Focussed Crawling of Environmental Web Resources: A Pilot Study on the Combination of Multimedia Evidence

Theodora Tsikrika Anastasia Moumtzidou

Stefanos Vrochidis Ioannis Kompatsiaris

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

ABSTRACT

This work investigates the use of focussed crawling techniques for the discovery of environmental multimedia Web resources that provide air quality measurements and forecasts. Focussed crawlers automatically navigate the hyperlinked structure of the Web and select the hyperlinks to follow by estimating their relevance to a given topic, based on evidence obtained from the already downloaded pages. Given that air quality measurements and particularly air quality forecasts are presented not only in textual form, but are most commonly encoded as multimedia, mainly in the form of heatmaps, we propose the combination of textual and visual evidence for predicting the benefit of fetching an unvisited Web resource. First, text classification is applied to select the relevant hyperlinks based on their anchor text, a surrounding text window, and URL terms. Further hyperlinks are selected by combining their text classification score with an image classification score that indicates the presence of heatmaps in their source page. A pilot evaluation indicates that the combination of textual and visual evidence results in improvements in the crawling precision over the use of textual features alone.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Design, Experimentation

Keywords

focussed crawling, environmental data, link context, image classification, heatmaps

1. INTRODUCTION

Environmental conditions, such as the weather, air quality, and pollen concentration, are considered as one of the factors with a strong impact on the quality of life, since they directly affect human health (e.g., allergies and asthma), a variety of human outdoor activities (ranging from agriculture to sports and travel planning), as well as major environmental issues (such as the greenhouse effect). In order to support both scientists in forecasting environmental phenomena and also people in everyday action planning, there is a need for services that provide access to information related to environmental conditions that is gathered from several sources, with a view to obtaining reliable data. Monitoring stations established by environmental organisations and agencies typically perform such measurements and make them available, most commonly, through Web resources, such as pages, sites, and portals. Assembling and integrating information from several such providers is a major challenge, which requires, as a first step, the automatic discovery of Web resources that contain environmental measurement data; this can be cast as a *domain–specific search* problem.

Domain–specific search is mainly addressed by techniques that fall into two categories: (i) the domain–specific query submission to a general–purpose search engine followed by post–retrieval filtering, and (ii) *focussed crawling*. Past research in the environmental domain (e.g., [12]) has mainly applied techniques from the first category, while the effectiveness of focussed crawlers for environmental Web resources has not been previouly investigated.

Focussed (or *topical*) crawlers exploit the graph structure of the Web for the discovery of resources about a given topic. Starting from one or more *seed* URLs on the topic, they download the Web pages addressed by them and mine their content so as to extract the hyperlinks contained therein and select the ones that would lead them to pages relevant to the topic. This process is iteratively repeated until a sufficient number of pages is fetched (i.e., downloaded). To predict the benefit of fetching an unvisited Web resource is a major challenge since crawlers need to estimate its relevance to the topic at hand based solely on evidence obtained from the already downloaded pages. To this end, state–of–the–art approaches (see [13] for a review) adopt classifier–guided crawling strategies based on supervised machine learning; the hyperlinks are classified based on their *local* context, such as their anchor text and the textual content surrounding them in the parent page from which they were extracted, as well on *global* evidence associated with the entire parent

In: S. Vrochidis, K. Karatzas, A. Karpinnen, A. Joly (eds.): Proceedings of the International Workshop on Environmental Multimedia Retrieval (EMR 2014), Glasgow, UK, April 1, 2014, published at http://ceur-ws.org

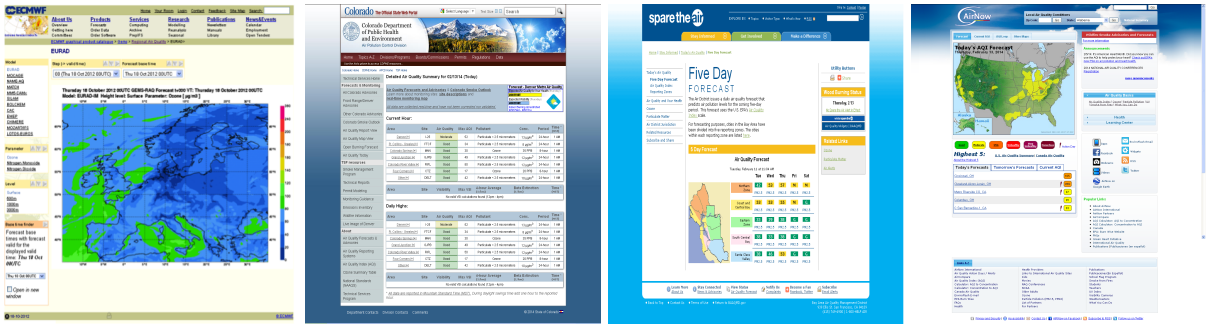

Figure 1: Examples of environmental Web resources providing air quality measurements and forecasts (left–to–right): http://gems.ecmwf.int/, http://www.colorado.gov/airquality/air_quality.aspx, http://www.sparetheair.org/Stay-Informed/Todays-Air-Quality/Five-Day-Forecast.aspx, http://airnow.gov.

page, such as its textual content or its hyperlink structure.

This work investigates focussed crawling for the automatic discovery of environmental Web resources, in particular those providing air quality measurements and forecasts; see Figure 1 for some characteristic examples. Such resources report the concentration values of several air pollutants, such as sulphur dioxide (SO2), nitrogen oxides and dioxide (NO+NO2), thoracic particles (PM10), fine particles (PM2.5) and ozone (O3), measured or forecast for specific regions [9]. Empirical studies [8, 17, 11] have revealed that such measurements and particularly air quality forecasts are presented not only in textual form, but are most commonly encoded as multimedia, mainly in the form of heatmaps (i.e., graphical representations of matrix data with colors representing pollutant concentrations over geographically bounded regions); see Figure 2 for an example.

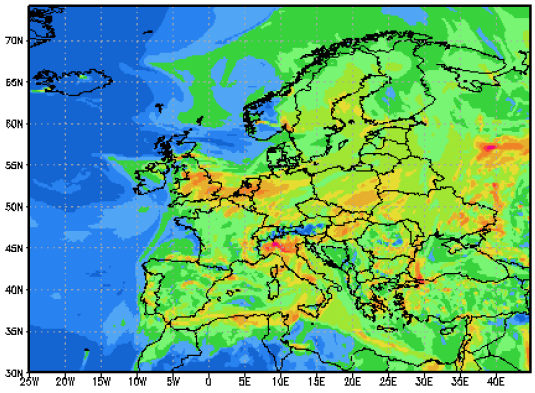

Figure 2: Heatmap example extracted from http://silam.fmi.fi/.

This motivates us to form the hypothesis that the presence of a heatmap in a page already estimated to be an air quality resource indicates that it is indeed highly relevant to the topic. Therefore, if such a page has already been downloaded by a crawler focussed on air quality, it would be a useful source of global evidence for the selections to be subsequently performed by such a focussed crawler. To this end, this work proposes a classifier–guided focussed crawling approach that estimates the relevance of a hyperlink to an unvisited Web resource based on the combination of (i) textual evidence from its local context and (ii) global visual evidence indicating the presence of a heatmap in its parent page. This is achieved by the late fusion of text and image classification confidence scores obtained by supervised machine learning methods based on Support Vector Machines (SVMs).

The main contribution of this work is a novel focussed crawling approach that takes into account multimedia (textual + visual) evidence for predicting the benefit of fetching an unvisited Web resource based on the combination of text and image classifiers. State–of–the–art classifier–guided focussed crawlers rely mainly on textual evidence [13] and, to the best of our knowledge, visual evidence has not been previously considered in this context. The proposed classifier–guided focussed crawler is evaluated in the domain of air quality environmental Web resources and the experimental results of our pilot study indicate improvements in the crawling precision when incorporating visual evidence, over the use of textual features alone.

The remainder of this paper is structured as follows. Section 2 discusses related work. Section 3 presents the proposed focussed crawling approach, Section 4 describes the evaluation setup, and Section 5 reports and analyses the experimental results. Section 6 concludes this work and outlines future research directions.

2. RELATED WORK

Focussed crawling techniques have been researched since the early days of the Web [7]. Based on the 'topical locality' observation that most Web pages link to other pages that are similar in content [6], focussed crawlers attempt to estimate the benefit of following a hyperlink extracted from an already downloaded page by mainly exploiting the (i) *local context* of the hyperlink and (ii) *global evidence* associated with its parent page.

Previous research has defined local context in textual terms as the lexical content that appears around a given hyperlink in its parent page. It may correspond to the anchor text of the hyperlink, a text window surrounding it, the words appearing in its URL, and combinations thereof. Virtually all focussed crawlers [7, 1, 20, 19, 15, 16, 13] use such textual evidence in one form or another. Global evidence, on the other hand, corresponds either to textual evidence, typically the lexical content of the parent page [16], or to hyperlink evidence, such as the centrality of the parent page within its neighbouring subgraph [1]. A systematic study of the

effectiveness of various definitions of link context has found that crawling techniques that exploit terms both in the immediate vicinity of a hyperlink, as well as in its entire parent page, perform significantly better than those depending on just one of those cues [16].

Earlier focussed crawlers (e.g., [5]) estimated the relevance of the hyperlinks pointing to unvisited pages by computing the textual similarity of the hyperlinks' local context to a query corresponding to a textual representation of the topic at hand; this relevance score could also be smoothed by the textual similarity of the parent page to the same query. State–of–the–art focussed crawlers, though, use supervised machine learning methods to decide whether a hyperlink is likely to lead to a Web page on the topic or not [13]. Classifier–guided focussed crawlers, introduced by Chakrabarti et al. [1], rely on models typically trained using the content of Web pages relevant to the topic; positive samples are usually obtained from existing topic directories such as the Open Directory Project[1] (ODP). A systematic evaluation on the relative merits of various classification schemes has shown that SVMs and Neural Network–based classifiers perform equally well in a focussed crawling application, with the former being more efficient, while Naive Bayes is a weak choice in this context [15]. This makes SVMs the classification scheme of choice in guiding focussed crawlers.

Focussed crawling has not really been previously explored in the environmental domain. The discovery of environmental Web resources has previously been addressed mainly through the submission of domain–specific queries to general–purpose search engines, followed by the application of a post–retrieval classification step for improving precision [12, 10]. The queries were generated using empirical information, including the incorporation of geographical terms [10], and were expanded using 'keyword spices' [14], i.e., a Boolean expression of domain–specific terms corresponding to the output of a decision tree trained on an appropriate corpus [12]. Post–retrieval classification was performed using SVMs trained on textual features extracted from a training corpus [12]. Such approaches are complementary to the discovery of Web resources using focussed classifiers and hybrid approaches that combine the two techniques in a common framework are a promising research direction [11].

3. MULTIMEDIA FOCUSSED CRAWLING

This work proposes a classifier–guided focussed crawling approach for the discovery of environmental Web resources providing air quality measurements and forecasts. To this end, it estimates the relevance of a hyperlink to an unvisited resource based on the combination of its local context with global evidence associated with its parent page. Local context refers to the textual content appearing in the vicinity of the hyperlink in the parent page. Motivated by the frequent occurrence of heatmaps in such Web resources, we consider the presence of a heatmap in a parent page as global evidence for its high relevance to the topic.

An overview of the proposed focussed crawling approach is depicted in Figure 3. First the seed pages are added to the list of URLs to fetch. In each iteration, a URL is picked from the list and the page corresponding to this URL is fetched (i.e., downloaded) and parsed to extract its hyperlinks. In the simple case that the focussed crawler estimates

[1] `http://www.dmoz.org/`.

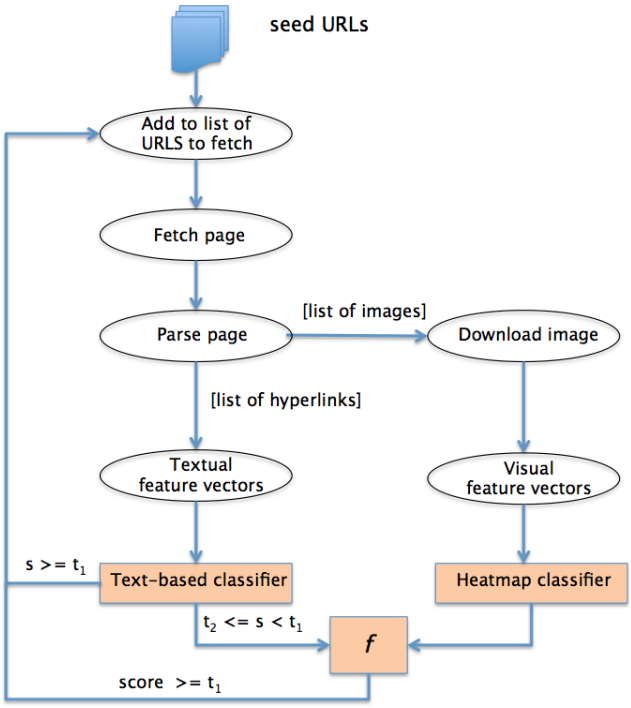

Figure 3: Multimedia focussed crawling.

the relevance of a hyperlink pointing to an unvisited page p based only on its local context, the decision to fetch p depends solely on the output of an appropriately trained text classifier. Therefore, a page is fetched if the confidence score s of the text–based classifier is above an experimentally set threshold t_1.

However, there are cases in which the local context is not sufficient to effectively represent relevant hyperlinks, leading them to obtain low confidence scores below the set threshold t_1, and thus to not being fetched by the focussed crawler. In this case, global evidence can be used for adjusting the estimate for the hyperlink's relevance. This is motivated by the 'topical locality' phenomenon of Web pages linking to other pages that are similar in content; therefore, if there is strong evidence of the parent's page relevance, then the relevance estimates of its children pages should be adjusted accordingly.

As mentioned before, the presence of heatmaps in a Web resource already assumed to be an air quality resource is a strong indication that it is indeed highly relevant to the topic. Therefore, we propose the consideration of heatmap presence in the parent page as global evidence to be used for adjusting the relevance estimate of hyperlinks with text-based confidence scores below the required threshold t_1 (in practice, a lower bound threshold t_2 is also set; this threshold is also experimentally tuned). In particular, the estimate of relevance of each hyperlink is adjusted to correspond to the late fusion of a text and a heatmap classifier: $score = f(text\_classifier, heatmap\_classifier)$, and the page is fetched if its $score \geq t_1$. In our case, a binary heatmap classifier is considered and the fusion function f is set to correspond to max. This results in a page being fetched if either its text-based confidence score is above t_1 or if its text-based confidence score is above t_2 ($t_2 < t_1$) and its

parent page contains at least one heatmap. Next, the text and heatmap classifiers employed in this work are described.

3.1 Text–Based Link Classification

Text–based link classification is performed using a supervised machine learning approach based on SVMs and a variety of textual features extracted from the hyperlink's local context. SVMs are applied due to their demonstrated effectiveness in similar applications [15].

Each hyperlink is represented using textual features extracted from the following fields:

- a: anchor text of the hyperlink,
- h: the terms extracted from the URL of the hyperlink; string sequences are split in punctuation marks and common URL extensions (e.g., `com`) and prefixes (e.g., `www`) are removed;
- s: the terms extracted from a text window of 50 characters surrounding the hyperlink; this text window does not contain the anchor text of adjacent links (i.e., the window stops as soon as it encounters another link),
- so: the terms extracted from a text window of 50 characters surrounding the hyperlink when overlap to the adjacent links is allowed.

Combinations of the above lead to the following five representations corresponding to concatenations of the respective fields: *a+s*, *a+so*, *a+h*, *a+h+s*, and *a+h+so*.

In the training phase, a list of positive and negative samples are collected first, so as to build a vocabulary for representing the samples in the textual feature space and also for training the model. Each sample corresponds to a hyperlink pointing to a Web page on air quality measurements and forecasts and its associated *a+so* representation. The vocabulary is built by accumulating all the terms from the *a+so* representations of the samples and eliminating all stopwords. This representation was selected so as to lead to a richer feature space, compared to the sparser *a*, *s*, and *a+s* representations, while also remaining relatively noise free compared to the *a+h+s* and *a+h+so* representations which are likely to contain more noise given the difficulties in successfully parsing URLs.

Each sample is represented in the textual feature space spanned by the created vocabulary using a $tf.idf = tf(t,d) \times log(\frac{n}{df(t)})$ weighting scheme, where $tf(t,d)$ is the frequency of term t in sample d and $idf(t)$ is the inverse document frequency of term t in the collection of n samples, where $df(t)$ is the number of samples containing that term. Furthermore, a feature representing the number of geographical terms in the sample's *a+so* representation is added, given the importance of such terms in the environmental domain [12]. To avoid overestimation of their effect, such geographical terms were previously removed from the vocabulary that was built. The SVM classifier is built using an RBF kernel and 5–fold cross–validation is performed on the training set to select the class weight parameters.

In the testing phase, each sample is represented as a feature vector based on the $tf.idf$ of the terms extracted from one of the proposed representation schemes (*a*, *a+s*, *a+so*, *a+h*, *a+h+s*, or *a+h+so*) and the number of geographical terms within the same representation. The text–based classification score of each hyperlink is then obatined by the employing the classifier on the feature vector and corresponds to a confidence value that reflects the distance of the testing sample to the hyperplane.

Our model was trained using 711 samples (100 positive, 611 negative). Each sample corresponds to a hyperlink pointing to page providing air quality measurements and forecasts; these hyperlinks were extracted from 26 pages about air quality obtained from ODP and previous empirical studies conducted by domain experts in the context of the project PESCaDO[2]. It should be noted that both the hyperlinks and their parent pages are different from the seed set used in the evaluation of the focussed crawler (see Section 4). The generated lexicon consists of 207 terms with the following being the 10 most frequent in the training corpus: *days*, *ozone*, *air*, *data*, *quality*, *today*, *forecast*, *yesterday*, *raw*, and *current*. The geographical lexicon consists of 3,625 terms obtained from a geographical database.

3.2 Heatmap Recognition

Heatmap recognition is performed by applying a recently developed approach by our research group [10]. That investigation on heatmap binary classification using SVMs and a variety of visual features indicated that, overall, the MPEG–7 [3] descriptors demonstrated a slightly better performance than the other tested visual features (SIFT [4] and AHDH[3] [18]).

In particular, the following three extracted MPEG–7 features that capture color and texture aspects of human perception were the most effective:

- **Scalable Color Descriptor (SC)**: a Haar–transform based encoding scheme that measures color distribution over an entire image, quantized uniformly to 256 bins,
- **Edge Histogram Descriptor (EH)**: a scale invariant visual texture descriptor that captures the spatial distribution of edges; it involves division of image into 16 non–overlapping blocks and edge information calculated for each block in five edge categories, and
- **Homogenous Texture Descriptor (HT)**: describing directionality, coarseness, and regularity of patterns in images based on a filter bank approach that employs scale and orientation sensitive filters.

Their early fusion (SC–EH–HT), as well as the feature EH on its own produced the best results when employing an SVM classifier with an RBF kernel. The evaluation was performed by training the classifier on a dataset of 2,200 images (600 relevant, i.e., heatmaps) and testing it on dataset of 2,860 images (1,170 heatmaps)[4].

In this work, both the EH and the SC–EH–HT models trained on the first dataset are employed. An image is classified as a heatmap if at least one of these classifiers considers it to be a heatmap, i.e., a late fusion approach based on a logical OR is applied.

[2]Personalised Environmental Service Configuration and Delivery Orchestration (`http://www.pescado-project.eu/`).
[3]Adaptive Hierarchical Density Histogram.
[4]Both datasets are available at: `http://mklab.iti.gr/project/heatmaps`.

Table 1: List of seed URLs.

	URL	heatmap present
1.	http://aircarecolorado.com/	
2.	http://airnow.gov/	✓
3.	http://db.eurad.uni-koeln.de/en/	✓
4.	http://gems.ecmwf.int/	✓
5.	http://maps.co.mecklenburg.nc.us/website/airquality/default.php	✓
6.	http://uk-air.defra.gov.uk/	
7.	http://www.baaqmd.gov/The-Air-District.aspx	
8.	http://www.eea.europa.eu/	
9.	http://www.gmes-atmosphere.eu/	
10.	http://www.londonair.org.uk/LondonAir/Default.aspx	✓

4. EVALUATION

A pilot study is performed for evaluating the performance of the proposed focussed crawling approach.

A set of 10 seeds[5] (listed in Table 1) was selected, similarly to before, i.e., using ODP and the outcomes of empirical studies conducted by domain experts in the context of the project PESCaDO. Half of them contain at least one heatmap. Starting from these 10 seeds, a crawl at depth 1 is performed. A total of 807 hyperlinks are extracted from these 10 seeds and several focussed crawling approaches are applied for deciding which ones to fetch. These are evaluated in the following two sets of experiments.

4.1 Experiments

Experiment 1: This experiment examines the relative merits of the different text–based representations of hyperlinks (i.e., *a*, *a+s*, *a+so*, *a+h*, *a+h+s*, and *a+h+so*). In this case, a text–based classifier–guided focussed crawling is applied for each representation and a page is fetched if its text–based confidence score is above a threshold t_1. Experiments are performed for t_1 values ranging from 0.0 to 0.9 at step 0.1. When $t_1 = 0.0$, the crawl corresponds to a breadth–first search where all hyperlinks are fetched and no focussed crawling is performed.

Experiment 2: This experiment investigates the effectiveness of incorporating multimedia evidence in the form of heatmaps in the crawling process. In this case, a page pointed by a hyperlink is fetched if the hyperlink's text–based confidence score is above t_1 or if its text–based confidence score is above t_2 ($t_2 < t_1$) and its parent page contains at least one heatmap. The text–based confidence scores are obtained from the best performing classifier in Experiment 1. Experiments are performed for t_1 and t_2 values ranging from 0.0 to 0.9 at step 0.1, while maintaining $t_2 < t_1$. These experimental results are compared against two baselines: (i) the results of the corresponding text–based focussed crawler for threshold t_1, and (ii) the results of the corresponding text–based focussed crawler for threshold t_2.

To determine the presence of a heatmap in the parent page of a hyperlink, the page is parsed (since it is already downloaded) and the hyperlinks pointing to images are compiled into a list. The crawler iteratively downloads each image in the list, extracts its visual features, and applies the heatmap classification until a heatmap is recognised or a maximum number of images is downloaded from each page (set to 20 in our experiments).

In both experiments, when a hyperlink appears more than once within a seed page, only the one with the highest score is taken into consideration for evaluation purposes.

[5]These URLs are different to the ones used when training the classifiers.

4.2 Performance Metrics

The standard retrieval evaluation metrics of *precision* and *recall* are typically applied for assessing the effectiveness of a focussed crawler. Precision corresponds to the proportion of fetched pages that are relevant and recall to the proportion of all relevant pages that are fetched. The latter requires knowledge of all relevant pages on a given topic, an impossible task in the context of the Web. To address this limitation, two recall–oriented evaluation techniques have been proposed [13]: (i) manually designate a few representative pages on the topic and measure what fraction of them are discovered by the crawler, and (ii) measure the overlap among independent crawls initiated from different seeds to see whether they converge on the same set of pages. Given the small scope of our study (i.e., a crawl at depth 1), these approaches are not applicable and therefore recall is not considered in our evaluation. In addition to precision, the *accuracy* of the classification of the crawled outlinks is also reported.

4.3 Relevance Assessments

All 807 extracted hyperlinks were manually assessed. After applying some light URL normalisation (e.g., deleting trailing slashes) and removing duplicates, 689 unique URLs remain. These correspond both to internal (within–site) and to external links that were assessed using the following three–point relevance scale:

- (*highly*) *relevant*: Web resources that provide air quality measurements and forecasts. These data should either be visible on the page or should appear after selecting a particular value from options (e.g., region, pollutant, time of day, etc.) provided from drop–down menus.
- *partially relevant*: 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, describe methods, approaches, and research for measuring, validating, and forecasting air quality data, or provide links to components, systems, and applications that measure air quality.
- *non–relevant*: Web resources that are not relevant to air quality measurements and forecasts, including resources that are about air quality and pollution in general, discussing, for instance, its causes and effects.

Overall, our crawled dataset contains 232 (33.7%) highly relevant pages, 51 (7.4%) partially relevant, and 406 (58.9%) non–relevant ones.

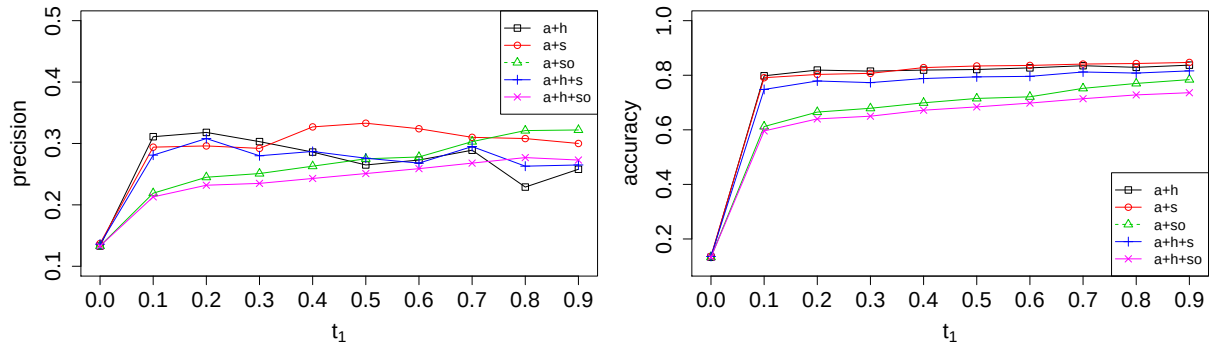

Figure 4: Precision and accuracy of the focussed crawl for each text–based link classification method (a+h, a+s, a+so, a+h+s, a+h+so) for threshold $t_1 \in \{0, 0.1, ..., 0.9\}$ when *strict* relevance assessments are employed.

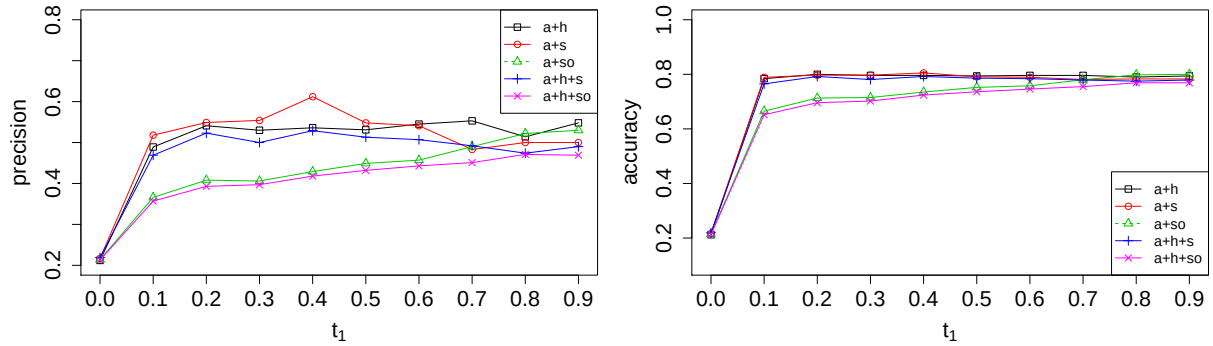

Figure 5: Precision and accuracy of the focussed crawl for each text–based link classification method (a+h, a+s, a+so, a+h+s, a+h+so) for threshold $t_1 \in \{0, 0.1, ..., 0.9\}$ when *lenient* relevance assessments are employed.

A closer inspection revealed that 162 (69.8%) of the highly relevant pages were all crawled from seed no. 2 in Table 1 (`http://airnow.gov/`). These correspond to internal links pointing to pages with air quality measurements/forecasts, each regarding a different U.S. region. This, in conjunction with the fact that all these links obtained really high scores (over 0.9) by our text classifier led us to remove them from further consideration as they would significantly skew the evaluation results. Therefore, the evaluation was performed only for the pages crawled from the nine remaining seeds and these are the results reported in Section 5[6]. Starting from the 9 seeds, our crawled dataset contains 526 URLs: 70 (13.3%) highly relevant pages, 50 (9.5%) partially relevant, and 406 (77.2%) non–relevant ones.

To apply the performance metrics presented above, these multiple grade relevance assessments are mapped into binary relevance judgements in two different ways, depending on whether we are strictly interested in discovering resources containing air quality data, or whether we would also be interested in information about air quality measurements and forecasts. In particular, two mappings are considered:

- *strict*: when considering only highly relevant Web resources as relevant and the rest (partially relevant and non–relevant) as non–relevant, and
- *lenient*: when considering both highly relevant and partially relevant Web resources as relevant.

The distributions of relevance assessments in these two cases are listed in Table 2.

Table 2: Relevance assessments distributions when the 3–point scale judgements are mapped to binary.

	Strict		Lenient	
Relevant	70	(13.3)%	120	(22.8)%
Non–Relevant	456	(86.7)%	406	(77.2)%
All	526	(100.0)%	526	(100.0)%

4.4 Implementation

Our implementation is based on Apache Nutch (`http://nutch.apache.org/`), a highly extensible and scalable open source Web crawler software project. To convert it to a focussed crawler, its parser was modified so as to filter the links being fetched based on our proposed approach. The text–based classifier was implemented using the libraries of the Weka machine learning software (`http://www.cs.waikato.ac.nz/ml/weka/`), while the implementation of the visual classifier was based on the LIBSVM [2] library.

5. RESULTS

Experiment 1: The results of this first experiment that evaluates the effectiveness of the different textual represen-

[6]It should be noted that `http://airnow.gov/` appears in the list of our crawled pages even when removed from the seed list, since it is linked from other seed pages. However, since crawling is performed at depth 1, its own outlinks are not considered any further.

Table 3: Precision of the focussed crawler that combines the a+s text–based link classifier with the heatmap classifier for thresholds $t_1 \in \{0.1, ..., 0.9\}$ and $t_2 \in \{0, 0.1, ..., 0.8\}$ when *strict* relevance assessments are employed.

t_1	t_2									Text–based baseline a+s (fetch if $s >= t_1$)
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
0.1	0.215									0.294
0.2	0.213	**0.314**								0.296
0.3	0.206	**0.299**	0.284							0.292
0.4	0.214	**0.346**	**0.340**	**0.354**						0.327
0.5	0.215	**0.353**	**0.347**	**0.362**	0.333					0.333
0.6	0.214	**0.362**	**0.356**	**0.372**	**0.341**	0.333				0.324
0.7	0.222	**0.405**	**0.400**	**0.421**	**0.385**	**0.382**	**0.364**			0.310
0.8	0.222	**0.405**	**0.400**	**0.421**	**0.385**	**0.382**	**0.364**	0.310		0.308
0.9	0.221	**0.421**	**0.417**	**0.441**	**0.400**	**0.400**	**0.379**	**0.320**	**0.318**	0.300
Text–based baseline a+s (fetch if $s >= t_2$)	0.137	0.294	0.296	0.292	0.327	0.333	0.324	0.310	0.308	

Table 4: Precision of the focussed crawler that combines the a+s text–based link classifier with the heatmap classifier for thresholds $t_1 \in \{0.1, ..., 0.9\}$ and $t_2 \in \{0, 0.1, ..., 0.8\}$ when *lenient* relevance assessments are employed.

t_1	t_2									Text–based baseline a+s (fetch if $s >= t_1$)
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
0.1	0.360									0.518
0.2	0.360	**0.571**								0.549
0.3	0.350	0.552	0.537							0.554
0.4	0.352	**0.615**	**0.620**	**0.646**						0.612
0.5	0.347	**0.608**	**0.612**	**0.638**	0.604					0.548
0.6	0.343	**0.617**	**0.622**	**0.651**	**0.614**	0.538				0.541
0.7	0.333	**0.619**	**0.625**	**0.658**	**0.615**	0.529	0.515			0.483
0.8	0.333	**0.619**	**0.625**	**0.658**	**0.615**	0.529	0.515	0.483		0.500
0.9	0.328	**0.632**	**0.639**	**0.676**	**0.629**	0.533	0.517	0.480	0.500	0.500
Text–based baseline a+s (fetch if $s >= t_2$)	0.217	0.518	0.549	0.554	0.612	0.548	0.541	0.483	0.500	

tations employed by the text–based focussed crawler are depicted in Figures 4 and 5, when applying strict and lenient relevance assessments, respectively.

The *a+s* classifier–guided focussed crawler achieves the highest overall precision, both for the strict and the lenient cases, and for $t_1 = 0.4$, indicating the benefits of combining the anchor text with the terms obtained from a non–overlapping text window. It also achieves the highest accuracy, which is equal to that of the *a+h* and *a+h+s* classifiers; these two classifiers have though slightly lower precision compared to that of *a+s*. This indicates that the URL is potentially a useful source of evidence and that application of more advanced techniques for extracting terms from an URL is probably required for reaching its full potential. The *a+so* and *a+h+so* classifiers are the least effective for lower t_1 values indicating that the additional terms present in the overlapping text window introduce noise that leads to the misclassification of non–relevant hyperlinks. Furthermore, all focussed crawlers improve upon precision for $t_1 = 0.0$ that corresponds to general–purpose crawling. As expected, the absolute values of precision are much higher in the lenient case, compared to the strict.

Experiment 2: The second experiment aims to allow us to gain insights into the feasibility and potential benefits of incorporating multimedia in the form of heatmaps in the crawling process. To this end, it combines *a+s*, the best performing text–based classifier from the first experiment, with results from the heatmap classifier. First, the results of the heatmap classification are presented.

Each of the nine seeds contains 15 images on average as identified by our parser. On average, 8 images are downloaded from each seed before a heatmap is found or the image list ends. Out of the 75 downloaded images, 74 were correctly classified, with 3 being heatmaps. This means that 8 of the 9 seeds were classified accurately for the presence of heatmaps in them (all apart from seed no. 10 in Table 1). This is probably due to the difficulty in parsing the specific Web resource and also in recognising its images as heatmaps, as they correspond to non–typical heatmaps, different to the ones in our training set. On average, 10 seconds were required per Web resource for the downloading, feature extraction, and classification of its images; however, this overhead could be reduced by applying parallelisation.

Tables 3 and 4 present the results of the second experiment, when applying strict and lenient relevance assessments, respectively, for t_1 and t_2 values ranging from 0.0 to 0.9 at step 0.1, while maintaining $t_2 < t_1$. The results are compared against the two baselines listed in the tables' last column and last row respectively. The values in bold correspond to improvements over both baselines.

The observed substantial improvements for multiple threshold values provide an indication of the benefits of incorporating visual evidence as global evidence in a focussed crawler. Consider the best performing classifier when strict relevance assessments are employed: it achieves precision of 0.44 for $t_1 = 0.9$ and $t_2 = 0.3$, while the text–based focussed crawler for the same $t_1 = 0.9$ achieves precision 0.30. An examination of the results shows that the improvements are due to the fact that 65% of the newly added hyperlinks, i.e., those with text–based classification score between 0.3 and 0.9, are relevant.

6. CONCLUSIONS

This work proposed a novel classifier–guided focussed crawling approach for the discovery of environmental Web resources providing air quality measurements and forecasts that combines multimedia (textual + visual) evidence for predicting the benefit of fetching an unvisited Web resource.

The results of our pilot study provide a first indication of the effectiveness of incorporating visual evidence in the focussed crawling process over the use of textual features alone.

Large–scale experiments are currently planned for fully assessing the potential benefits of the proposed multimedia focussed crawling approach, including experiments for improving the effectiveness of the textual classification by taking into account also the textual content of the entire parent page, similar to previous research [16]. Further future work includes the consideration of other types of images common in environmental Web resources, such as diagrams, simple filtering mechanisms for removing prior to classification small–size images that are unlikely to contain useful information (e.g., logos and layout elements), and the incorporation of additional local evidence, such as the distance of the hyperlink to the heatmap image. Finally, we aim to investigate the application of the proposed focussed crawler in other domains where information is commonly encoded in multimedia form, such as food recipes.

7. ACKNOWLEDGMENTS

This work was supported by MULTISENSOR (contract no. FP7–610411) and HOMER (contract no. FP7–312388) projects, partially funded by the European Commission.

8. REFERENCES

[1] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the 8th International Conference on World Wide Web*, (WWW 1999), pages 1623–1640, 1999.

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

[3] 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.

[4] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference (BMVC 2011)*, pages 1–12, 2011.

[5] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks*, 30(1-7):161–172, 1998.

[6] B. D. Davison. Topical locality in the web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR 2000), pages 272–279, 2000.

[7] P. De Bra and R. D. J. Post. Information retrieval in the world-wide web: Making client-based searching feasible. *Computer Networks and ISDN Systems*, 27(2):183–192, 1994.

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

[9] 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.

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

[11] A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris. Discovery, analysis and retrieval of multimodal environmental information. In *Encyclopedia of Information Science and Technology (in press)*. IGI Global, 2013.

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

[13] C. Olston and M. Najork. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.

[14] 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, Jan. 2004.

[15] G. Pant and P. Srinivasan. Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems*, 23(4):430–462, 2005.

[16] G. Pant and P. Srinivasan. Link contexts in classifier-guided topical crawlers. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):107–122, 2006.

[17] R. San José, A. Baklanov, R. Sokhi, K. Karatzas, and J. Pérez. Computational air quality modelling. *Developments in Integrated Environmental Assessment*, 3:247–267, 2008.

[18] P. Sidiropoulos, S. Vrochidis, and I. Kompatsiaris. Content-based binary image retrieval using the adaptive hierarchical density histogram. *Pattern Recognition*, 44(4):739 – 750, 2011.

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

[20] T. T. Tang, D. Hawking, N. Craswell, and R. S. Sankaranarayana. Focused crawling in depression portal search: A feasibility study. In *Proceedings of the 9th Australasian Document Computing Symposium (ADCS 2004)*, pages 1–9, 2004.