Discovery of Environmental Web Resources Based on the Combination of Multimedia Evidence

Theodora Tsikrika Antonis Latas Anastasia Moumtzidou

Elisavet Chatzilari Stefanos Vrochidis Yiannis Kompatsiaris

Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
{theodora.tsikrika, latas, moumtzid, ehatzi, stefanos, ikom}@iti.gr

ABSTRACT

This work proposes a framework for the discovery of environmental Web resources providing air quality measurements and forecasts. Motivated by the frequent occurrence of heatmaps in such Web resources, it exploits multimedia evidence at different stages of the discovery process. Domain-specific queries generated using empirical information and machine learning driven query expansion are submitted both to the Web and Image search services of a general-purpose search engine. Post-retrieval filtering is performed by combining textual and visual (heatmap-related) evidence in a supervised machine learning framework. Our experimental results indicate improvements in the effectiveness when performing heatmap recognition based on SURF and SIFT descriptors using VLAD encoding and when combining multimedia evidence in the discovery process.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Algorithms, Performance, Design, Experimentation

Keywords

domain-specific search, keyword spices, heatmap recognition, environmental data, multimedia classification

1. INTRODUCTION

Environmental Web resources provide data (e.g., measurements and forecasts) and possibly additional information on environmental conditions, such as the weather, pollen concentration, and air quality. In particular, Web resources (e.g., sites, pages, and portals) that provide air quality measurements and forecasts are of great importance given the significant impact of air quality on life quality, due to its effect on health and on numerous outdoor activities. The need for access to reliable air quality data is thus of high value to several and diverse segments of society. This can be addressed by gathering and integrating data from air quality Web resources provided by different environmental agencies and organisations; this is a challenging task that requires the automatic discovery of such resources as its first step.

Such air quality Web resources typically report the concentration values of several air pollutants, such as sulphur dioxide, fine particles, and ozone, measured or forecact for specific regions and time periods [11]. Empirical studies [7] have revealed that air quality measurements, and particularly forecasts, are not only provided in textual form, but are also commonly encoded as multimedia, mainly in the form of *heatmaps*. Heatmaps are graphical representations of matrix data with a colour index scale illustrating pollutant concentrations over geographically bounded regions; Figure 1 provides an example. This observation motivates us to form the hypothesis that methods for the discovery of Web resources providing air quality measurements and forecasts would benefit by taking into account not only their textual content, but also their visual content, and especially evidence associated with the presence of heatmaps in them.

The automatic discovery of Web resources on any given topic is generally viewed as a *domain-specific search* problem [15] and is mainly addressed by two categories of techniques: (i) the domain-specific query submission to a general-purpose search engine, possibly followed by post-retrieval filtering

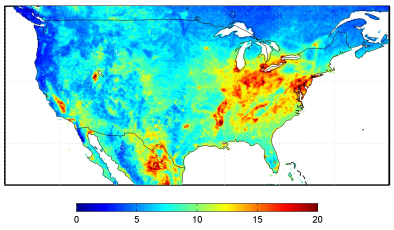

Figure 1: Heatmap example extracted from http://nasa.gov/ indicating satellite-derived measurements for fine particles ($PM_{2.5}$ [$\mu g/m^3$])

DOI: http://dx.doi.org/10.1145/2764873.2764876.

[22, 20, 14], and (ii) focussed crawling [17, 24]. In the environmental domain, resource discovery has previously been addressed through the application of techniques from both categories. Early techniques of the first category (e.g., [19]) have relied only on textual evidence for the classification of the retrieval results, while more recent approaches [18] have used visual evidence for performing this post-retrieval filtering; however, the combination of multimedia evidence has not been considered. On the other hand, recent focussed crawling approaches [25] have taken into account both textual and visual evidence for selecting the links to follow during their traversal of the Web graph.

This work proposes a framework for the discovery of Web resources that provide air quality measurements and forecasts based on techniques of the first category that combine multimedia evidence at several steps of the process. In particular, given the frequent presence of multimedia items relevant to the topic within such resources, namely heatmaps, it proposes the submission of domain-specific queries not only to the Web search component of general-purpose search engines (as done so far [22, 20, 19, 18]), but also to their Image search component, and the fusion of the two result sets. These submitted queries are either formulated manually based on domain-specific terms that are empirically identified, or are also automatically expanded by applying machine learning techniques for extracting domain-specific expressions (referred to as 'keyword spices' [22]) from positive and negative samples of such Web resources. By considering the results of image search to correspond to the Web resources containing the retrieved images, the Web search results are complemented with resources that are focussed on the domain, not only in terms of their textual content, but also with respect to their visual content. Moreover, the proposed framework combines multimedia evidence for performing the post-retrieval classification; this is achieved by the late fusion of textual and visual classification confidence scores obtained by supervised machine learning methods. The visual classification, in particular, considers the presence of heatmaps within these resources, and, to this end, several state-of-the-art descriptors (SURF, SIFT, and features extracted from Convolutional Neural Networks), are investigated using classifiers based on Support Vector Machines (SVMs) and Logistic Regression.

The main contributions of this work are thus: (i) a discovery framework that introduces the combination of multimedia evidence during search engine querying and post-retrieval filtering, and (ii) heatmap recognition based on several state-of-the-art descriptors that have not been previously investigated in this context.

2. RELATED WORK

Previous discovery approaches that submit domain-specific queries to general-purpose search engines have typically generated such queries using empirical information and then expanded them using, for instance, the concepts in an ontology [14] or machine learning driven techniques for extracting 'keyword spices', i.e., Boolean expressions of domain-specific terms corresponding to the output of a decision tree trained on an appropriate manually [22] or semi-automatically [20] annotated corpus. Our approach also follows this domain-specific search paradigm, but extends existing approaches by querying both the Web and Image search components of general-purpose search engines; it also combines textual and visual evidence, in the form of heatmaps, for further classifying the discovered resources at a post-retrieval stage.

Moreover, heatmap recognition has not been extensively researched. Relevant work in the related area of map recognition includes the use of knowledge of the colourisation schemes in maps for automatically segmenting them based on their semantic contents (e.g., roads) [9], and the development of techniques for improving segmentation quality of text and graphics in colour maps through the cleaning up of possible errors (e.g., dashed lines) [2]. Map recognition has also been investigated at TRECVID (`http://trecvid.nist.gov/`) through the evaluation of the concept 'maps' in the high level concept feature extraction task of TRECVID 2007 [21], where the best performing system employed a supervised machine learning method and fused visual descriptors [26]. Regarding heatmaps, research has mainly focussed on the information extraction from them [7]. More recently, a method for heatmap recognition that uses SVMs to build classifiers based on several visual features (MPEG-7, SIFT, AHDH) has been investigated [18]. Our approach considers further state-of-the-art visual features (including SURF and SIFT descriptors using VLAD encoding, as well as features extracted from Convolutional Neural Networks) and employs both SVMs and also Logistic Regression Classifiers.

3. DISCOVERY FRAMEWORK

This work proposes a framework for the discovery of environmental Web resources providing air quality measurements and forecasts that exploits multimedia evidence both when querying general-purpose search engines and also for the post-retrieval classification of the search results. Motivated by the frequent occurrence of heatmaps in such Web resources, it submits domain-specific queries (either 'basic' manually formulated queries or their automatic expanded versions) both to the Web search and also to the Image search components of a general-purpose search engine, and merges the retrieval results into a single set; for Image search, we consider that the retrieval results correspond to the Web resources containing the retrieved images, rather than to the actual images. This set of discovered resources is further filtered to reduce the noise by removing non-relevant items; to this end, both textual and visual evidence are considered.

After duplicate elimination, each resource is parsed to obtain its textual content and its embedded images (if any). The textual feature vector extracted from its content is used as input to an appropriately trained text classifier for obtaining an estimate of its relevance to the topic; this real-valued classification confidence score is denoted as $score_T$. On the other hand, the visual feature vectors extracted from the set of images downloaded from the resource are used as input to a heatmap classifier. The confidence scores for the individual images are then fused into a single score $score_V$ that reflects the overall presence of image(s) relevant to the topic within this resource. This work estimates the visual relevance of each resource to be equal to that of its most relevant image. As a result, detection of even a single heatmap ensures high estimates of relevance. The overall relevance of each discovered resource is then estimated by the late fusion of these two scores $f(score_T, score_V)$; both scores are converted to probabilistic outputs and thus the threshold for classifying a resource as positive is set to 0.5.

An overview of the discovery framework is depicted in Figure 2 and its main modules are described next.

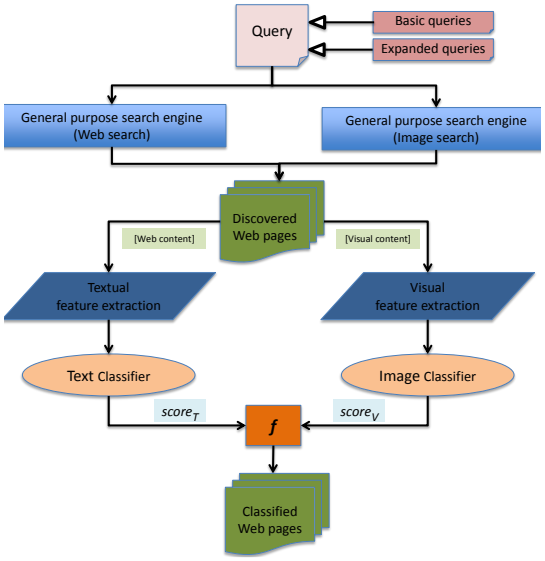

Figure 2: Discovery framework

3.1 Search Engine Querying

The first step comprises the formulation of two types of domain-specific queries, the basic and the expanded, and their submission to the Web search and to the Image search components of a general-purpose search engine.

3.1.1 *Basic Queries*

The formulation of basic queries is based upon the knowledge of domain experts or/and the findings of empirical studies on the particular domain. Examples of such queries for the air quality domain are listed in Table 1; these are based on the findings of empirical studies [18] and they are the ones that will be employed in our experiments (Section 4).

3.1.2 *Expanded Queries*

To automatically generate high precision and high recall queries in a domain of interest, such basic queries are expanded using the 'keyword spice' approach [22]. Based on machine learning techniques and a given set of Web resources annotated as relevant or non-relevant to the domain, this approach generates Boolean expressions (referred to as 'keyword spices') that aim to characterise in an effective manner the domain. To this end, the following process is applied.

First, the annotated Web resources are split into two disjoint subsets: (i) the training set for generating the initial keyword spices, and (ii) the validation set for simplifying them. Then, the nouns found in these Web resources are extracted so as to be used as domain-specific keywords. Based on the training set, a (binary) decision tree is constructed using the information gain measure without any pruning techniques and a decision tree learning algorithm is applied for discovering the keyword spices. The internal nodes of the decision tree correspond to the extracted keywords and its leaves to class labels. This results in a decision tree that can be expressed as a set of rules or as a Boolean disjunctive normal form; these are the initial keyword spices. Table 2 contains an example of such keyword spices for the air quality domain as generated in our experiments (see Section 4).

Similar to rule post-pruning, these initial keyword spices are simplified by iteratively removing keywords (or entire conjunctions) if their removal increases the F-measure (i.e., the harmonic mean of precision and recall) over the validation set, i.e., if their removal improves the effectiveness compared to them occurring in the query. This process is repeated until there is no keyword (or conjunction) that can be removed without decreasing the F-measure. Table 2 also contains the simplification of the aforementioned example of initial keyword spices.

Table 1: Air quality basic queries

q_1 = air quality	q_6 = ozone
q_2 = air pollutant	q_7 = PM2.5 particles
q_3 = carbon monoxide	q_8 = particulate matter
q_4 = nitrogen dioxide	q_9 = sulphur dioxide
q_5 = breathable particulate matter	q_{10} = fine particles

3.2 Post-Retrieval Classification

The resources discovered through search engine querying are then classified by combining their multimedia evidence.

3.2.1 *Text-Based Classification*

A text-based classifier is trained on a set of Web resources annotated as relevant or non-relevant to the domain. Each resource is parsed, its textual content is extracted, stopword removal and stemming are applied, and its textual feature vector is generated using the tf.idf term weighting scheme. Then, an SVM classifier is built using an RBF kernel, while 10-fold cross-validation is performed for selecting the class weight parameters. The text-based classifier is implemented using the libraries of the Weka machine learning software (`http://www.cs.waikato.ac.nz/ml/weka/`).

3.2.2 *Heatmap Recognition*

Heatmap recognition is performed within a binary supervised classification framework, where classifiers are trained using various visual features. This work employs two approaches for the visual representation of images: *shallow* and *deep*. For the shallow representation, the popular and widely adopted SIFT [13] and SURF [1] descriptors are used. There are several options for encoding these local features into a single feature vector, such as BoW, VLAD, LLC, and Fisher. VLAD [10] was preferred over the other approaches, as the best compromise between speed and performance. More specifically, compared to BoW, which can be faster, VLAD is significantly more effective, without adding significant computational load. On the other hand, the popular Fisher based encoding, which performs significantly better than VLAD, is much slower to compute, while when reducing its high dimensions with PCA-alike methods, its performance decreases significantly. This, however, is not the case for VLAD, which maintains its performance even with significantly reduced dimensions; see discussion in [16] on differences among encoding methods. For the deep representations, features are extracted using Convolutional Neural Networks (CNNs), which have shown remarkable performance the past years in several computer vision problems, including image annotation and object detection, while only requiring minimal computation time [12, 5].

Regarding the first case, each image is represented using six types of local descriptors: SIFT and its colour variants (i.e., RGB-SIFT and opponent-SIFT), and SURF and its colour variants (i.e., RGB-SURF and opponent-SURF); these were selected given their effectiveness in recent research [16]. Then, VLAD feature encoding is performed.

Table 2: Output of the keyword spices approach for the air quality domain (ˆsignifies the NOT operator)

Initial	ˆpollutant OR (forecast AND ˆengine AND ˆscience AND ˆauthority AND map AND ˆpark AND ˆstate) OR (forecast AND ˆengine AND ˆscience AND ˆauthority AND map AND ˆpark AND ˆtool) OR (forecast AND ˆengine AND city)
Simplified	forecast AND ˆengine AND ˆscience AND ˆauthority AND map AND ˆpark AND ˆstate

However, the VLAD feature encodings are high-dimensional and their dimensionality is affected by the dimensionality of the local descriptors they encode. Therefore, in order to compact the image representation, PCA, a common dimensionality reduction technique, is applied prior to feature encoding. Thus, after feature extraction, the local descriptors are compacted using PCA and subsequently aggregated using the VLAD encoding. Then, each image is divided into eight regions using spatial binning and sum pooling is applied for combining the features from the different regions. The result of the above process corresponds to a VLAD vector of 163,840 elements for SIFT or SURF. Eventually, the VLAD vectors are compressed into 4,000-element vectors by applying a modification of the random projection matrix. The aforementioned approach is described and evaluated thoroughly in [16]. These reduced VLAD vectors serve as input to the classification step which involves the use of Logistic Regression Classifiers. Therefore, six different classifiers are trained, one for each local descriptor, and their output is combined by applying late fusion techniques and specifically by averaging the classifier output scores; this approach is denoted as SSV based on the initials of the underlying descriptors, i.e., SIFT and SURF using VLAD encoding.

The second approach involves the use of the low dimensional CNN-based features. Initially, in order to train and find the parameters of the CNNs, a large-scale manually labeled dataset is employed. In our case, the pre-trained networks of [5] are used, where the ImageNet database was used to find the parameters [6]. The last layer is then removed and the vector of activities of the penultimate layer is used as the feature vector. We have tested two different architectures for the CNNs that explore a different accuracy/speed trade-off: (i) the medium architecture (CNN-M), and (ii) the slow architecture (CNN-S) [6]. Moreover, the application of data augmentation is also tested. This involves the perturbation of an image (i.e., flipping and cropping it in our case) resulting in 10 different versions of the same image; the features of all the 10 perturbed images are then combined using sum pooling. The resulting feature vectors had in all cases 4,096 dimensions. Finally, Linear SVM models are trained for both feature spaces using the LIBSVM library [3].

The evaluation was performed on a publicly available image collection[1] of manually annotated heatmaps that consists of two datasets [18]: (i) dataset A consisting of 2,200 images (600 relevant, i.e., heatmaps, and 1,600 irrelevant), and (ii) dataset B consisting of 2,860 images (1,170 heatmaps). To assess the discrimination power of the described low-level features, several experiments were carried out using dataset A for training and dataset B for testing.

Table 3 presents the results of these experiments. While all approaches achieve high precision, only the SSV approach manages to achieve high recall, and thus also high F-measure, as well as accuracy and Average Precision (AP) values. Moreover, this approach also outperforms the best performing descriptors on the same datasets [18], and in particular the early fusion of the MPEG-7 [4] descriptors Scalable Color (SC), Edge Histogram (EH), and Homogenous Texture (HT), as well as the EH feature on its own; see [18] for further details. It is also interesting to note the low performance of the CNN-based features, which have shown remarkable performance on generic object recognition tasks compared to shallow representations. This can be attributed to the fact that the networks have been pre-trained on generic images (ImageNet) which are not expected to include heatmaps in the training set. In order for CNN-based features to perform better, the networks would have to be refined with large sets of labelled heatmap images, which however are not easy to gather. Given these results, this work will apply the approach that employs the SIFT and SURF descriptors using VLAD encoding for identifying the heatmaps among the images downloaded from the discovered Web pages.

[1]Available at: `http://mklab.iti.gr/project/heatmaps`.

Table 3: Effectiveness of heatmap recognition

Descriptors	Precision	Recall	F-measure	Accuracy	AP
SSV	**99.83**	**97.95**	**98.88**	**99.09**	**97.94**
CNN-M	97.52	67.18	79.55	85.88	86.99
CNN-S	99.51	52.22	68.50	80.36	87.25
CNN-M augm.	99.50	51.11	67.53	79.90	85.56
CNN-S augm.	99.18	51.62	67.90	80.04	85.12

3.2.3 *Multimedia Combination*

The combination is performed based on the late fusion of the textual and visual classification confidence scores; as discussed, the visual score of a Web resource corresponds to the maximum of the scores of its images. To this end, the following combination methods [8], shown to be robust in various settings, are applied: *CombMIN*, *CombMAX*, and *CombSUM* denoting the minimum, maximum, and summation of all scores, as well as *CombMNZ* that denotes the sum of all scores multiplied by the number of nonzero scores.

4. EVALUATION

An experimental study is performed for evaluating the performance of the proposed approach.

4.1 Experiments

Our experiments employ the 10 air quality queries listed in Table 1 as the set of basic queries; this query set is denoted as Q1. Given that our main aim is to discover Web resources providing air quality measurements and forecasts, we expand these basic queries towards this direction by adding to each: (i) the term "measurements", or (ii) the term "forecasts", or (iii) the (simplified) keyword spice listed in Table 2; these query sets are denoted Q2, Q3, and Q4, respectively.

The keyword spices listed in Table 2 are generated using an annotated set of 664 air quality Web resources (284 positive, 380 negative). These were obtained by performing focussed crawling while starting from a set of seed pages

that provide air quality measurements and forecasts [25], and manually annotating the crawled results using the relevance scale presented in the next section. This annotated dataset is split in half for training and validation. Keywords are extracted from the positive samples (following the strict interpretation of relevance - see Section 4.2) after performing stopword removal, but not stemming, while nouns are identified through the use of the Part-Of-Speech tagger of the Stanford NLP group (`http://nlp.stanford.edu/software/tagger.shtml`); only the 160 most frequent nouns are kept. The C4.5 decision tree learning algorithm [23], and in particular its J48 implementation in Weka, is applied for discovering the initial keyword spices which are then simplified as described in Section 3.1.2.

Each query is submitted both to the Web search and to the Image search components of the Yahoo! Search BOSS API (`https://boss.yahoo.com/`) and the top 20 results are retrieved in each case. Therefore, for each query set, a maximum of 200 results are retrieved by each search component, and a maximum of 400 resources are discovered following the merge of the retrieval results from the two components.

A text-based classifier is trained based on an annotated dataset of 702 samples (194 positive, 508 negative) obtained through focussed crawling, similarly to above. Post-retrieval filtering is performed by combining the text-based classification scores with the visual classification scores (Section 3.2.3).

4.2 Relevance Assessments & Metrics

The discovered Web pages are manually assessed using the following four-point relevance scale:

- *highly relevant* (relevance scale = 3): Web resources that provide air quality measurements and forecasts.
- *partially relevant* (relevance scale = 2): Web resources that are about air quality measurements and forecasts, but do not provide actual data. Examples include Web resources that list monitoring sites and the pollutants being measured, explain what such measurements mean etc.
- *weakly* (or *softly*) *relevant* (relevance scale = 1): Web resources that are about air quality in general, discussing, for instance, the causes and effects of air pollution, but are not about air quality measurements and forecasts.
- *non-relevant* (relevance scale = 0): Web resources that are not about air quality.

Based on these multiple grade relevance assessments, the Normalized Discounted Cumulative Gain (NDCG) evaluation metric is applied. To apply binary metrics (e.g., MAP), these multiple grade assessments are mapped into binary relevance judgements in three different ways, depending on (i) whether we are strictly interested in discovering resources containing air quality data, (ii) whether we would also be interested in information about air quality measurements and forecasts, or (iii) whether we would also like information about air quality in general. These three mappings are:

- *strict*: when considering only highly relevant Web resources as relevant and the rest (partially relevant, weakly relevant, and non-relevant) as non-relevant,
- *lenient*: when considering both highly relevant and partially relevant Web resources as relevant and the rest (weakly relevant and non-relevant) as non-relevant, and
- *soft*: when considering all Web resources even with some slight degree of relevance (i.e., highly relevant, partially relevant, and weakly relevant) as relevant.

Next, the results of our experiments are presented.

5. RESULTS

For each search component, Table 4 presents the binary and multiple-level evaluation metrics averaged over the 10 queries for each query set Qi. First of all, the MAP and the NDCG that are based on the binary relevance assessments are always very high (over 90%) when considering the soft interpretation of relevance for all query sets. This indicates that general-purpose Web or Image search manages to discover within the top 20 results Web resources that are, at least generally speaking, about air quality; this is also indicated by the high values of the non-binary NDCG. To discover, though, resources providing air quality measurements and forecasts (i.e., to adopt the strict interpretation of relevance), query sets Q1 and Q2 are rather ineffective, while Q3 and Q4 perform significantly better. Although this was expected for Q1 since its queries are not geared towards the disovery of such resources, it is rather surprising for Q2 since its queries contain the term "measurements". This indicates that this is not a discriminative term and it is highly likely that it does not appear within such resources. On the other hand, the term "forecasts" that appears both in the Q3 queries and also in the keyword spice (see Table 2) used in Q4 for expanding Q1 queries appears to be highly beneficial.

Table 4: Average effectiveness for each Qi query set

	Total	MAP			NDCG (binary)			NDCG
		strict	lenient	soft	strict	lenient	soft	
	Web search							
Q1	200	0.26	0.27	0.97	0.38	0.44	0.98	0.88
Q2	200	0.20	0.60	0.98	0.28	0.81	0.98	0.90
Q3	200	0.47	0.73	**0.99**	0.65	0.85	**0.99**	0.89
Q4	190	**0.58**	**0.90**	**0.99**	**0.71**	**0.95**	**0.99**	**0.95**
	Image search							
Q1	200	0.05	0.14	0.95	0.09	0.27	**0.99**	0.90
Q2	200	0.16	0.54	0.92	0.28	0.73	0.96	0.85
Q3	200	0.30	0.68	**0.98**	0.48	0.83	0.98	0.87
Q4	177	**0.63**	**0.79**	**0.98**	**0.77**	**0.88**	**0.99**	**0.94**

A comparison between Web and Image search indicates that the former is more effective across all query sets and relevance interpretations, apart from the strict case for query set Q4. Given that Web resources that provide air quality measurements and forecasts are likely to encode data and information in visual form (typically heatmaps), this indicates that such automatically expanded queries are able to gear retrieval towards such resources when targeting the ones with visual content (i.e., the ones considered by Image search).

Next, the resources discovered by each query set Qi are merged into a single set and duplicates are removed. Table 5 presents the number of resources across the four relevance scales, as well as the precision with respect to the different interpetations of relevance. The overlap between Web and Image search results is not particularly high, ranging from 8% for Q1, to 17% for Q2 and Q3, and reaching 23% for Q4; therefore, these can be considered as complementary searches. Similarly to before, Q4 is the most effective query set, especially for the discovery of Web resources providing or describing measurements and forecasts (i.e., the strict and lenient cases).

Table 6 presents the results of the classification of the set of resources discovered by Q4 when submitted both to Web and Image search. Both for the strict and lenient cases, the classification (irrespective of the evidence being used, i.e., textual, visual, or their combination) improves significantly

Table 5: Web resources set discovered by each Qi

	Unique	Relevance				Precision		
		3	2	1	0	strict	lenient	soft
	Web+Image search							
Q1	369	13	31	301	24	0.04	0.12	0.93
Q2	336	19	124	181	12	0.06	0.43	0.96
Q3	332	105	87	134	6	0.32	0.58	**0.98**
Q4	285	129	74	68	14	**0.45**	**0.71**	0.95

Table 6: Multimedia classification for Q4

	Precision	Recall	F-measure	Accuracy	AP	NDCG
	strict					
textual	0.70	0.65	0.67	0.68	0.52	0.68
visual	0.81	0.35	0.49	0.62	0.32	0.47
CombMIN	**0.91**	0.26	0.41	0.61	0.24	0.38
CombMAX	0.69	0.74	**0.71**	**0.69**	0.62	0.75
CombSUM	0.66	0.76	**0.71**	0.68	0.65	0.79
CombMNZ	0.60	**0.87**	**0.71**	0.64	**0.72**	**0.86**
	lenient					
textual	0.76	0.76	0.76	0.67	0.61	0.78
visual	0.92	0.30	0.45	0.49	0.30	0.42
CombMIN	**0.95**	0.25	0.40	0.47	0.25	0.38
CombMAX	0.76	0.81	0.78	0.69	0.68	0.83
CombSUM	0.76	0.83	0.79	0.70	0.75	0.86
CombMNZ	0.73	**0.93**	**0.82**	**0.71**	**0.82**	**0.93**

over the corresponding search precision in Table 5. Moreover, the classification based on visual evidence is more precise than the text-based one, but has significantly lower recall. The combination of these multimedia evidence further improves the precision when using CombMIN, indicating the complementarity of the two approaches, while all other combination methods manage to improve all other metrics, thus reaching a balance between precision and recall.

6. CONCLUSIONS

This work proposed a framework for the discovery of Web resources providing air quality measurements and forecasts that combines multimedia (textual- and heatmap-based) evidence during search engine querying and post-retrieval filtering. It also investigated heatmap recognition using several state-of-the-art descriptors. Our experimental results indicate improvements in the effectiveness when performing heatmap recognition based on SURF and SIFT descriptors using VLAD encoding and when combining multimedia evidence in the discovery process. Future work includes the automatic identification of representative images from annotated resources (e.g., through clustering) and their use in the discovery process (e.g., as image queries submitted to the Image search service of a general-purpose search engine).

7. ACKNOWLEDGMENTS

This work was supported by MULTISENSOR (610411) and HOMER (312388) FP7 projects, partially funded by the European Commission.

8. REFERENCES

[1] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *Proc. of the 9th European Conference on Computer Vision (ECCV)*, pages 404–417. 2006.

[2] R. Cao and C. Tan. Text/graphics separation in maps. In *Proc. of 4th IAPR International Workshop on Graphics Recognition (GREC)*, pages 167–177. 2002.

[3] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[4] S. F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695, 2001.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. of the British Machine Vision Conference (BMVC)*, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[7] V. Epitropou, K. Karatzas, and A. Bassoukos. A method for the inverse reconstruction of environmental data applicable at the chemical weather portal. In *Proc. of the GI-Forum Symposium and Exhibit on Applied Geoinformatics*, pages 58–68, 2010.

[8] E. Fox and J. Shaw. Combination of multiple searches. In *Proc. of TREC-2*, pages 243–252, 1994.

[9] T. C. Henderson and T. Linton. Raster map image analysis. In *Proc. of the 10th International Conference on Document Analysis and Recognition (ICDAR)*, pages 376–380, 2009.

[10] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311, 2010.

[11] K. Karatzas and N. Moussiopoulos. Urban air quality management and information systems in Europe: legal framework and information access. *Journal of Environmental Assessment Policy and Management*, 2(02):263–272, 2000.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 1097–1105. 2012.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] H. P. Luong, S. Gauch, and Q. Wang. Ontology-based focused crawling. In *Proc. of the International Conference on Information, Process, and Knowledge Management*, pages 123–128, 2009.

[15] M. Lupu, M. Salampasis, and A. Hanbury. Domain specific search. In *Professional Search in the Modern World*, pages 96–117. 2014.

[16] F. Markatopoulou, N. Pittaras, O. Papadopoulou, V. Mezaris, and I. Patras. A study on the use of a binary local descriptor and color extensions of local descriptors for video concept detection. In *Proc. of the 21st International Conference on Multimedia Modeling (MMM)*, pages 282–293, 2015.

[17] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

[18] A. Moumtzidou, S. Vrochidis, E. Chatzilari, and I. Kompatsiaris. Discovery of environmental nodes based on heatmap recognition. In *Proc. of the 20th IEEE International Conference on Image Processing (ICIP)*, 2013.

[19] A. Moumtzidou, S. Vrochidis, S. Tonelli, I. Kompatsiaris, and E. Pianta. Discovery of environmental nodes in the web. In *Proc. of the 5th International Retrieval Facility Conference (IRFC)*, volume 7356, pages 58–72, 2012.

[20] H. Nabeshima, R. Miyagawa, Y. Suzuki, and K. Iwanuma. Rapid synthesis of domain-specific web search engines based on semi-automatic training-example generation. In *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 769–772, 2006.

[21] P. Over, G. Awad, W. Kraaij, and A. F. Smeaton. TRECVID 2007– overview. In *TRECVID 2007 workshop participants notebook papers*, 2007.

[22] S. Oyama, T. Kokubo, and T. Ishida. Domain-specific web search with keyword spices. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):17–27, 2004.

[23] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[24] T. T. Tang, D. Hawking, N. Craswell, and K. Griffiths. Focused crawling for both topical relevance and quality of medical information. In *Proc. of the 14th ACM International Conference on Information and Knowledge Management*, pages 147–154, 2005.

[25] T. Tsikrika, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris. Focussed crawling of environmental web resources: A pilot study on the combination of multimedia evidence. In *Proc. of the 1st International Workshop on Environmental Multimedia Retrieval (EMR)*, pages 61–68, 2014.

[26] J. Yuan and etal. THU and ICRC at TRECVID 2007. In *TRECVID 2007 workshop participants notebook papers*, 2007.