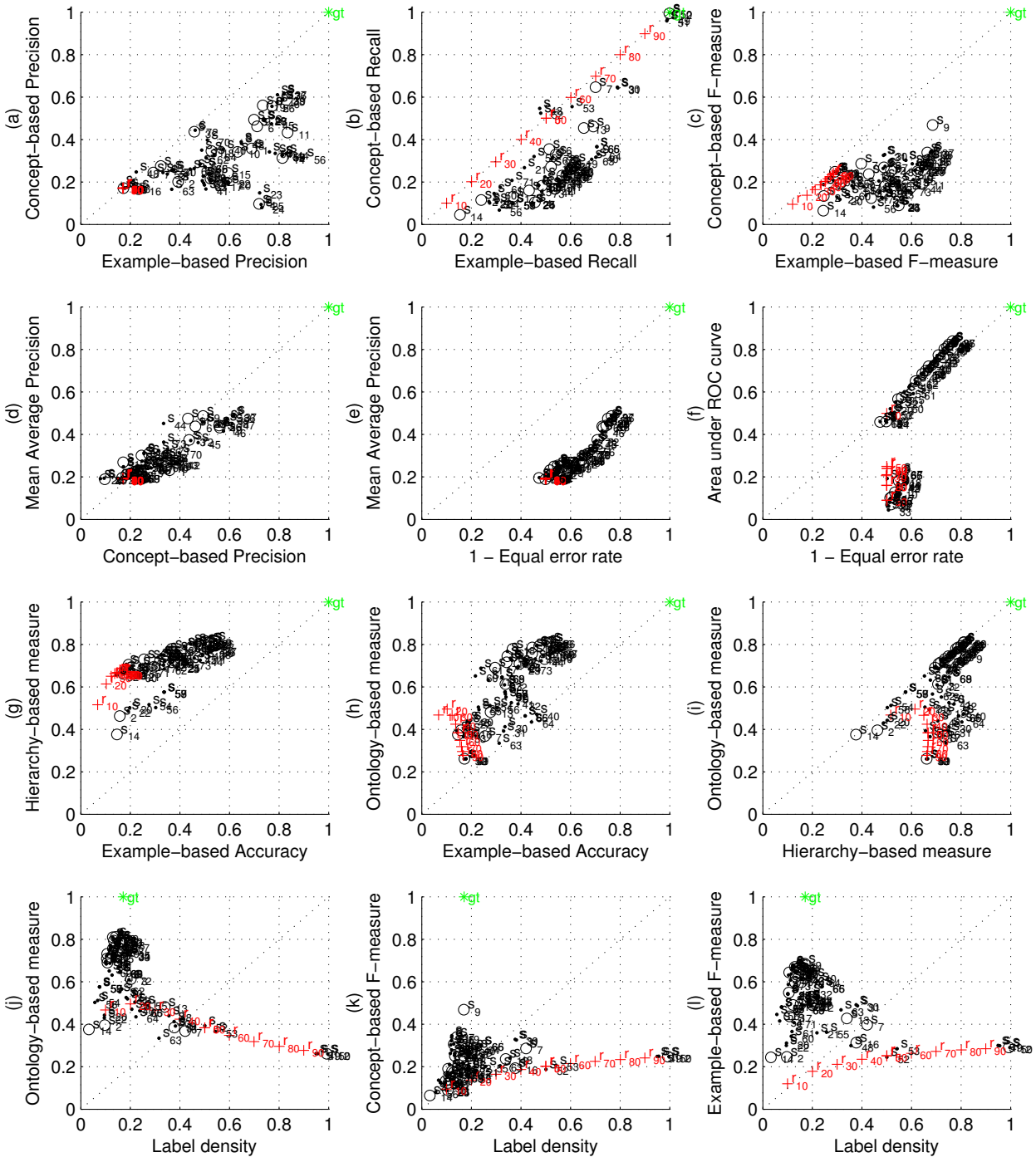

Figure 3: The scatter plots visualize the results for all run configurations for some of the evaluation measures. The submitted run configurations are presented as circles. Big circles denote the runs that were utilized in the bar plots of Fig. 2 and small circles were utilized for all other submissions. The results for the random runs are presented as crosses and the score of the ground truth as star.

extent. Therefore both measures assign similar scores. In contrast to HS and OS, the example-based Accuracy assigns stricter scores. The OS measure tends to give good results to configurations with similar LD as the ground truth. The LD of the dataset is 0.17. This means in average 17% of the concepts are annotated per photo. The random runs with 30%, 20% and 10% annotated concepts, get the highest scores compared to the other random runs. But through the restrictions of the relationships the results of these random runs are nevertheless not better than 0.5.

4.2 Discussion

In the previous section, the results of 13 evaluation measures are presented and analysed. The following enumeration summarizes the most important characteristics of evaluation measures that were found in our case study:

- Example-based evaluation: HS and OS

  The HS does not satisfy the needs of a good evaluation measure. With random numbers good results can be achieved and the difference to results of well-working classification systems is not apparent. The OS assigns good scores to systems that got also good ranks in the other measures like the example-based F-measure. However, it tends to give better results to systems that follow all ontology rules but got only average ranks in the traditional measures. Further, the OS assigns better scores if a number of concepts is annotated which is close to the LC of the ground truth.

- Example-based evaluation: α-evaluation

  The results of the α-evaluation are dependent on the threshold chosen for α. If α is equal to 1 or to 0.5, the results for our case study show the best distribution without assigning good scores to the random runs.

- Precision, Recall and F-Measure

  The example-based Precision, Recall and F-measure assign higher scores as their concept-based counterparts. As outlined, this is due to the uniformly distributed annotation quality per example for most systems, but the variying quality per concept. While the concept-based Precision shows good evaluation characteristics, the concept-based Recall is not adequate as evaluation measure for multilabel evaluation. Further, our study shows that a simple adaption of thresholds of randomly generated scores can achieve high ranked results in an evaluation with the concept-based F-measure. In contrast, the scores for random runs are not of comparable quality in comparison to the scores achieved by systems for evaluation with the example-based F-measure. In both cases there is no major influence on the scores by the number of labels.

- Concept-based evaluation: AUC

  AUC clusters the scores for submissions with binary values and submissions with confidence values in two clusters. In consequence of the definition of the measure, binary submissions can get significantly worse results than 0.5 as achieved in case random values in the interval [0:1] are used. This leads to the conclusion that either confidences for the annotations should be applied as default in a benchmark scenario or another measure should be (additionally) utilized, as the AUC does not allow for a comparison of both kinds of submissions.

- Concept-based evaluation: MAP

  The scores obtained by MAP show a correlation to the concept-based Precison and the EER. Especially compared to EER it assigns stricter scores. Further, MAP is a stable measure as it is robust against manipulations from random runs and not dependent on the percentage of labels that are set.

In summary, no preference to example-based or concept-based evaluation measures can be given. Depending on the application the one or the other may be more appropriate. If using example-based evaluation, we suggest to use the example-based F-measure, Accuracy or OS, as depending on the needs these measures show good characteristics. The HS should not be utilized as evaluation measure. Also the adaptation of the α-evaluation with different parameters cannot convince for α values despite 1. For concept-based evaluation, we recommend using MAP. If it is assured that the evaluation is performed solely on submissions with confidence values, the AUC measure shows good characteristics.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented the results of a multilabel image annotation benchmark and analysed its results with several evaluation measures. We highlighted the differences between concept-based and example-based multilabel evaluation and the results achieved. Altogether 13 evaluation measures were utilized to establish a profound comparison about their strengths and weaknesses.

Concluding, the example-based F-measure, Accuracy or OS showed promising results in terms of example-based evaluation. In contrast, the HS cannot cope with traditional evaluation measures. Also the adaptation of the α-evaluation with different parameters cannot convince for α values despite 1. For concept-based evaluation, we recommend using MAP. The AUC measure also shows good evaluation characteristics in case all annotations contain confidence values.

In case of confidence-based annotation, all presented evaluation measures need a threshold to obtain a binary decision about the presence and absence of concepts despite AUC, EER and MAP. These thresholds have a major influence on the results of the evaluation.

In contrast to all other evaluation measures, the OS does not only perform a binary decision when comparing label sets, but calculates scores for each label also when just contained in one of both sets. This evaluation approach seems promising, as concepts annotated semantically close to the correct one are not regarded as incorrect, but partly correct. The degree of correctness is deducted from the length of the path in the ontology between both concepts. Because of that, it is important to find a common agreed way of structuring the concepts in the ontology in future work. This could be performed e.g. with the help of user studies or by utilizing a method based on semantic similarity to calculate costs. It has to be investigated if the existing semantic ontology-based measures that were tested on quite different ontologies in size and structure like Gene Ontology or WordNet, can be applied to the Photo Tagging Ontology.

Further, the OS would benefit if it is independent from a threshold for confidence-based annotations.

To improve the concept-based evaluation measures one could modify the averaging procedure by introducing additional weights to the concepts according to their importance and the annotator agreement.

6. ACKNOWLEDGMENTS

This work was supported by grant 01MQ07017 of the German research program THESEUS funded by the Ministry of Economics. The work was funded by a German Academic Exchange Service (DAAD) scholarship and partly performed at Knowledge Media Institute at Open University.

7. REFERENCES

[1] A. Bernstein, E. Kaufmann, C. Bürki, and M. Klein. How similar is it? Towards personalized similarity measures in ontologies. In *7th Intern. Conference Wirtschaftsinformatik, Germany*. Springer, 2005.

[2] A. Binder and M. Kawanabe. Fraunhofer FIRST's Submission to ImageCLEF2009 Photo Annotation Task: Non-sparse Multiple Kernel Learning. *CLEF working notes*, 2009.

[3] H. Blockeel, M. Bruynooghe, S. Džeroski, J. Ramon, and J. Struyf. Hierarchical multi-classification. In *SIGKDD Workshop on Multi-Relational Data Mining*, pages 21–35, 2002.

[4] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Incremental algorithms for hierarchical classification. *Journal of Machine Learning Research*, 7:31–54, 2006.

[5] B. Daroczy, I. Petras, A. Benczur, Z. Fekete, D. Nemeskey, D. Siklosi, and Z. Weiner. SZTAKI @ ImageCLEF 2009. *CLEF working notes*, 2009.

[6] M. Douze, M. Guillaumin, T. Mensink, C. Schmid, and J. Verbeek. INRIA-LEARs participation to ImageCLEF 2009. *CLEF working notes*, 2009.

[7] H. Escalante, J. Gonzalez, C. Hernandez, A. Lopez, M. Montex, E. Morales, E. Ruiz, L. Sucar, and L. Villasenor. TIA-INAOE's Participation at ImageCLEF 2009. *CLEF working notes*, 2009.

[8] A. Fakeri-Tabrizi, S. Tollari, L. Denoyer, and P. Gallinari. UPMC/LIP6 at ImageCLEF annotation 2009: Large Scale Visual Concept Detection and Annotation. *CLEF working notes*, 2009.

[9] J. Fan, Y. Gao, H. Luo, and R. Jain. Mining multilevel image semantics via hierarchical classification. *IEEE Trans. on Multimedia*, 10(2):167, 2008.

[10] M. Ferecatu and H. Sahbi. TELECOM ParisTech at ImageClef 2009: Large Scale Visual Concept Detection and Annotation Task. *CLEF working notes*, 2009.

[11] A. Freitas and A. de Carvalho. A tutorial on hierarchical classification with applications in bioinformatics. *Intelligent Information Technologies: Concepts, Methodologies, Tools and Applications*, 2007.

[12] H. Glotin, A. Fakeri-Tabrizi, P. Mulhem, M. Ferecatu, Z. Zhao, S. Tollari, G. Quenot, H. Sahbi, E. Dumont, and P. Gallinari. Comparison of Various AVEIR Visual Concept Detectors with an Index of Carefulness. *CLEF working notes*, 2009.

[13] J. Hare and P. Lewis. IAM@ImageCLEF Photo Annotation 2009: Naive application of a linear-algebraic semantic space. *CLEF working notes*, 2009.

[14] M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. In *MIR '08: Proceedings of the 2008 ACM Intern. Conf. on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[15] A. Iftene, L. Vamanu, and C. Croitoru. UAIC at ImageCLEF 2009 Photo Annotation Task. *CLEF working notes*, 2009.

[16] Y. Liu and E. Shriberg. Comparing evaluation metrics for sentence boundary detection. In *Intern. Conf. on Acoustics, Speech and Signal Processing*, 2007.

[17] A. Llorente, S. Little, and S. Rüger. MMIS at ImageCLEF 2009: Non-parametric Density Estimation Algorithms. *CLEF working notes*, 2009.

[18] P. Lord, R. Stevens, A. Brass, and C. Goble. Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. volume 19. Oxford Univ Press, 2003.

[19] D. Lowe. Object recognition from local scale-invariant features. In *Intern. Conf. on Computer Vision*, volume 2, pages 1150–1157. Corfu, Greece, 1999.

[20] C. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval [Draft]*. Cambridge, UK: Cambridge University Press, April 2009. http://www.informationretrieval.org/.

[21] P. Mulhem, J.-P. Chevallet, G. Quenon, and R. Al Batal. MRIM-LIG at ImageCLEF 2009: Photo Retrieval and Photo Annotation tasks. *CLEF working notes*, 2009.

[22] J. Ngiam and H. Goh. I2R ImageCLEF Photo Annotation 2009 Working Notes. *CLEF working notes*, 2009.

[23] S. Nowak and P. Dunker. A Consumer Photo Tagging Ontology: Concepts and Annotations. In *THESEUS/ImageCLEF Pre-Workshop*, 2009.

[24] S. Nowak and P. Dunker. Overview of the CLEF 2009 Large-Scale Visual Concept Detection and Annotation Task. *CLEF working notes*, 2009.

[25] S. Nowak and H. Lukashevich. Multilabel Classification Evaluation using Ontology Information. In *Proc. of IRMLeS Workshop, ESWC*, 2009.

[26] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 1999.

[27] S. Sarin and W. Kameyama. Joint Contribution of Global and Local Features for Image Annotation. *CLEF working notes*, 2009.

[28] X. Shen, M. Boutell, J. Luo, and C. Brown. Multi-label machine learning and its application to semantic scene classification. In *Intern. Symp. on Electronic Imaging, San Jose, CA*, 2004.

[29] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. *Lecture Notes in Computer Science*, 4701:406, 2007.

[30] K. van de Sande, T. Gevers, and A. Smeulders. The University of Amsterdam's Concept Detection System at ImageCLEF 2009. *CLEF working notes*, 2009.

[31] Z.-Q. Zhao, H. Glotin, and E. Dumont. LSIS Scale Photo Annotations: Discriminant Features SVM versus Visual Dictionary based on Image Frequency. *CLEF working notes*, 2009.