BTU DBIS at ImageCLEF2013 Plant Identification Task

Sascha Saretz and Thomas Böttcher

Brandenburg Technical University
Chair of Database and Information Systems
Walther-Pauer-Str. 2, 03046 Cottbus
ssaretz@informatik.tu-cottbus.de, tboettcher@tu-cottbus.de

Abstract. In this paper we summarize the results of our second participation in the ImageCLEF2013 plant identification task. Again we used the combination of low-level features to identify similar pictures, identify adequate matchings and thus learn classifiers. This year, instead of using our workgroup's similarity query language "Commuting Quantum Query Language" (CQQL), we utilized support vector machines (SVMs) to classify the data. So we used classification on split subsets of the data instead of clustering similar results with the *k*-medoid method. For our experiments we used many different parameter combinations and feature combinations on the 2012 and 2013 data to compile four different runs.

Keywords: Plant identification, support vector machine, classification, feature combination, content-based image retrieval, experiments

1 Introduction

In this paper we present the participation of the Database and Information Systems Group (DBIS[1]) of the Brandenburg Technical University Cottbus, Germany to the ImageCLEF 2013 plant identification task[2] [1]. In the last years the DBIS working group was focused on experiments with low-level visual image features and their contribution towards a good retrieval performance. Furthermore our query language CQQL [11] was developed which allows a logical combination of various features. For more detailed information about CQQL we refer to the central CQQL publication [11]. Additional information, e.g., the relation of CQQL to fuzzy logic can be found in [12]. Its relation to probabilistic IR models is discussed in [18].

For the scope of this paper we take some distance to the CQQL part and try to gain some experience of classification techniques. We want to acquire some basic knowledge and in future combine the mechanisms of CQQL and classification techniques. The basis of our experiments in the plant identification task are still low-level visual features extracted from images.

[1] http://dbis.informatik.tu-cottbus.de
[2] http://www.imageclef.org/2013/plant

In our contribution we use our own developed retrieval system PythiaSearch [16,15,17] to extract a set of visual features from the Pl@ntLeaves[3] data set and create high dimensional feature vectors. Based on these feature vectors a multi class classification using support vector machines (SVM) is performed to predict the correct plant species.

For the task this year DBIS uses both, the last year's data and experiences as well as the training data from 2013. One goal was to find out whether an optimization on last year data (including training and test data) yields good results. This point is interesting because an analysis of the last year´s data reveals a big difference of classification results between training and test data.

Our main approach was to use the high dimensional classification technique SVM on high dimensional low-level features. Our submitted runs are based on several studies on different combinations of low-level features and various SVM kernels and parameters.

1.1 Plant Identification Task

The ImageCLEF Plant Identification Task 2013 is the third iteration of this task and is focused on plant species identification based on images. There are some changes compared to the challenge last year. The number of species has increased from 126 to 250, the number of pictures has drastically gone up from 11572 to 26077. New subclasses were introduced to offer more differentiation. The exact distribution can be found in the next subsection.

Furthermore the main goal changed from pure classification to a plant species retrieval task. The data and their identification are described in the following section. The complete description of the Plant Identification Task 2013 can be found in [7].

1.2 Training and Test Data

This year the Plant Identification Task is again based on the Pl@ntLeaves data set [10] which is divided into train and test data. The training subset was built by including the training and test subsets of last year Pl@ntLeaves data set and randomly selecting approximately 2/3 of the individual plant classes [9]. The complete data set contains 26077 pictures and the correspondent metadata files, 250 plant species from mainly French Mediterranean area, subdivided into the two main subclasses "SheetAsBackground" and "NaturalBackground". The distribution of training and test data of the Pl@ntLeaves data set is depicted in table 1.

These two classes are again divided into smaller subclasses. SheetAsBackground contains "Scan" (white paper as background) and "Scan-like" (laying on a table) pictures. "NaturalBackground" are pictures of different parts on the plants. It includes the five subclasses: "Entire", "Flower", "Fruit", "Leaf" and "Stem". The distribution of these subclasses is depicted in table 2.

[3] http://www.plantnet-project.org/

Table 1: Distribution of the Pl@ntLeaves data set

	SheetAsBackground	**Natural Background**	Σ
training data	9781	11204	20985
test data	1250	3842	5092
complete data set	11031	15046	26077

Table 2: Distribution of the Pl@ntLeaves class "Natural Background"

	Entire	**Flower**	**Fruit**	**Leaf**	**Stem**	**Natural Background**
training data	1455	3522	1387	3503	1337	11204
test data	694	1233	520	790	605	3842
complete data set	2149	4755	1907	4294	1942	15046

1.3 Identification and Evaluation

The goal of this task is to identify the plant species which is shown on each test image. So for each combination of test image and plant species a prediction should be made, where a prediction is a score value between 0 and 1. The prediction scores of the different species for each image have to be ranked in descending order.

To evaluate the predictions the task organizers calculate a score which is related to the rank of the correct species in the result list. Thereby a mean value is built per author and plant in the collection. Here, an author is a person which helped to built up the Pl@ntLeaves collection by contributing pictures and the corresponding metadata. The score values are calculated with the following formula [9]:

$$S = \frac{1}{U} \sum_{u=1}^{U} \frac{1}{P_u} \sum_{p=1}^{P_u} \frac{1}{N_{u,p}} \sum_{n=1}^{N_{u,p}} s_{u,p,n} \tag{1}$$

U : number of users (who have at least one image in the test data)
P_u : number of individual plants observed by the u-th user
$N_{u,p}$: number of pictures taken from the p-th plant observed by the u-th user
$s_{u,p,n}$: score between 1 and 0 equals to the inverse of the rank of the correct species (for the n-th picture taken from the p-th plant observed by the u-th user)

2 Strategy Overview

In this section we give a short introduction to the used techniques of our participation. Firstly, a short description to our retrieval system PythiaSearch is

given where the focus lies on the extraction of visual low-level features. Afterwards, the topic of image classification and the state of the art of support vector machiens is described.

2.1 PythiaSearch

Our retrieval system PythiaSearch is used to extract visual features from the given plant images. PythiaSearch is a multimedia information retrieval system supporting multiple search strategies. In order to formulate an information need (e.g. using Query by Example) the user can choose from multimodal data such as images, (multilingual) texts and various meta data formats. Additionally, it features a relevance feedback process that can be used to adjust query results based on the user's interaction with the system. A detailed description of the system can be found in [16,15,17]. A base system is freely available[4].

Feature Extraction

PythiaSearch supports the extraction of low-level global and local visual features, e.g., color, edge and texture features or local features like SIFT and SURF. In total, the extraction component offers more than 30 visual features.

Additionally, PythiaSearch allows the extraction of the most widely spread meta data collections EXIF, IPTC and XMP. These meta data, e.g. GPS coordinates, camera model, orientation, extend the variety of available features. Finally a full featured IR component is integrated.

For the scope of this paper we only extract available visual features. In a preparatory study (see section 3.1) we check the performance of all features and pick the best ones to create SVM models. The features were normalized to the interval $[0;1]$ so the SVMs can generate better results and the different dimensions have similar impact on the score values.

2.2 Support Vector Machines

Classification is an important field in data mining with an ever increasing relevance for real-world applications. Support vector machines (SVMs) are a popular and powerful type of classification algorithms which achieves excellent results. Furthermore this widespread technique can also classify large data sets rapidly. Currently major database vendors are integrating SVMs and other data mining capabilities in their systems, e.g. into Oracle Data Miner[5] [13] and SAP HANA[6] [6].

[4] http://saffron.informatik.tu-cottbus.de/iclef2013/personal_photo_baseline_sys.zip

[5] http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/

[6] http://www.saphana.com

The basic idea is to map the raw data into a high dimensional vector space and separate the data using a hyperplane. All data on one side of the hyperplane is added to class +1, the rest to class −1. There are also several approaches to generalize this behavior to multi-class classification [4,14,5,8].

For this paper we used the LIBSVM[7] library which is a popular collection of different SVM and regression models, e.g. radial basis functions (RBF), linear and polynomial SVMs [3]. It enables efficient multi-class classification, cross validation for model selection, weighted SVMs and probability estimates.

3 Description of Experiments

In this section we describe the selected features and how they were chosen. Afterwards, the type of SVMs and the final parameter combination are specified.

3.1 Feature Selection

As mentioned in our last year's participation in the plant identification task a good selection of visual features is crucial for a satisfying result. This year the challenge contains more complex and a much larger amount of data, thus reductions of the parameter combinations have to be performed. For our last year's evaluation it was sufficient to calculate a distance measure over the feature data. This year the SVM works directly on the raw feature data. Due to the fact that some features have a variable length it was impossible to use them directly as input for the SVMs. Another challenge was revealed in preliminary tests. We found out that feature combinations can lead to strongly varying results when changing the SVM kernels and parameters. This complicates obtaining a good model and increases the spanned space of feasible parameters combination enormously.

Nevertheless we used the last year's experiments [2] to preselect a set of possible features. We tested the performance of single features as well as of feature combinations. Additionally we tested the classification performance using the 2012 and 2013 data. For the 2012 data we checked the classification scores executing cross validation runs as well as training a model based on the official training data and evaluate the resulting models on the test data. For the 2013 data set we ran cross validations on the training data. Finally, several of the feature combinations were found to perform well (see table 3).

3.2 Run Overview

To construct the final runs we used SVM multi-class classification by applying the SVM library LIBSVM. A broad range of parameter combinations were tested and the best resulting runs selected. Parameters were amongst others the SVM type, kernel type, weights and costs for construction and movement of the hyperplane.

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 3: Composition of tested feature combinations

Feature	FC_00	FC_01	FC_02	FC_03	FC_04	FC_05	FC_07	FC_08	FC_09	FC_10	FC_11	FC_12	FC_13	FC_14
AutoColorCorrelogram				x	x									
BorderInteriorColor						x								
CEDD	x	x										x		x
ColorHistogram	x		x	x	x			x	x	x	x	x	x	x
ColorLayout			x	x	x		x		x	x	x	x	x	x
ColorStructure		x	x	x	x	x	x	x		x	x	x	x	x
EdgeHistogram	x		x				x	x	x		x	x	x	x
FCTH	x	x		x	x								x	x
Tamura	x		x		x	x	x	x	x	x		x	x	x
#dimensions	690	464	602	952	970	274	346	482	474	522	584	746	794	938

Table 4: Official run names

run	official run name
run1	1368038646069__DBISForMaT_run1_train2012_svm_Scan4_Photo2_1_2_3
run2	1368038721036__DBISForMaT_run2_train2012_svm_Scan12_Photo4_-_1_4
run3	1368045672892__DBISForMaT_run3_crossval2013_svm_feature4_config60_1_2_3
run4	1368045820175__DBISForMaT_run4_crossval2013_svm_feature5_config80_Photo14_1_3_3

The pictures were split into six subclasses, one for each of the five subcategories of "Natural Background" and the sixth for pictures in "SheetAsBackground". For each of these six classes different models were learned.

We assign the test images deterministically to the predicted classes. That means we choose the best class (plant species) and ranked it #1 with confidence score 1.0, the remaining classes (species) are not present in the result for that picture, meaning the confidence score is 0.

We published four runs for this task. The association between the internal and the official names of the runs is presented in table 4. The official names are important to find these runs on the task website [9] and in the lab proceedings [7]. The cryptic run names contain information about the feature selection (see section 3.1) and the used SVM parameters which are described in the next subsections.

All resulting runs are computed without probability estimates because in our tests they archived worse results than the deterministic runs and the computation time is partly prohibitively high. For all runs the shrinking heuristics supplied by LIBSVM are used because it is recommended by the LIBSVM's authors. This parameter seems to have no or a negligible effect on the result. The parameter `epsilon_loss_function` was not set because it is only needed in the SVM type `epsilon-SVR` which is a regression type algorithm. For the plant identification task we needed no regression but classification.

Table 5: LIBSVM parameter selection for `run1`

parameter	Natural Background	SheetAsBackground
feature combination	FC_02	FC_04
svm_type	nu-SVC (multi-class classification)	
kernel_type	RBF: $e^{-\gamma * \lvert u-v \rvert^2}$	
coef0	n/a	
γ	0.05	0.02
epsilon_termination_criterion	0.001	
degree	n/a	
cost	n/a	

Table 6: LIBSVM parameter selection for `run2`

parameter	Natural Background	SheetAsBackground
feature combination	FC_04	FC_12
svm_type	C-SVC	nu-SVC
kernel_type	polynomial: $\left(\gamma * u' * v + \text{coef0}\right)^{\text{degree}}$	
coef0	2	0
γ	0.01	
epsilon_termination_criterion	0.001	
degree	4	
cost	5	n/a

3.3 Runs 1 and 2

For our first two runs we used the score calculation script of the 2012 plant identification task using the 2012 training and test data to build classifiers. Our hope was that this script works analogously to its counterpart from the 2013 task. These approach holds two advantages: We learned with the ground truth of other data which can hopefully construct more general models reducing overfitting. Furthermore, using the score computation we can avoid problems which could arise when score computation and SVM classification results do not match completely. We chose the best performing parameter combinations for each subclass and applied the corresponding models to the 2013 test data. The parameters used for the first two runs are depicted in the tables 5 and 6.

3.4 Runs 3 and 4

In contrast to the first two runs `run3` and `run4` are constructed using the 2013 data set. This is done by performing 5-fold cross validations on the training data using many different parameter combinations. For this we computed the results and chose two promising feature combinations. The corresponding learned models are applied to the test data producing `run3` and `run4`. The parameters used for the last two runs are depicted in table 7.

Table 7: LIBSVM parameter selection for `run3` and `run4`

parameter	`run3`	`run4`
feature combination	FC_04	several[8]
svm_type	nu-SVC (multi-class classification)	
kernel_type	RBF: $e^{-\gamma * \lvert u-v \rvert^2}$	sigmoid: $\tanh(\gamma * u' * v + \text{coef0})$
coef0	n/a	several[9]
γ	0.01	
epsilon_termination_criterion	0.0001	
degree	n/a	
cost	n/a	

Table 8: Official score of submitted automatic visual runs

run	**SheetAsBG**	**Natural BG**	**Entire**	**Flower**	**Fruit**	**Leaf**	**Stem**
run1	0.191	0.12	0.067	0.168	0.1	0.052	0.103
run2	0.311	0.159	0.102	0.264	0.082	0.034	0.095
run3	0.193	0.158	0.109	0.256	0.079	0.035	0.095
run4	0.281	0.141	0.152	0.206	0.104	0.027	0.042
`max of runs`	0.607	0.393	0.297	0.494	0.311	0.275	0.285
`median`	0.191	0.089	0.068	0.105	0.081	0.049	0.095
`mean`	0.234	0.130	0.093	0.162	0.098	0.089	0.099

4 Results

In this section we compare the official scores of our four runs (see table 8) with the results of our competitors. For two of the five "Natural Background" subclasses we achieved good results. We were ranked the 4th best group for "Flower" (see figure 1) and 3th for "Entire" (see figure 2). Unfortunately, in the aggregated rank for "Natural Background" we were just 5th (see figure 3) because the results for the other three classes "Fruit" (5th), "Leaf" (7th) and "Stem" (6th) were just moderate. In the category of Scans and Scan-like photos of leafs (class "SheetAsBackground") we were ranked 6th over all runs and 5th when just considering the automatic runs (see figure 4).

From our competitors the groups Inria and NlabUTokyo were always ranked better than we were and Sabanci Okan in most categories. This could be motivated in their additional experience and knowledge in the plant identification task[10] or superior classification strategies. After release of the working notes

[8] feature combination: FC_05 for class "Natural Background", FC_14 for class "SheetAsBackground"

[9] coef0: -0.5 for class "Entire"; -0.2 for class "Fruit"; -1 for other classes

[10] Inria and Sabanci Okan also had excellent results in the 2011 and 2012 versions of the plant identification task.

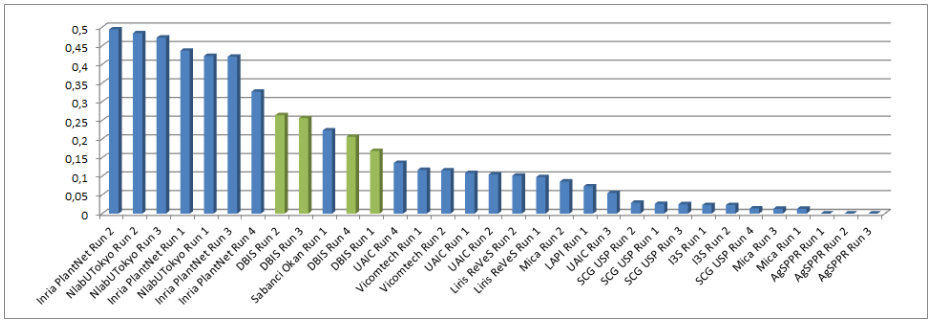

Fig. 1: Ranking scores for all runs considering class “Flower”

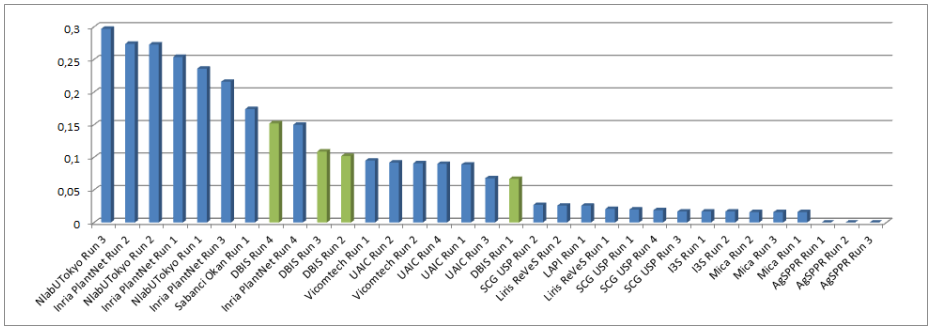

Fig. 2: Ranking scores for all runs considering class “Entire”

we will compare the different classification approaches for better results in the future.

We are pleased with the results of the class “Entire” because it is a difficult class, but also important for real-world applications. The poor results for “Leaf” (with natural background) and “SheetAsBackground” (scanned leafs) could be caused by the extensive research for this use case by other groups. Another reason could be the selection of wrong or not enough features for the different subclasses. Here, contour and shape features should promise better results.

It is interesting to observe that the results of the two different approaches for the run parametrization – training with the official score computation on the 2012 data and cross validation on the 2013 data – led to similar results. So the runs 1 and 3 have selected feature combination FC_04 for the class “SheetAsBackground” and the runs 2 and 3 selected FC_04 for the “Natural Background” subclasses. This means with the setup used by us both strategies can be used without losing too much classification success. Aggregated over all four runs `run2` was the best, `run1` the worst. This could mean that training with the last year’s ground truth has the potential to be better than with just the current training set, but the results and the small number of runs means that this result is not statistically significant and needs further research.

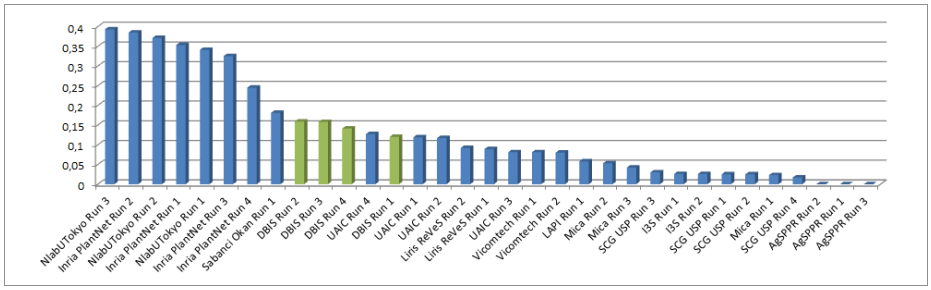

Fig. 3: Ranking scores for all runs considering class "Natural Background"

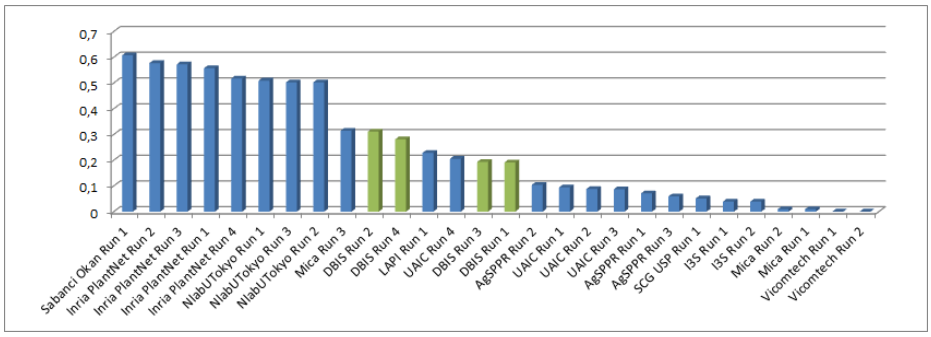

Fig. 4: Ranking scores for all fully automatic runs considering class "SheetAs-Background"

5 Conclusion and Future Work

Instead of using a weight-learning algorithm with CQQL we used SVMs to classify data this year. The above-average results show that this strategy is viable and further research in the parameter and feature selection should be considered.

In our 2012 participation [2] we suggested concentrating on each individual subclass and learn a separate classifier instead of just a general classifier applicable on all picture classes. This approach led to better result scores. Just one of the four submitted runs uses the same feature combination for all subclasses (`run3`), while this run did not obtain the best results.

Future Work

While performing the task and after evaluating the results many possibilities for future changes and enhancements of our approach emerged. In order to obtain a deeper understanding of the classification results we will rerun the tests using the 2013 version of the score computation when available. We are interested whether we will achieve better results with other SVM parameter combinations.

For this year's task we used deterministic SVMs because the probabilistic runs achieves a lower percentage of correct classifications on the 2013 data and lower score values on the 2012 data set. Because the 2013 data set is very different to the last year's version (see section 1) a probabilistic rerun with score computation of the 2013 makes sense. For both tests we need to wait for the ground truth and 2013 score computation script.

The usage of more and a larger spectrum of features could benefit the result. We could e.g. use good performing features described by other participants or more standard features. Furthermore instead of using several fixed feature combinations we could also test all reasonable feature combinations, train, test and choose the best one for each image subtype. Unfortunately, this procedure is very time-consuming and could also lead to overfitting of the classifiers.

For this task we restricted the usage of features to low-level visual features. We could improve our results by using also the metadata of the images, e.g. the GPS coordinates, time of shooting and possibly the photographer.

A further approach is the combination of SVM with the weighted learning approach of our 2012 contribution [2], which uses the downhill-Simplex method on our Commuting Quantum Query Lanage (CQQL). CQQL is a query language designed by our chair which can combine a broad range of similarity features based on the theoretical foundation of quantum logic.

In the next year we will try to generate an own CQQL kernel which can be easily used within LIBSVM. This kernel shall provide the capabilities of quantum logic an enable us to combine the different low-level features. We hope to gain more possibilities to adjust the training set and so separate the data in a more convenient way.

References

1. B. Caputo, H. Muller, B. Thomee, M. Villegas, R. Paredes, D. Zellhofer, H. Goeau, A. Joly, P. Bonnet, J. Martinez Gomez, I. Garcia Varea, M. Cazorla: ImageCLEF 2013: the vision, the data and the open challenges. In: Proc CLEF 2013. LNCS (2013)
2. Böttcher, T., Schmidt, C., Zellhöfer, D., Schmitt, I.: BTU DBIS' Plant Identification Runs at ImageCLEF 2012. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Duan, K.b., Keerthi, S.S.: Which is the best multiclass SVM method? An empirical study. In: Proceedings of the Sixth International Workshop on Multiple Classifier Systems. pp. 278–285 (2005)
5. Franc, V., Hlavac, V.: Multi-class Support Vector Machine. pp. 236–239 (2002)
6. Große, P., Lehner, W., Weichert, T., Färber, F., Li, W.S.: Bridging Two Worlds with RICE Integrating R into the SAP In-Memory Computing Engine. PVLDB 4(12), 1307–1317 (2011)
7. Hervé Goëau, Pierre Bonnet, Alexis Joly, Vera Bakić, Daniel Barthelemy, Nozha Boujemaa, Jean-François Molino: The ImageCLEF 2013 plant identification task, CLEF 2013 working notes, Valencia, Spain (2013)

8. Hsu, C.W., Lin, C.J.: A Comparison of Methods for Multiclass Support Vector Machines (2002)
9. ImageCLEF: Plant Identification Task 2013. http://imageclef.org/2013/plant, 2013-06-14.
10. Pl@ntNet: Interactive plant identification and collaborative information system. http://www.plantnet-project.org/papyrus.php?langue=en, 2013-06-14
11. Schmitt, I.: QQL: A DB&IR Query Language. The VLDB Journal 17(1), 39–56 (2008)
12. Schmitt, I., Zellhöfer, D., Nürnberger, A.: Towards quantum logic based multimedia retrieval. In: IEEE (ed.) Proceedings of the Fuzzy Information Processing Society (NAFIPS). pp. 1–6. IEEE (2008), 10.1109/NAFIPS.2008.4531329
13. Tamayo, P. et al.: Oracle Data Mining – Data Mining in the Database Environment. In: Maimon, O., Rokach, L. (eds.) The Data Mining and Knowledge Discovery Handbook, pp. 1315–1329. Springer (2005)
14. Weston, Jason an Watkins, C.: Multi-class Support Vector Machines (1998)
15. Zellhöfer, D.: A permeable expert search strategy approach to multimodal retrieval. In: Proceedings of the 4th Information Interaction in Context Symposium. pp. 62–71. IIIX '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2362724.2362739
16. Zellhöfer, D., Bertram, M., Böttcher, T., Schmidt, C., Tillmann, C., Schmitt, I.: PythiaSearch – A Multiple Search Strategy-supportive Multimedia Retrieval System. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. p. 59. ICMR '12, ACM (2012)
17. Zellhöfer, D., Böttcher, T., Bertram, M., Schmidt, C., Tillmann, C., Uhlig, M., Zierenberg, M., Schmitt, I.: PythiaSearch – Interaktives, Multimodales Multimedia-Retrieval. In: BTW. pp. 495–498 (2013)
18. Zellhöfer, D., Frommholz, I., Schmitt, I., Lalmas, M., van Rijsbergen, K.: Towards Quantum-Based DB+IR Processing Based on the Principle of Polyrepresentation. In: Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., Murdoch, V. (eds.) Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings, Lecture Notes in Computer Science, vol. 6611, pp. 729–732. Springer (2011)