Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Focussed Crawling of Environmental Web Resources Based on the Combination of Multimedia Evidence

Theodora Tsikrika · Anastasia Moumtzidou · Stefanos Vrochidis · Ioannis Kompatsiaris

Received: date / Accepted: date

Abstract Focussed crawlers enable the automatic discovery of Web resources about a given topic by automatically navigating the Web link structure and selecting the hyperlinks to follow by estimating their relevance to the topic based on evidence obtained from the already downloaded pages. 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 textual evidence representing its local context, namely the textual content appearing in its vicinity in the parent page, with visual evidence associated with its global context, namely the presence of images relevant to the topic within the parent page. The proposed focussed crawling approach is applied towards the discovery of environmental Web resources that provide air quality measurements and forecasts, since such measurements (and particularly the forecasts) are not only provided in textual form, but are also commonly encoded as multimedia, mainly in the form of heatmaps. Our evaluation experiments indicate the effectiveness of incorporating visual evidence in the link selection process applied by the focussed crawler over the use of textual features alone, particularly in conjunction with hyperlink exploration strategies that allow for the discovery of highly relevant pages that lie behind apparently irrelevant ones.

Keywords focussed crawling · environmental data · link context · image classification · heatmaps

T. Tsikrika
Information Technologies Institute, CERTH, Thessaloniki, Greece
Tel.: +30 2311 257793 Fax: +30 2310 474128 E-mail: theodora.tsikrika@iti.gr

A. Moumtzidou · S. Vrocidis · I. Kompatsiaris
Information Technologies Institute, CERTH, Thessaloniki, Greece
E-mail: {moumtzid, stefanos, ikom}@iti.gr

1 Introduction

Focussed (or topical) *crawlers* enable the discovery of Web resources about a given topic by automatically navigating the Web graph structure and selecting to follow only the hyperlinks that would lead them to resources relevant to the topic. The process starts by adding a set of *seed* Web pages on the topic to the *frontier* [16], i.e., to the list that contains the URLs of the pages discovered, but not yet downloaded by the crawler. Each of the pages in the frontier is fetched (i.e., downloaded) so that the crawler can extract the hyperlinks it contains and select for navigation the most promising ones, i.e., those with the highest likelihood of pointing to other, yet unvisited pages relevant to the topic. The URLs of these selected pages are also added to the frontier and this process is iteratively repeated until a sufficient number of pages is fetched, or other criteria are satisfied, such as a limit on the crawling depth (i.e., the maximum distance allowed between the current and seed pages) is reached.

To predict the benefit of fetching an unvisited Web resource is a challenging task, since its relevance to the topic at hand needs to be estimated based solely on evidence obtained from the already downloaded pages. To this end, state-of-the-art approaches (see [16] for a review) adopt classifier-guided crawling strategies based on supervised machine learning methods that rely on two sources of evidence for selecting which hyperlinks to follow for reaching relevant Web resources: (i) the *local* context of hyperlinks, typically the textual content appearing in their vicinity within their parent page (e.g., their anchor text and part of its surrounding text), and (ii) the *global* context of hyperlinks, i.e., evidence typically associated with the entire parent page (e.g., its textual content or its hyperlink connectivity). The use of the latter is motivated by the regularity of topicality observed over the Web, whereby Web pages tend to link to other pages that are similar in content and thus likely to be relevant to the same topics [7, 21]. Therefore, given that strong evidence of relevance for the parent page increases the likelihood of relevance of its children pages, the global context of hyperlinks pointing to these childen pages could be used for adjusting the estimates of relevance determined by the local context.

Existing focussed crawling approaches have predominantly employed textual evidence for representing and estimating the relevance of the global context of hyperlinks [20, 16]. Motivated by the frequent (and continuously increasing) occurrence of multimedia items, such as images and videos, in present-day Web resources, either for complementing the available textual content with illustrations relevant to the topic (e.g., Figure 1(left)), or for encoding much of the information to be conveyed in a visual form (e.g., Figure 1(right)), this work proposes the consideration of such multimedia evidence in the representation and estimation of relevance of the global context in the link selection process of a focussed crawler. In particular, it proposes a classifier-guided focussed crawling approach that estimates the relevance of a hyperlink to an unvisited Web resource based on the combination of textual evidence representing its local context, namely the textual content appearing in its vicinity in the parent page, with visual evidence associated with its global context,

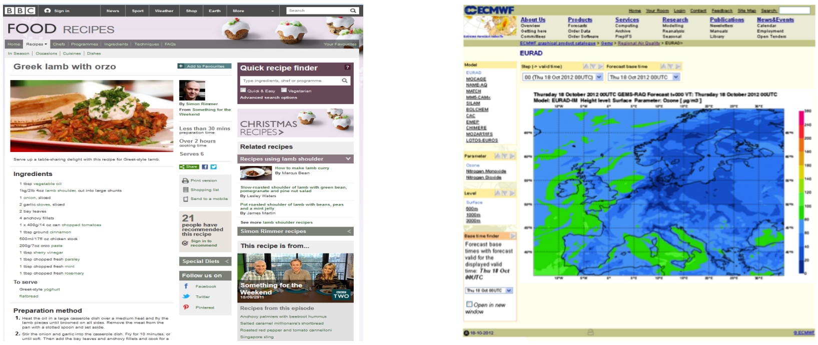

Fig. 1 Examples of Web pages containing images that either complement their textual content, such as the Web page providing a food recipe on the left (extracted from *http://www.bbc.co.uk/food/recipes/greek_lamb_with_orzo_36568*), or encode their informational content in a visual form, such as the Web page providing air quality measurements and forecasts on the right (extracted from *http://gems.ecmwf.int/*).

namely the presence of multimedia items relevant to the topic within the parent page. This is achieved by the late fusion of textual and visual classification confidence scores obtained by supervised machine learning methods based on Support Vector Machines (SVMs).

The proposed focussed crawling approach aims to be beneficial for the domain-specific discovery of Web resources on a wide range of topics and especially in cases where: (i) the Web resources on the given topic commonly encode (a significant part of) their informational content as multimedia, while their textual content may often be rather sparse, (ii) the multimedia items encoding the topical information share common characteristics and can therefore be grouped into (a few) categories that are representative of the domain, and (iii) effective and efficient classifiers are available for recognising such categories of multimedia items. To this end, this work investigates the efficacy of the proposed focussed crawler in a domain that exhibits such characteristics: the domain of environmental Web resources that provide air quality measurements and forecasts; see Figure 1(right) for a characteristic example.

Environmental Web resources that provide air quality measurements and forecasts report the concentration values of several air pollutants, such as sulphur dioxide, nitrogen oxides and dioxide, thoracic particles, fine particles and ozone, measured or forecact for specific regions and time periods [11]. Needless to say that the automatic discovery of such resources is of great importance given the significant impact of air quality on life quality, due to its effect on human health (e.g., allergies, asthma, and respiratory diseases in general) and on numerous human outdoor activities (ranging from professional endeavours, such as agriculture, to leisure pursuits, such as sports and holiday planning). Empirical studies on air quality Web resources [9,22,14] have revealed that such 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 correspond to graphical representations of matrix data

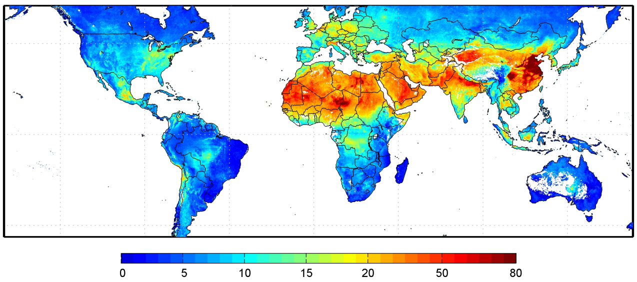

Fig. 2 Heatmap showing the concentration ($\mu g/m^3$) of fine particles $\mathrm{PM}_{2.5}$ over the globe; extracted from *http://www.nasa.gov/*.

with a colour index scale illustrating pollutant concentrations over geographically bounded regions; see Figure 2 for a typical example. This observation motivates us to form the hypothesis that the presence of a heatmap in a Web resource already fetched by a crawler focussed on air quality indicates (i) that the resource is indeed highly relevant to the topic, and (ii) that the heatmap would be a useful representation of the global context for estimating the relevance of hyperlinks to be subsequently selected by such a focussed crawler.

This work proposes a classifier-guided focussed crawling approach for the discovery of Web resources about a given topic that estimates the relevance of hyperlinks to unvisited resources based on the combination of multimedia evidence. This is particularly novel, since state-of-the-art classifier-guided focussed crawlers rely mainly on textual evidence [16] and, to the best of our knowledge, visual evidence has not been previously considered in this context (see Section 2 for a review of related work). As an extension of a preliminary study [27], this work has the following additional contributions:

1. It expresses the proposed focussed crawling approach that combines multimedia evidence for estimating the relevance of a hyperlink pointing to an unvisited Web resource, and thus predicting the benefit of selecting this hyperlink, in a general, parameterised form that allows for the modelling of different hyperlink selection methods by varying the parameter settings (see Section 3).
2. In addition to the hyperlink selection methods, it also introduces hyperlink exploration strategies that are applied during the focussed crawling, so as to discover highly relevant pages that lie behind apparently irrelevant pages, not selected by the focussed crawler; these hyperlink exploration strategies are also based on the combination of multimedia evidence (see Section 3).
3. It evaluates the hyperlink selection methods and the hyperlink exploration strategies of the proposed classifier-guided multimedia focussed crawler in the domain of air quality environmental Web resources by performing experiments that indicate the benefits of incorporating visual evidence, over the use of textual features alone (see Section 4 for a presentation of the textual and visual evidence taken into account, Section 5 for a description of the experimental set-up, and Section 6 for a discussion on the experimental

results). These experiments rely on much larger focussed crawls and are much more extensive, compared to the preliminary study [27], leading thus to more reliable findings.

4. Finally, it refines several aspects of the implementation for improving the effectiveness of the focussed crawling, including the application of (i) a filtering mechanism for removing, prior to classification, small-size images that are unlikely to contain useful information (e.g., logos and layout elements), and (ii) a more comprehensive URL normalisation method for improving the process of URL duplicate detection (see Section 5).

2 Related Work

The automatic discovery of Web resources about a given topic is generally viewed as a *domain-specific search* problem. Domain-specific search has been mainly addressed by two categories of techniques: (i) the domain-specific query submission to a general-purpose search engine followed by post-retrieval filtering, and (ii) focussed crawling. In the environmental domain, resource discovery has previously been addressed mainly through the application of techniques from the first category [15,13], while techniques from the second category, as is the focussed crawler proposed in this work, have not really been explored.

In previous approaches from the first category applied in the environmental domain [15,13], the domain-specific queries were generated using empirical information, including the incorporation of geographical terms [13], and were expanded using 'keyword spices' [18], i.e., Boolean expressions of domain-specific terms corresponding to the output of decision trees trained on an appropriate corpus [15]. Post-retrieval classification was performed using SVMs trained on textual features extracted from a suitable training corpus [15]. Such approaches are complementary to the discovery of Web resources using focussed crawlers, and hybrid approaches that combine the two techniques in a common framework are a promising research direction [14].

On the other hand, focussed crawling techniques have been researched since the early days of the Web [8]. Based on the 'topical locality' observation that most Web pages link to other pages that are similar in content [7], 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 [8,2,26,25,19,20,16] 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 [20], or to hyperlink evidence, such as the centrality of the parent page within its neighbouring subgraph [2]; to the best of our knowledge visual evidence has not been considered in this context.

A systematic study on 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 [20].

Earlier focussed crawlers (e.g., [6]) 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 [16]. Classifier-guided focussed crawlers, introduced by Chakrabarti et al. [2], 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 (ODP, `http://www.dmoz.org/`), an open, non-commercially biased resource that is maintained by a large number of diverse editors. 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 [19]; this makes SVMs the classification scheme of choice in guiding focussed crawlers.

Our approach also follows the classifier-guided focussed crawling paradigm, discussed above, and employs SVMs, given their demonstrated effectiveness [19]. Similar to the majority of focussed crawlers reviewed above, our approach represents the local context of hyperlinks based on textual features (such as their anchor text and surrounding text window), but extends existing approaches by proposing the use of visual evidence, in the form of heatmaps, as global evidence in the selection and exploration mechanisms.

Finally, heatmap recognition, which is incorporated by the proposed focussed crawler, has not been extensively researched. Relevant work in the related area of map recognition includes the use of knowledge of the colourisation schemes in U.S. Geological Survey maps in order ro automatically segment them based on their semantic contents (e.g., roads) [10], and the development of techniques for improving segmentation quality of text and graphics in colour map images through the cleaning up of possible errors (e.g., dashed lines) [1]. 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 [17], where the best performing system employed a supervised machine learning method and fused visual descriptors [28]. Regarding heatmaps, in particular, research has mainly focussed on the information extraction from them; this includes methods for reconstructing environmental data out of chemical weather images [9]. More recently, our research group has proposed a method for heatmap recognition [13], which is incorporated in the proposed approach and described in Section 4.2.

3 Multimedia Focussed Crawling

This work proposes a classifier-guided focussed crawling approach for the discovery of Web resources about a given topic that estimates the relevance of hyperlinks to unvisited resources based on the combination of multimedia evidence. To this end, it combines textual evidence representing the *local context* of each hyperlink, namely the textual content appearing in its vicinity in the parent page, with visual evidence associated with its *global context*, namely the presence of images relevant to the topic within the parent page. Driven by the combination of such multimedia evidence, the proposed focussed crawler selects to follow the most promising hyperlinks, i.e., those with high estimates of relevance. Hyperlinks with low estimates of relevance may simply be discarded, given that they appear to be irrelevant, or, alternatively, they may be taken into consideration and explored further, since they might lead to relevant resources, even though they themselves appear to be suboptimal choices.

An overview of the proposed focussed crawling approach is depicted in Figure 3. First the seed pages are added to the frontier. In each iteration, a URL is picked from the list and the page corresponding to this URL is fetched and parsed to obtain its hyperlinks and its embedded images (if any). The textual feature vector extracted from the representation of each hyperlink h based on its local textual context is used as input to an appropriately trained text classifier for obtaining an estimate of its relevance to the topic; the real-valued confidence scores produced by such classifiers are denoted as $score_T(h)$. On the other hand, the visual feature vectors extracted from the set of images $I = \{I_1, \ldots, I_N\}$ downloaded from the parent page of hyperlink h, denoted as $parent(h)$, are used as input to an image classifier appropriately trained for

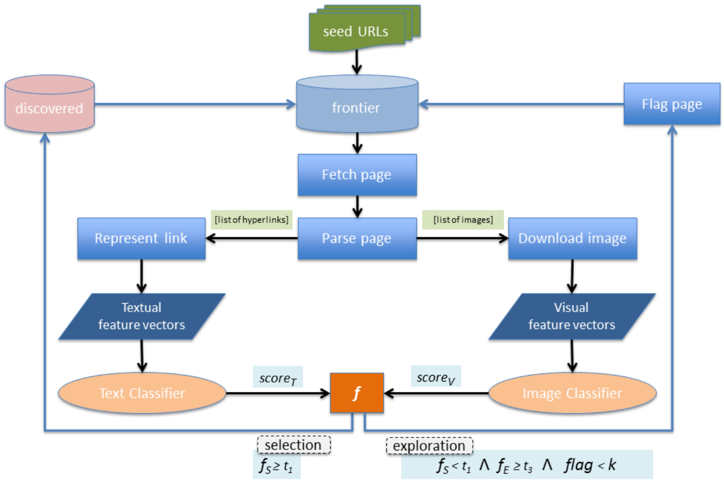

Fig. 3 Focussed crawling based on the combination of multimedia evidence

the topic at hand. The confidence scores for the individual images are then fused into a single score that corresponds to an estimate of the relevance of the global context of h, as this is reflected by the overall presence of image(s) relevant to the topic within its parent page. This work estimates the relevance of the global context of h based on visual evidence as:

$$score_V(parent(h)) = max_j\{score_V(I_j)\}_{j=1}^{N} \tag{1}$$

i.e., considers it to be as relevant as the most relevant image within its parent page. As a result, the presence of even a single representative image of the topic at hand in the parent page, that is recognised by the employed image classifier as such, is sufficient for ensuring high estimates of relevance for the global context. In the case of binary image classifiers, similar to the ones employed in this work (see Section 4.2), the $score_V(parent(h))$ is equal to 1 if the parent page contains at least one image relevant to the topic, and 0 otherwise. The relevance of each hyperlink h is then estimated by the late fusion of these two scores:

$$f(h) = \alpha_1 \times score_T(h) + \alpha_2 \times score_V(parent(h)) \tag{2}$$

where α_1, α_2 are parameters that are appropriately set for expressing different strategies that decide the selection of a hyperlink (Section 3.1) or merely its exploration (Section 3.2), as discussed next.

3.1 Link Selection Based on Multimedia Evidence

The estimate of relevance to the topic at hand of a hyperlink h pointing to an unvisited page p is used for deciding whether to select p for further crawling by adding it to the frontier and also to the list of URLs discovered by the crawler; in particular, a hyperlink is selected if its estimate of relevance is over a set threshold value t_1. This section first describes the proposed link selection approach and then models this process using Equation (2), by appropriately setting the parameters α_1 and α_2.

In the simplest case, the focussed crawler decides to consider only the local context of a hyperlink h, and thus select it, if its text-based classification confidence score, $score_T(h)$, is over t_1. This could happen when the focussed crawler is unable or unwilling to download and process any of the available images (e.g., for efficiency purposes) or when the parent page contains no images; in such cases, the relevance of the global context, $score_V(parent(h))$, cannot be estimated and is defined as being equal to 0.

There are cases, though, in which the local context may not be sufficient for effectively representing relevant hyperlinks, leading them to obtain low confidence scores below the set threshold t_1, and thus to not being selected and fetched; in such cases, the global evidence can be used for adjusting the existing relevance estimates that depend solely on the hyperlinks' local context. As previously mentioned, this is motivated by the 'topical locality' [7, 21] 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 could be accordingly adjusted.

This work considers that the presence of images classified as relevant to the topic in a Web resource that has been already fetched, and hence estimated by the focussed crawler as being relevant, is a strong indication that this resource is indeed highly relevant to the topic. Therefore, this global evidence can be used for adjusting the relevance estimate of hyperlinks so that those with text-based classification confidence scores below the required threshold t_1 can exceed it; in practice, a lower bound threshold t_2 is also set, i.e., only hyperlinks with $score_T(h) \geq t_2$ ($t_2 < t_1$) are considered for adjustment, so that the crawling still largely maintains its focus on the topic. When such adjustments occur, the focussed crawler may as a result also select hyperlinks h with text-based classification confidence scores below t_1, but over t_2, provided that their parent page contains at least one relevant image.

By focussing on binary image classifiers similar to the ones employed in this work (see Section 4.2), the link selection process (S) proposed above can be modelled by setting $\alpha_1 = 1$, $\alpha_2 = t_1 - t_2$ in Equation (2):

$$f_S(h) = score_T(h) + (t_1 - t_2) \times score_V(parent(h)) \tag{3}$$

when requiring $f_S(h) \geq t_1$. In the simple case where no images are available or have been estimated as relevant, $f_S(h) = score_T(h)$ and hyperlinks are selected when their $score_T(h) \geq t_1$. When, on the other hand, at least one relevant image is present in the parent page of the hyperlink, $score_V(parent(h)) = 1$ and $f_S(h) = score_T(h) + t_1 - t_2$; hyperlinks in such pages are selected when their $score_T(h) \geq t_2$. In the case of non-binary image classifiers, Equation (3) can still model the link selection process described above by using an appropriate step function to map $score_V(parent(h))$ into the binary space. In this work, both thresholds, t_1 and t_2, are experimentally set.

Different link selection approaches can be modelled by different settings of α_1 and α_2 in Equation (2), while requiring that the estimated relevance is above a threshold t_1; to this end, the following approaches are investigated:

1. $\alpha_1 = 1$ and $\alpha_2 = 0$: $f_S^{'}(h) = score_T(h)$, i.e., a hyperlink h is selected if its text-based classification confidence score is over t_1, irrespective of the presence of relevant images within its parent page,
2. $\alpha_1 = 0$ and $\alpha_2 = 1$: $f_S^{''}(h) = score_V(parent(h))$, i.e., a hyperlink h is selected if its parent page contains at least one image with confidence score over t_1; since this work considers binary image classifiers, this approach selects all hyperlinks within pages containing at least one relevant image, and
3. $\alpha_1 = score_V(parent(h))$ and $\alpha_2 = 0$: $f_S^{'''}(h) = f_S^{'}(h) * f_S^{'''}(h)$; when considering binary image classifiers, this approach selects all hyperlinks h with $score_T(h) \geq t_1$ embedded within pages containing at least one relevant image.

It should be noted that some of these approaches are subsumed by Equation (3), e.g., $f_S^{'}(h)$ is equivalent to the simple case considered by Equation (3) where the focussed crawler considers that estimates of relevance of the global context are unavailable. The additional notation is introduced for clarity and for better distinguishing between the different cases.

To achieve high precision, the proposed link selection approach would need to employ high threshold values, so as to filter out pages not likely to contain relevant information. To be effective, such an approach requires the existence of URL trails that start from the seed pages and lead to other relevant pages exclusively through links with estimates of relevance over these high thresholds. This high level of greediness that concentrates the link selection process only on hyperlinks with the highest estimates of relevance may lead to missing other promising points that lie behind apparently irrelevant pages, including both pages that are indeed non-relevant, and also pages that have been falsely estimated as non-relevant due to noisy and/or sparse evidence. The incorporation of multiple sources of evidence in the link selection process aims to address the latter case. The former case requires, however, the adoption of strategies that also explore suboptimal links that could, though, lead to further relevant pages, as discussed next.

3.2 Link Exploration Based on Multimedia Evidence

The rather restrictive nature of the proposed link selection policy is evident in situations similar to the one depicted in Figure 4; for simplicity, we assume that there is a single incoming hyperlink h to each page and that the number over each hyperlink denotes the relevance estimation $f_S(h)$ based on Equation (3), with t_1 set to 0.7. If we consider that pages A, B, D_1, D_5, E_1, and E_2 are the ones relevant to the topic, then the URL trail starting from page A and leading to D_1 and D_5, through pages B and C, is "blocked" by page C, whereas the one leading to E_1 and E_2 is in addition "blocked" by D_2. This is caused by both C and D_2 having low scores, either because they are actually not relevant (as assumed here) or due to unreliable predictions. Such situations may lead the focussed crawler to miss out on the discovery of further relevant pages.

To address this issue, we introduce link exploration strategies that may be applied in conjunction with the proposed link selection approaches. When the focussed crawler encounters a hyperlink h pointing to an unvisited page p with estimated relevance $f_S(h)$ below the required threshold t_1, the crawler may still decide to explore its potential. To maintain some focus, not all links with $f_S(h) \leq t_1$ are necessarily explored, but additional thresholds may be set such that there is also a selection process for the links to be explored. Similarly to before, this could be modelled by appropriately setting α_1 and α_2 in Equation (2) and requiring that the estimated relevance is above a threshold t_3, where $t_3 < t_1$. To this end and equivalently to the approaches defined in the previous section, the following exploration strategies (E) are proposed: (i) $f_E^{'}(h) = score_T(h)$, (ii) $f_E^{''}(h) = score_V(parent(h))$, and (iii) $f_E^{'''}(h) = f_E^{'}(h) * f_E^{'''}(h)$. Further exploration strategies could be devised for

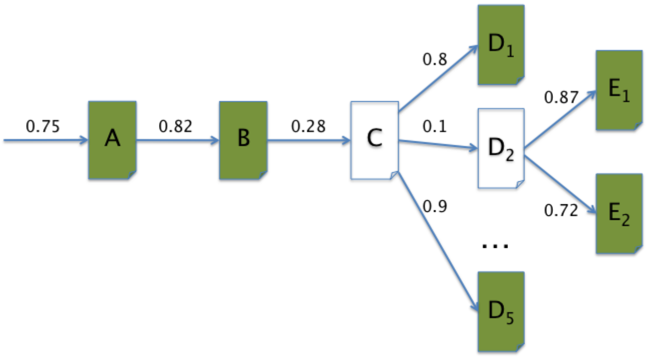

Fig. 4 Illustration of linked Web pages and their scores assigned by the focussed crawler

different settings of the available parameters. In this work, the threshold t_3 is experimentally set.

A page p pointed by h that is selected only for exploration is not considered as relevant and, as a result, it is not added to the collection of pages discovered by the focussed crawler (see Figure 3). It is simply fetched and parsed so as to classify its hyperlinks using their local and global context, with the hope that they point to pages with relevant information. It is also flagged to denote that it is part of an exploration trail and assigned a number indicating its depth within that trail; for example, page C in Figure 4 would be assigned the number 1. If any of its hyperlinks achieves a relevance score over t_1, then the link is selected and the focussed crawling continues as normal. Otherwise, if none of its hyperlinks are considered relevant, we first examine whether the exploration trail has reached the maximum allowed exploration depth k. If that is the case, the exploration stops there and all pages in that trail are discarded; otherwise, the exploration continues. Consider for instance the example in Figure 4; for $k = 1$, page C would be explored and D_1 and D_5 would be crawled, whereas page D_2 would only be explored for $k \geq 2$. Given that this strategy allows us to skip up to k pages that may "block" a trail to relevant information, it is referred to as *skip-k*. To achieve a balance between the crawler's need to focus on the selection of the most promising links and the need for exploration, this work considers $k = 1$ and the employed exploration strategy is referred to as *skip-one*.

This work employs the proposed classifier-guided focussed crawling approach that combines multimedia evidence for the selection and exploration of hyperlinks in order to discover Web resources that belong to the environmental domain, and in particular those that provide air quality measurements and forecasts. To this end, it combines the local context of each hyperlink, namely the textual content appearing in its vicinity in the parent page, with its global context, namely the occurrence of heatmaps, regarded as the characteristic images of this domain, within the parent page. The classifiers employed for estimating the relevance of each of these types of context are described next.

4 Context Classification in Environmental Web Resources

The relevance of the local and global context of hyperlinks appearing within environmental Web resources providing air quality measurements and forecasts is estimated using appropriately trained classifiers based on textual (Section 4.1) and visual evidence (Section 4.2), respectively, as discussed next.

4.1 Text-Based Link Classification

Focussed crawling applications typically represent the local context of hyperlinks by textual features extracted from the following fields (and concatenations thereof) [16]: (i) the anchor text of the hyperlink, (ii) a text window of x characters surrounding the hyperlink that may or may not contain the anchor text of adjacent links, and (iii) the URL of the hyperlink. Text-based classification of this local context is typically performed using a supervised machine learning approach based on SVMs, given their demonstrated effectiveness in such applications [19]. Based on the best performing results of our preliminary study [27], this work represents the local context of each hyperlink using its anchor text a together with a text window s of $x = 50$ characters that does not overlap with the anchor text of adjacent links (i.e., it stops as soon as it encounters another link); this a+s representation is classified using SVMs.

In the training phase, 711 samples (100 positive, 611 negative) were collected, so as to first build a vocabulary for representing the samples in the textual feature space and also for training the model. Each sample corresponds to the a+s representation of a hyperlink pointing to a 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[1]. 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 5). The vocabulary is built by accumulating all the terms from the a+s representations of the positive samples and eliminating all stopwords. The generated lexicon consists of 207 terms with the following being the 10 most frequent: *days*, *ozone*, *air*, *data*, *quality*, *today*, *forecast*, *yesterday*, *raw*, and *current*.

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. A further feature representing the number of the geographical terms in the sample's a+s representation is added, given the importance of such terms in the environmental domain [15]. The presence of such terms is determined based on a geographical lexicon of 3,625 terms obtained from a geographical database. To avoid overestimation of their

[1] Personalised Environmental Service Configuration and Delivery Orchestration (`http://www.pescado-project.eu/`).

effect, such geographical terms were previously removed from the vocabulary. 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 the *a+s* representation and the number of geographical terms within the same representation. The text-based classification confidence score $score_T(h)$ of each hyperlink h is then obtained by employing the classifier on the feature vector and reflects the distance of the testing sample to the hyperplane in the feature space induced by the kernel; this work considers that the values of these confidence scores are between 0 and 1 (see Section 5.4).

4.2 Heatmap Recognition

The representation of the global context of hyperlinks embedded within environmental Web resources providing air quality measurements and forecasts is based on the presence of heatmaps within such pages. Heatmap recognition is performed by applying a recently developed approach by our research group [13]. That investigation on heatmap binary classification using SVMs and a variety of visual features indicated that, overall, the MPEG-7 [4] descriptors demonstrated a slightly better performance than the other tested visual features (SIFT [5] and AHDH[2] [23]).

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

- Scalable Color Descriptor (SC): a Haar-transform based encoding scheme that measures colour distribution over an entire image, quantised 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 each 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.

The evaluation was performed by training an SVM classifier with an RBF kernel for each of the aforementioned features and their early fusion combinations on dataset A consisting of 2,200 images (600 relevant, i.e., heatmaps) and testing it on dataset B consisting of 2,860 images (1,170 heatmaps)[3]. Their early fusion (SC-EH-HT), as well as the feature EH on its own produced the best results by achieving accuracy equal to 0.98 and 0.96, respectively.

In this work, both the EH and the SC-EH-HT models trained on dataset A are employed. An image is classified as a heatmap if at least one of these

[2] Adaptive Hierarchical Density Histogram.

[3] Both datasets are available at: `http://mklab.iti.gr/project/heatmaps`.

classifiers considers it to be a heatmap, i.e., a late fusion approach based on a logical OR is applied for estimating $score_V$.

5 Evaluation

To assess the effectiveness of the proposed focussed crawling approaches, a series of experiments are performed using a set of seed URLs 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. This section describes the experimental set-up (Section 5.1), the criteria applied for assessing the relevance of the experimental results (Section 5.2), and the employed performance metrics (Section 5.3), while it also provides some implementation details (Section 5.4).

5.1 Experiments

Table 1 lists the seed URLs[4] used in the experiments; half of them contain at least one heatmap. A relatively small seed set is employed so as to keep the evaluation tractable while investigating larger crawling depths, compared to our preliminary study [27]. Starting from these 6 seeds, crawls at depth $\in \{1, 2, 3\}$ were performed in June 2014 and were locally stored.

Table 1 List of seed URLs

URL	heatmap present
1. `http://aircarecolorado.com/`	
2. `http://db.eurad.uni-koeln.de/en/`	✓
3. `http://gems.ecmwf.int/`	✓
4. `http://uk-air.defra.gov.uk/`	
5. `http://www.baaqmd.gov/The-Air-District.aspx`	
6. `http://www.londonair.org.uk/LondonAir/Default.aspx`	✓

First, the effectiveness of the text-based link selection approach $f_S^{'}(h)$ applied by a classifier-guided focussed crawler that employs the *a+s* representation of hyperlinks is examined; in this case, no images are considered and a page is fetched if its text-based classification confidence score is above a threshold t_1. Experiments are performed at different crawling depths (depth $\in \{1, 2, 3\}$) and for t_1 values ranging from 0.1 to 0.9 at step 0.1. The case when $t_1 = 0.0$ is not considered, since this crawl corresponds to a breadth-first search where all hyperlinks are fetched and no focussed crawling is performed.

When a hyperlink appears more than once within a Web resource, only the one with the highest score is taken into account for evaluation purposes. Duplicates are identified after applying URL normalisation that converts the URL scheme and host to lower case, capitalises letters in escape sequences,

[4] These URLs are different to the ones used for training the classifiers.

decodes percent-encoded octets of characters, and also removes from the URL the default port, well known directory indexes (e.g., "index.html"), empty queries, fragments, and also duplicate and trailing slashes.

Next, the effectiveness of incorporating multimedia evidence, in the form of heatmaps, in the focussed crawler is investigated through the application of the $f_S(h)$, $f_S^{''}(h)$, and $f_S^{'''}(h)$ link selection approaches. In the case of $f_S(h)$, a page pointed by a hyperlink h is fetched if its $score_T(h) \geq t_1$, or if its $score_T(h) \geq t_2$ ($t_2 < t_1$) and its parent page contains at least one heatmap. For $f_S^{''}(h)$, a hyperlink is selected if its parent page contains at least one heatmap, whereas for $f_S^{'''}(h)$, link selection takes place when if its $score_T(h) \geq t_1$ and its parent page contains at least one heatmap, i.e., the former is equivalent to the latter for $t_1 = 0.0$. Experiments are performed for t_1 and t_2 values ranging from 0.0 to 0.9 at step 0.1, while maintaining that $t_2 < t_1$. These experimental results are compared against two baselines: (i) the results of the corresponding text-based focussed crawler $f_S^{'}(h)$ for threshold t_1, and (ii) the results of the corresponding text-based focussed crawler $f_S^{'}(h)$ 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), its images are extracted and compiled into a list. The crawler iteratively downloads each image in the list and extracts its visual features only if its size exceeds certain thresholds, so as to exclude small-size images that do not likely contain useful information. These size thresholds have been determined based on an analysis of the images downloaded and classified in our preliminary study, and were set in our experiments to 50 and 100 pixels for height and width, respectively. Then, the heatmap classification method is applied and the iteration continues until a heatmap is recognised or a maximum number of images is downloaded from each page (set to 20 in our experiments).

Finally, the *skip-one* link exploration strategies $f_E^{'}(h)$, $f_E^{''}(h)$, and $f_E^{'''}(h)$ are applied in conjunction with the above link selection approaches for t_3 values ranging from 0.1 to 0.9 at step 0.1, while maintaining that $t_3 < t_1$ and $t_3 \leq t_2$ (where applicable).

5.2 Relevance Assessments

Given the large scale of crawling experiments, automatic relevance assessments are often employed based, for instance, on the lexical similarity of the textual content of the discovered Web pages to a detailed textual description of the topic, or on the confidence score of a classifier trained on the topic [24,19]. Since this introduces some level of noise that may inadvertently affect the performance evaluation, and thus not allow us to fully gauge the effectiveness of the proposed approaches, we decided to manually assess the Web pages pointed to by the extracted hyperlinks using the following three-point relevance scale:

- (*highly*) *relevant*: Web resources that provide air quality measurements or/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, for example, from drop-down menus or through tick boxes.

- *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 and the standards they should adhere to, 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.

Given that most performance metrics require binary relevance judgements, these multiple grade relevance assessments are also mapped onto the two dimensional space. This mapping is performed in two different ways depending on: (i) whether we are strictly interested in discovering resources containing air quality data, or (ii) whether we would also be interested in information *about* air quality measurements and forecasts. These two mappings are denoted as:

- *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.

5.3 Performance Metrics

The standard retrieval evaluation metrics of *precision* and *recall*, as well as their harmonic mean, the F_1-*measure*, are typically applied for assessing the effectiveness of a focussed crawler [24]. 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 [16]: (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; only the former could apply in our experimental set-up. Given though the size of our seed set and the relatively small crawling depth, this approach is not applicable, as it would need much larger crawling depths to ensure that many of these pages are found. As an alternative, this work considers the relevant pages crawled by our most comprehensive link selection approach (i.e., $f_S(h)$ in Equation 3) as our recall base; statistics regarding the number of relevant pages in our crawled datasets are listed in Table 2 (see also the discussion in Section 6). This enables us to approximate recall for all the focussed crawling approaches examined in our experiments.

5.4 Implementation

The implementation of the proposed focussed crawler 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 the proposed approaches. The text-based classifier was implemented using the LIBSVM package of the Weka machine learning software (`http://www.cs.waikato.ac.nz/ml/weka/`), with the text-based classification scores corresponding to its probabilistic outputs [12]; the implementation of the heatmap classifier was based on the LIBSVM [3] library.

6 Results

Starting from the seeds in Table 1, a general-purpose (non-focussed) crawler would fetch a total of 444, 3912, and 19564 unique pages, at depth $\in \{1, 2, 3\}$, respectively, illustrating an exponential increase in the number of pages fetched at deeper crawls. The application of the proposed link selection approaches $f_S^{'}(h)$, $f_S(h)$, and $f_S^{'''}(h)$ during focussed crawling results in fetching the number of pages depicted in Figure 5, clearly illustrating a similar exponential increase for deeper crawls, particularly for low values of t_1. Given also that the lower the t_1 thresholds, the higher the number of pages that are crawled, the focussed crawling approach that employs the $f_S^{'}(h)$ text-based link classifier with $t_1 = 0.1$ at depth $= 3$ crawls the largest number of pages among all text-based focussed crawling approaches (796 pages), while its combination $f_S(h)$ with the heatmap classifier for $t_2 = 0$ achieves the same among all multimedia-based focussed crawling approaches (1609 pages). A comparison to the numbers of pages crawled by the same two approaches in our preliminary study [27], which though only employed a depth $= 1$, indicates a ten-fold increase (85 vs. 796, 176 vs. 1609), and thus the much larger scale of the experiments in this work. Moreover, the extremely low numbers of pages crawled by the $f_S^{'''}(h)$ $t_1 \geq 0.1$ leads us to remove this link selection approach for further consideration. For $t_1 = 0$, $f_S^{'''}(h)$ is equivalent to $f_S^{''}(h)$, a link selection approach that does not depend on any thresholds; since $f_S^{''}(h)$ crawls a reasonable number of pages, it will be further examined.

The set of 1609 pages fetched by the $f_S(h)$ multimedia focussed crawler corresponds to a superset of the pages fetched by all other approaches for all configurations of $0.1 \leq t_1, t_2, t_3 \leq 0.9$ (s.t. $t_2 < t_1$, $t_3 < t_1$, and $t_3 \leq t_2$); therefore, we only need to assess this in order to evaluate all proposed approaches. Assessments of the 1609 pages on the basis of the criteria discussed in Section 5.2 resulted in: 461 (28.65%) highly relevant pages, 210 (13.05%) partially relevant, and 938 (58.30%) non-relevant ones. Their *strict* and *lenient* mappings lead to the distributions of relevance assessments listed in Table 2.

Table 2 Relevance assessments when the 3-point scale judgements are mapped to binary.

	Strict		Lenient	
Relevant	461	(28.65)%	671	(41.70)%
Non-Relevant	1148	(71.35)%	938	(58.30)%
All	1609	(100.0)%	1609	(100.0)%

6.1 Effectiveness of Link Selection in Focussed Crawling

The results of the experiments that evaluate the effectiveness of the text-based focussed crawler (i.e., a focussed crawler that employs the $f^{'}_{S}(h)$ link selection method) at depth $\in \{1, 2, 3\}$, when applying strict or lenient relevance assessments, are depicted in Figure 6. The precision achieved is comparable to that of state-of-the-art text-based classifier-guided focussed crawling approaches that start from a similar (or larger) number of seeds and crawl a similar number of pages, while employing SVMs as their classification scheme [19]. The F1-measure also achieves similar results. As expected, the absolute values for both effectiveness metrics are much higher in the lenient case, compared to the strict. Moreover, an indication of the accuracy of the underlying text-based link classifier can be obtained by considering the hyperlinks selected at depth = 1 (67 hyperlinks: 29 positive and 38 negative) and is close to 0.8.

The highest overall precision for each depth value is achieved for $t_1 = 0.2$ for the case of lenient relevance assessments, while the results are mixed for the strict case, where the best results are observed for $t_1 = 0.4$, $t_1 = 0.7$, and $t_1 = 0.2$ for depth $\in \{1, 2, 3\}$, respectively. For the F1-measure, the best results are observed for $t_1 = 0.2$ for both lenient and strict relevance assessments. A further examination reveals that there is also a significant drop in the performance for $t_1 = 0.3$, for both metrics and for both the strict and lenient cases. To gain an understanding of this phenomenon, we investigated the relevant pages that are crawled when employing each threshold value, so as to determine the reasons behind their low percentages for some t_1 values.

Figure 7 shows the number of relevant pages crawled for each $t_1 = x$ that are not crawled though for any $t_1 > x$; each bar also shows the distribution of these relevant pages across the different crawling depths. For both the strict and the lenient case, a large number of relevant pages are crawled when t_1 is set to 0.2, which are not crawled for higher values of t_1, with many of these

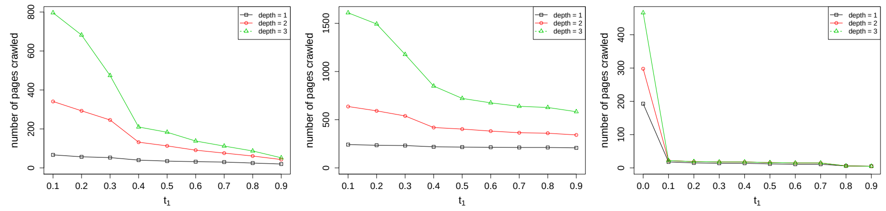

Fig. 5 Number of pages crawled at depth $\in \{1, 2, 3\}$ by the $f^{'}_{S}(h)$ (left), $f_S(h)$ ($t_2 = 0$) (middle), and $f^{'''}_{S}(h)$ (right) link selection approaches for threshold $t_1 \in \{0.1, ..., 0.9\}$.

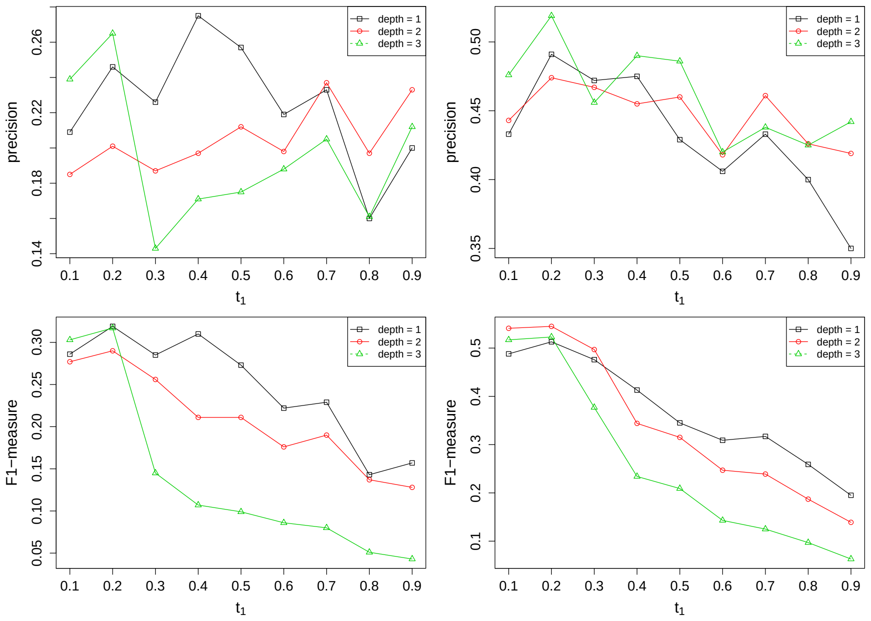

Fig. 6 Precision (top) and F1-measure (bottom) of the focussed crawl for the $f'_S(h)$ link selection method at depth $\in \{1, 2, 3\}$ for threshold $t_1 \in \{0.1, ..., 0.9\}$ when *strict* relevance assessments (left) and *lenient* relevance assessments (right) are employed.

crawled at depth $= 3$. A closer inspection of the URL trails followed by the crawler reveals situations similar to that depicted in Figure 4, where a page, such as C, with a score between 0.2 and 0.3 links to highly relevant pages at depth $= 3$. However, these pages can not be crawled when $t_1 \geq 0.3$, unless the *skip-one* exploration strategy is employed; the application of these strategies will be discussed in the next section.

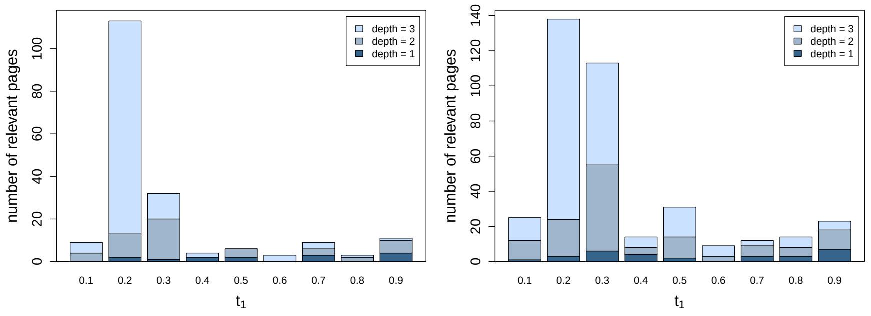

Fig. 7 Number of relevant pages only crawled when the $f'_S(h)$ link selection method is employed for threshold $t_1 = x \in \{0.1, ..., 0.9\}$ and not for $t_1 > x$, when *strict* relevance assessments (left) and *lenient* relevance assessments (right) are employed. Each bar shows the distribution of relevant pages across depth $\in \{1, 2, 3\}$.

We now examine the potential benefits of incorporating multimedia in the form of heatmaps in the crawling process; to this end, the focussed crawler combines the text-based classifier with results from the heatmap classifier. Tables 3-6 present the results of the $f_S(h)$ experiments for both effectiveness metrics at depth = 3, when applying strict and lenient relevance assessments, and 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, whereas the values in bold correspond to improvements over both baselines. Since the ground truth has only been built for the pages with $score_T \geq 0.1$, the precision when fetching pages with $0.0 \leq score_T < 0.1$ cannot be accurately estimated; this results in an empty first cell in the last row in both tables.

For precision, the significant improvements that are observed for multiple threshold values indicate the benefits of incorporating visual evidence (in the form of heatmaps) as global evidence in a focussed crawler for the environmental domain; our preliminary study [27] had also made similar observations, but in focussed crawls of much smaller scale. For the F1-measure, significant improvements over both baselines are only observed for $t_2 = 0.0$ over all values of t_1, whereas improvements (or equivalent performance) are achieved for all threshold values w.r.t. the $f'_S(h) \geq t_1$ baseline. This indicates that the addi-

Table 3 Precision of the focussed crawler for the $f_S(h)$ link selection method at depth = 3 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									$f'_S(h) \geq t_1$
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
0.1	0.287									0.239
0.2	**0.303**	0.263								0.265
0.3	0.278	0.143	0.144							0.143
0.4	**0.339**	0.169	0.173	0.171						0.171
0.5	**0.397**	0.177	0.182	**0.183**	**0.183**					0.175
0.6	**0.416**	0.196	0.203	**0.204**	**0.204**	**0.194**				0.188
0.7	**0.423**	0.226	0.235	**0.237**	**0.237**	**0.226**	**0.219**			0.205
0.8	**0.426**	0.230	0.241	**0.243**	**0.243**	**0.229**	**0.221**	**0.206**		0.161
0.9	**0.451**	**0.287**	**0.307**	**0.311**	**0.311**	**0.300**	**0.290**	**0.269**	0.204	0.212
$f'_S(h) \geq t_2$	–	0.239	0.265	0.143	0.171	0.175	0.188	0.205	0.161	

Table 4 F1-measure of the focussed crawler for the $f_S(h)$ link selection method at depth = 3 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									$f'_S(h) \geq t_1$
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
0.1	**0.446**									0.303
0.2	**0.463**	0.315								0.317
0.3	**0.399**	0.146	0.147							0.145
0.4	**0.440**	0.109	0.109	0.107						0.107
0.5	**0.484**	0.104	0.105	0.105	0.105					0.099
0.6	**0.495**	0.095	0.096	0.096	0.096	0.090				0.086
0.7	**0.491**	0.096	0.097	0.097	0.097	0.090	0.087			0.080
0.8	**0.491**	0.090	0.091	0.091	0.091	0.085	0.082	0.075		0.051
0.9	**0.504**	0.085	0.086	0.086	0.086	0.080	0.075	0.068	0.043	0.043
$f'_S(h) \geq t_2$	–	0.303	0.317	0.145	0.107	0.099	0.086	0.080	0.051	

tional pages that are considered by the multimedia-based focussed crawling $f_S(h) \geq t_1$ compared to the text-based crawling $f_S^{'}(h) \geq t_1$, are beneficial both in terms of precision and recall, but they still cannot reach the coverage achieved by the text-based crawling $f_S^{'}(h) \geq t_2$, with the exception of $t_2 = 0.0$. Therefore focussed crawls that are also oriented towards recall should take into account hyperlinks in pages containing heatmaps irrespective of their text-based score. However, considering only such pages (i.e., applying the visual-based $f_S^{''}(h)$ link selection method) does not achieve better results than the multimedia-based $f_S(h)$; the F1-measure for the $f_S^{''}(h)$ is 0.482 and 0.409 for the strict and lenient cases, respectively. Therefore, the incorporation of visual evidence is indeed beneficial, but particularly in combination with the textual evidence.

Examining in more detail Tables 3 and 5 indicates that, for the strict case, in particular, the improvements in precision are substantial, with the best precision 0.451 (achieved for $t_1 = 0.9$ and $t_2 = 0$) being far higher than the precision achieved by the text-only approach over any threshold value; this corresponds to 0.265 for $t_1 = 0.2$. Moreover, the precision improves as t_1 increases, indicating the merits of reducing the noise by imposing more stringent criteria on the pages being fetched. For the lenient case, there also substantial improvements over the corresponding baselines; e.g., for $t_1 = 0.9$ and $t_2 = 0.3$,

Table 5 Precision of the focussed crawler for the $f_S(h)$ link selection method at depth = 3 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									$f_S^{'}(h) \geq t_1$
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
0.1	0.417									0.476
0.2	0.433	0.515								0.519
0.3	0.420	0.449	0.454							0.456
0.4	0.437	0.475	0.486	0.490						0.490
0.5	**0.497**	0.474	0.487	0.489	0.489					0.486
0.6	**0.484**	0.412	0.427	0.430	0.430	0.424				0.420
0.7	**0.484**	0.435	0.454	**0.458**	0.458	0.452	**0.447**			0.438
0.8	**0.485**	0.442	0.463	**0.467**	0.467	0.457	**0.452**	**0.441**		0.425
0.9	**0.491**	0.463	0.493	**0.500**	**0.500**	**0.500**	**0.493**	**0.478**	**0.444**	0.442
$f_S^{'}(h) \geq t_2$	–	0.476	0.519	0.456	0.490	0.486	0.420	0.438	0.425	

Table 6 F1-measure of the focussed crawler for the $f_S(h)$ link selection method at depth = 3 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									$f_S^{'}(h) \geq t_1$
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
0.1	**0.589**									0.517
0.2	**0.597**	0.521								0.523
0.3	**0.535**	0.376	0.377							0.377
0.4	**0.488**	0.234	0.235	0.234						0.234
0.5	**0.515**	0.211	0.213	0.213	0.213					0.209
0.6	**0.485**	0.149	0.150	0.150	0.150	0.146				0.143
0.7	**0.472**	0.135	0.136	0.136	0.136	0.132	0.130			0.125
0.8	**0.468**	0.128	0.129	0.129	0.129	0.124	0.121	0.116		0.097
0.9	**0.456**	0.098	0.099	0.099	0.099	0.094	0.092	0.087	0.067	0.063
$f_S^{'}(h) \geq t_2$	–	0.517	0.523	0.377	0.234	0.209	0.143	0.125	0.097	

precision reaches 0.500 compared to the 0.442 and 0.456 baselines. However, although this corresponds to the best overall precision for the multimedia focussed crawling approaches, it does not manage to outperform the best overall precision for the text-based focussed crawling (0.519 for $t_1 = 0.2$). As this issue is due to the reasons explained above, it is addressed with the application of the skip-one exploration strategy.

6.2 Effectiveness of Link Exploration in Focussed Crawling

Experiments that employ the *skip-one* exploration are only performed for depth = 3, since the application of the link exploration strategies by the focussed crawling approaches is only meaningful for larger crawling depths. Figure 8 presents the precision of the focussed crawler that employs the text-based $f'_S(h)$ link selection method in conjunction with the skip-one $f'_E(h)$, $f''_E(h)$, and $f'''_E(h)$ exploration strategies, when employing lenient relevance assessments; t_1 and t_3 values range from 0.1 to 0.9 at step 0.1, while maintaining $t_3 < t_1$. These are compared against a baseline corresponding to the best performing text-based focussed crawling approach over all t_1 values.

Improvements in the effectiveness are observed only for the text-based exploration strategy $f'_E(h)$ and for low values of t_3 ($t_3 \leq 0.2$) across all t_1 values (apart from $t_1 = 0.2$). Furthermore, in all these cases, the precision improves, as t_1 increases, leading to significant improvements over the baseline for some

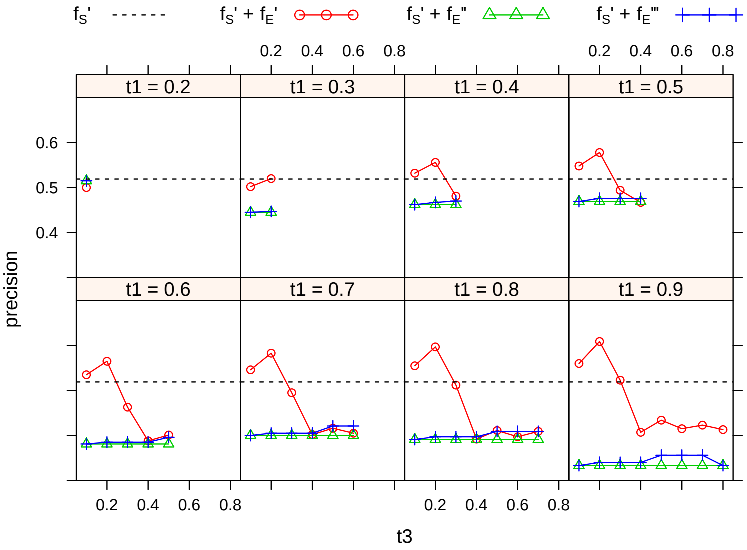

Fig. 8 Precision of the $f'_E(h)$, $f''_E(h)$, and $f'''_E(h)$ exploration strategies by a focussed crawler that employs the $f'_S(h)$ text-based link selection method at depth = 3 for thresholds $t_1, t_3 \in \{0.1, ..., 0.9\}$, s.t. $t_3 < t_3$, when *lenient* relevance assessments are employed.

configurations of t_1, t_3; e.g., for $t_1 = 0.9$ and $t_3 = 0.2$, the improvement is 17% over the best performing text-based focussed crawler (0.609 vs. 0.519) and reaches 38% when compared against the text-based focussed crawler for the same $t_1 = 0.9$ (0.609 vs. 0.442). Similar trends are also observed both for the case of strict relevance assessments and for the F1-measure (results not shown). The results indicate that the text-based link exploration strategy in conjunction with the text-based link selection method is beneficial, as it allows to have the best of both worlds: on the one hand, it ensures that Web pages classified with a high degree of confidence are fetched (by using high t_1 values), while on the other hand, it provides the freedom to remove potential "blocks" in the trails towards such pages (through appropriate t_3 values).

Figure 9 presents the precision of the application of the $f_E^{'}(h)$, $f_E^{''}(h)$, and $f_E^{'''}(h)$ exploration strategies by the multimedia focussed crawler $f_S(h)$ (acting as a baseline), for depth = 3, when applying lenient relevance assessments; $t_3 = 0.2$, while t_1 and t_2 values range from 0.3 to 0.9 at step 0.1, such that $t_2 < t_1$. In actual fact, experiments were conducted for all t_3 values ranging from 0.1 to 0.8 at step 0.1, but only the results achieved for $t_3 = 0.2$ are presented since this is the value that achieved the best results in Figure 8. A comparison against the baseline indicates improvements in the effectiveness by the text-based and the visual-based exploration strategies, $f_E^{'}(h)$ and $f_E^{''}(h)$, respectively, for all t_1 and t_2 threshold values, with the former performing slightly better than the latter. Similar trends are also observed for the strict relevance assessments

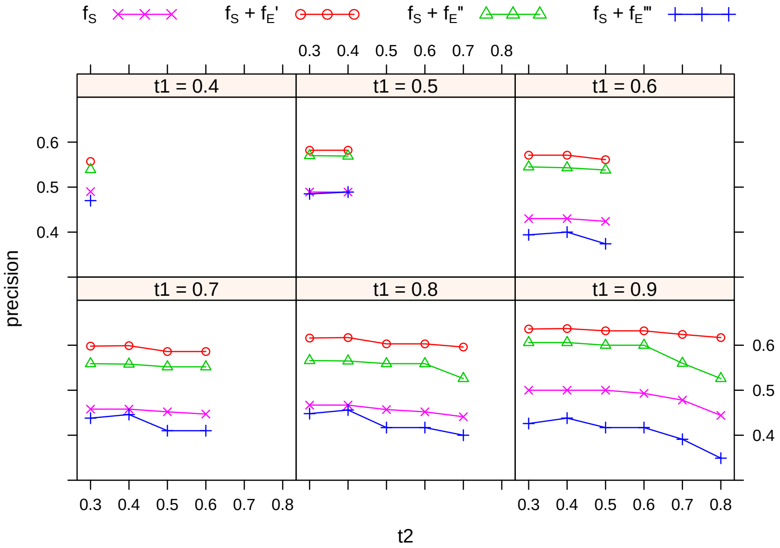

Fig. 9 Precision of the $f_E^{'}(h)$, $f_E^{''}(h)$, and $f_E^{'''}(h)$ exploration strategies by a focussed crawler that employs the $f_S(h)$ multimedia-based link selection method at depth = 3 for thresholds $t_3 = 0.2$ and $t_1, t_2 \in \{0.3, ..., 0.9\}$, s.t. $t_2 < t_1$, when *lenient* relevance assessments are employed.

and for the F1-measure (results not shown). These results clearly indicate the benefits of applying link exploration strategies while also incorporating multimedia evidence in the link selection process of the focussed crawling.

7 Conclusions

This work proposed a novel classifier-guided focussed crawling approach for the discovery of resources about a given topic that combines multimedia (textual + visual) evidence for selecting the hyperlinks to follow by predicting the benefit of fetching an unvisited Web resource. In addition, it introduced hyperlink exploration strategies that can be applied during the focussed crawling, so as to take into account highly relevant pages that lie behind apparently irrelevant pages, not selected by the focussed crawler; these hyperlink exploration strategies are also based on the combination of multimedia evidence. The proposed focussed crawling approach was employed for the discovery of environmental Web resources providing air quality measurements and forecasts. The results of our evaluation experiments indicated the effectiveness of incorporating visual evidence in the hyperlink link selection process of focussed crawling, over the use of textual features alone, particularly in conjunction with the application of a text-based hyperlink exploration strategy.

Future work includes the consideration of other types of images common in environmental Web resources, such as diagrams, the incorporation of additional local evidence, such as the distance of the hyperlink to the heatmap, and research towards improving the effectiveness of the textual classification by taking into account also the textual content of the entire parent page. Furthermore, we plan to investigate the adaptation of classifiers during the crawling process, and also decision mechanisms for determining the cases when the incorporation of global evidence would be benificial. Finally, we aim to apply the proposed focussed crawler in other domains where information is commonly encoded in multimedia form, such as food recipes.

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

References

1. Cao, R., Tan, C.: Text/graphics separation in maps. In: D. Blostein, Y.B. Kwon (eds.) Graphics Recognition: Algorithms and Applications, 4th IAPR International Workshop on Graphics Recognition (GREC 2001), Selected Papers, *Lecture Notes in Computer Science*, vol. 2390, pp. 167–177. Springer Berlin Heidelberg (2002)
2. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: A new approach to topic-specific web resource discovery. In: Proceedings of the 8th International Conference on World Wide Web, (WWW 1999), pp. 1623–1640 (1999)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(3), 27 (2011)

4. Chang, S.F., Sikora, T., Puri, A.: Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology **11**(6), 688–695 (2001)
5. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: Proceedings of the British Machine Vision Conference (BMVC 2011), pp. 1–12 (2011)
6. Cho, J., Garcia-Molina, H., Page, L.: Efficient crawling through URL ordering. Computer Networks **30**(1-7), 161–172 (1998)
7. Davison, B.D.: Topical locality in the web. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR 2000), pp. 272–279 (2000)
8. De Bra, P., Post, R.D.J.: Information retrieval in the world-wide web: Making client-based searching feasible. Computer Networks and ISDN Systems **27**(2), 183–192 (1994)
9. Epitropou, V., Karatzas, K., Bassoukos, A.: 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, pp. 58–68 (2010)
10. Henderson, T.C., Linton, T.: Raster map image analysis. In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009), pp. 376–380. IEEE Computer Society (2009)
11. Karatzas, K., Moussiopoulos, N.: 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. Lin, H.-T., Lin, C.-J., Weng, R. C.: A note on Platts probabilistic outputs for support vector machines. Machine learning **68**(3), 267–276 (2007)
13. Moumtzidou, A., Vrochidis, S., Chatzilari, E., Kompatsiaris, I.: Discovery of environmental nodes based on heatmap recognition. In: Proceedings of the 20th IEEE International Conference on Image Processing (ICIP 2013) (2013)
14. Moumtzidou, A., Vrochidis, S., Kompatsiaris, I.: Discovery, analysis and retrieval of multimodal environmental information. In: Encyclopedia of Information Science and Technology (in press). IGI Global (2013)
15. Moumtzidou, A., Vrochidis, S., Tonelli, S., Kompatsiaris, I., Pianta, E.: Discovery of environmental nodes in the web. In: Multidisciplinary Information Retrieval, Proceedings of the 5th International Retrieval Facility Conference (IRFC 2012), *LNCS*, vol. 7356, pp. 58–72 (2012)
16. Olston, C., Najork, M.: Web crawling. Foundations and Trends in Information Retrieval **4**(3), 175–246 (2010)
17. Over, P., Awad, G., Kraaij, W., Smeaton, A.F.: TRECVID 2007–overview. In: TRECVID 2007 workshop participants notebook papers. National Institute of Standards and Technology (NIST) (2007)
18. Oyama, S., Kokubo, T., Ishida, T.: Domain-specific web search with keyword spices. IEEE Transactions on Knowledge and Data Engineering **16**(1), 17–27 (2004)
19. Pant, G., Srinivasan, P.: Learning to crawl: Comparing classification schemes. ACM Transactions on Information Systems **23**(4), 430–462 (2005)
20. Pant, G., Srinivasan, P.: Link contexts in classifier-guided topical crawlers. IEEE Transactions on Knowledge and Data Engineering **18**(1), 107–122 (2006)
21. Pant, G., Srinivasan, P., Menczer, F.: Exploration versus exploitation in topic driven crawlers. In: M. Levene, A. Poulovassilis (eds.) Proceedings of the 2nd International Workshop on Web Dynamics, in conjunction with the World Wide Web Conference (WWW 2002) (2002)
22. San José, R., Baklanov, A., Sokhi, R., Karatzas, K., Pérez, J.: Computational air quality modelling. Developments in Integrated Environmental Assessment **3**, 247–267 (2008)
23. Sidiropoulos, P., Vrochidis, S., Kompatsiaris, I.: Content-based binary image retrieval using the adaptive hierarchical density histogram. Pattern Recognition **44**(4), 739 – 750 (2011)
24. Srinivasan, P., Menczer, F., Pant, G.: A general evaluation framework for topical crawlers. Information Retrieval **8**(3), 417–447 (2005). DOI 10.1007/s10791-005-6993-5. URL `http://dx.doi.org/10.1007/s10791-005-6993-5`
25. Tang, T.T., Hawking, D., Craswell, N., Griffiths, K.: 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), pp. 147–154 (2005)

26. Tang, T.T., Hawking, D., Craswell, N., Sankaranarayana, R.S.: Focused crawling in depression portal search: A feasibility study. In: Proceedings of the 9th Australasian Document Computing Symposium (ADCS 2004), pp. 1–9 (2004)
27. Tsikrika, T., Moumtzidou, A., Vrochidis, S., Kompatsiaris, I.: Focussed crawling of environmental web resources: A pilot study on the combination of multimedia evidence. In: Proceedings of the International Workshop on Environmental Multimedia Retrieval (EMR 2014), pp. 61–68 (2014)
28. Yuan, J., etal.: THU and ICRC at TRECVID 2007. In: P. Over, G. Awad, W. Kraaij, A.F. Smeaton (eds.) TRECVID 2007 workshop participants notebook papers. National Institute of Standards and Technology (NIST) (2007)