Landmark	Method			
	Baseline	QE1	QE2	Scene maps
La Pedrera(a)	0.326	0.588	0.377	0.901
Park Guell(a)	0.795	0.794	0.812	0.847
Museu Nat. d' Art	0.590	0.702	0.602	0.637
Columbus Monument	0.505	0.658	0.558	0.698
Carrer B.I.-El Gotic	0.449	0.917	0.555	0.739
Port Vell	0.332	0.746	0.380	0.480
Sagrada Familia	0.857	0.889	0.864	0.881
Casa Batllo	0.759	0.792	0.767	0.798
Arc de Triomf	0.840	0.889	0.847	0.882
La Pedrera(b)	0.651	0.921	0.939	0.903
Hotel Arts	0.560	0.773	0.573	0.633
Hosp. de San Pau(a)	0.317	0.580	0.423	0.838
Hosp. de San Pau(b)	0.421	0.776	0.502	0.709
Park Guell(b)	0.500	0.886	0.526	0.634
Torre Agbar	0.310	0.617	0.378	0.630
Placa de Catalunya	0.794	0.853	0.798	0.812
Cathedral (side)	0.487	0.864	0.546	0.972

Table 6: Mean Average Precision per landmark for the four methods. For each landmark 5 query images were used.

The annotated dataset used contains variable sized groups of images depicting the same scene. Small ones usually correspond to non-landmark scenes while large ones to well known landmarks. Achieving high recall scores is challenging when we deal with a large group of similar images. Re-ranking is only performed on the top ranked images and this can lead to missing quite a few images with the baseline method. Figure 17 shows mAP values for each query based on the size of the group of similar images corresponding to the query image. Observe that scene maps can yield total recall even for scenes containing more than 100 images. For the same scenes, the otherwise powerful QE1 fails to retrieve all the scene instances, since some images were lost from the initial query before the expansion. Furthermore, almost total recall is observed in the small clusters for scene maps, the images of which are usually contained in a very small number of scene maps, usually one or two.

Figures 18 and 19 show a query image, of a non-landmark and a landmark respectively and top ranked retrieved and geometrically verified images. Geometrically verified images are more for scene maps leading to higher recall. Tables 6 and 7 contain mAP values for each group of the landmarks and non-landmarks ground truth respectively. Remarkable is the fact that for many groups scene map achieved perfect mAP equal to 1.0 while other methods achieved a worse ranking of the similar images.

Location recognition evaluation. All *European Cities 1M* dataset images are geo-tagged. Thus, given the outcome of visual retrieval, location recognition is performed (see D2.3). To evaluate the proposed scheme, we compare the resulting estimation against the hand-picked geographic location information of each annotated group of images in our *European Cities 1M*. Localization accuracy in comparison to baseline and other methods is shown in Table 8. As we see, localization percentage is already high even for the baseline method. Still, our method using scene maps reaches the highest percentage.

Scene	Method			
	Baseline	QE1	QE2	Scene maps
Scene1	0.618	0.648	0.654	0.884
Scene2	0.667	0.847	0.730	1.000
Scene3	0.399	0.458	0.451	0.880
Scene4	1.000	1.000	1.000	1.000
Scene5	1.000	1.000	1.000	1.000
Scene6	0.800	0.969	0.848	0.802
Scene7	0.876	0.979	0.940	1.000
Scene8	1.000	1.000	1.000	1.000
Scene9	0.339	0.557	0.357	0.754
Scene10	0.351	0.482	0.428	0.687
Scene11	0.557	0.843	0.575	0.633
Scene12	0.577	0.857	0.639	0.755
Scene13	0.681	0.846	0.746	1.000
Scene14	0.875	1.000	0.880	0.885
Scene15	1.000	1.000	1.000	1.000
Scene16	0.791	0.883	0.798	0.812
Scene17	1.000	1.000	1.000	1.000
Scene18	0.800	0.972	0.810	1.000

Table 7: Mean Average Precision per scene for the four methods. For each scene 5 query images were used or less if the total group size is below 5.

Table 8: Percentage of correctly localized queries within at most 150 meters from the ground truth location.

Method	Distance threshold		
	$< 50m$	$< 100m$	$< 150m$
Baseline BoW	82.5%	91.6%	94.2%
QE1	86.3%	93.5%	96.2%
QE2	86.7%	93.3%	96.5%
Scene maps	87.8%	94.2%	97.1%

Samples of query images depicting well known landmarks and the corresponding localization result on the map are presented in Figure 20. The first 6 cases achieve successful recognition. However in the last two cases we present two examples, coming from the evaluation queries, of unsuccessful recognition based on the ground truth geo-tag which is the exact location of the landmark. Final estimation is far from the ground truth location. This is derived from the fact that geo-tags of user images correspond to the location where the photo was taken from. Thus, these are unsuccessful examples of localizing the landmark but successful ones of localizing the photo.

Landmark recognition evaluation. Since most photographers are taking pictures of well known landmarks, we can safely assume that some of the annotated groups of images in our European Cities 1M dataset can be linked with Wikipedia articles. Given that the metadata of the images in our European Cities 1M dataset contain user tags, we use the method proposed in D2.3 to analyze them and effectively identify the landmark and suggest Wikipedia articles for each query.

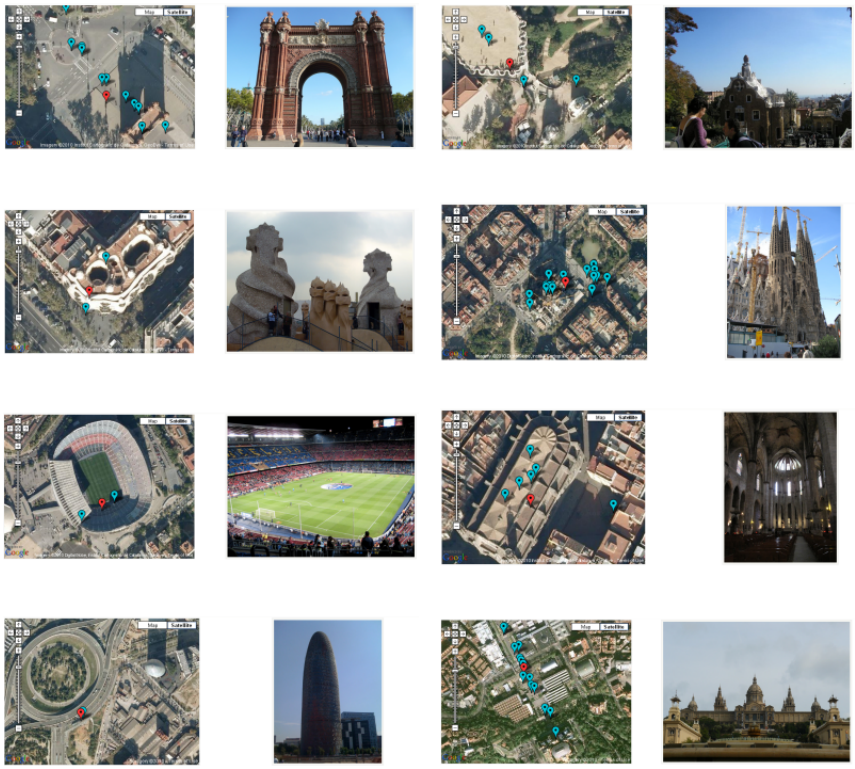

Figure 20: Samples of query images and location recognition result on the map. For each pair there is the map on the left and the initial query image on the right. Blue marker: Retrieved image. Red marker: Geo-tag estimation. Grey marker: Discarded after RNN clustering.

The performance of the approach is shown in Table 9, where we see the percentage of correctly discovered links. Experiments are carried on 17 of the groups, that is the dataset subset which depict landmarks and has corresponding Wikipedia articles. We regard a query link suggestion as correct, if the ground truth article link is one of those suggested from the landmark recognition process. As the table shows, recognition for landmark queries is really efficient both with the use of scene maps and query expansion. Samples of query images and the corresponding suggested and frequent tags are presented in Figure 21. These are examples of successful landmark recognition.

4.3 Travel photo sequences

Given a set of user photos we localize them on the map and form a travel route approximating the one followed when these photos were taken. We also associate parts of the route with landmarks depicted and link them to Wikipedia articles.

ViRaL for location and landmark estimation. Let $P = \{p_i\}$, $i = 1 \ldots N$ be the sequence of N user images. Initially we use each individual image p_i of the sequence as a query to the ViRaL tool. This process is described in D2.1.2. As a consequence, a set of localized images $L = \{p_i : \ell(p_i) = 1\}$, $L \subset P$ is formed, where $\ell(p_i)$ is one if image p_i is successfully localized on the map by the ViRaL tool, otherwise it is zero. All remaining images $\hat{L} = P \backslash L$, are not successfully localized yet. A sample of user images taken during a day in London are shown in Figure 22. After using the ViRaL tool, we get the initial subset of localized images on the map

Table 9: Percentage of correct Wikipedia article suggestions for each landmark and average percentage for the four methods.

Landmark	Method			
	Baseline	QE1	QE2	Scene maps
La Pedrera(a)	100%	100%	100%	100%
Park Guell(a)	100%	100%	100%	100%
Museu Nat. d' Art	40%	100%	60%	80%
Columbus Monument	100%	100%	100%	100%
Carrer del Bisbe Irurit-El Gotic	100%	100%	100%	100%
Port Vell	80%	100%	80%	100%
Sagrada Familia(b)	100%	100%	100%	100%
Casa Batllo	100%	100%	100%	100%
Arc de Triomf	100%	100%	100%	100%
La Pedrera(b)	60%	100%	80%	80%
Hotel Arts	40%	40%	40%	60%
Hospital de Sant Pau(a)	100%	100%	100%	100%
Hospital de Sant Pau(b)	80%	80%	80%	100%
Park Guell(b)	100%	100%	100%	100%
Torre Agbar	100%	100%	100%	100%
Placa de Catalunya	100%	100%	100%	100%
Cathedral (side)	80%	80%	80%	80%
Average	**87%**	**95%**	**90%**	**95%**

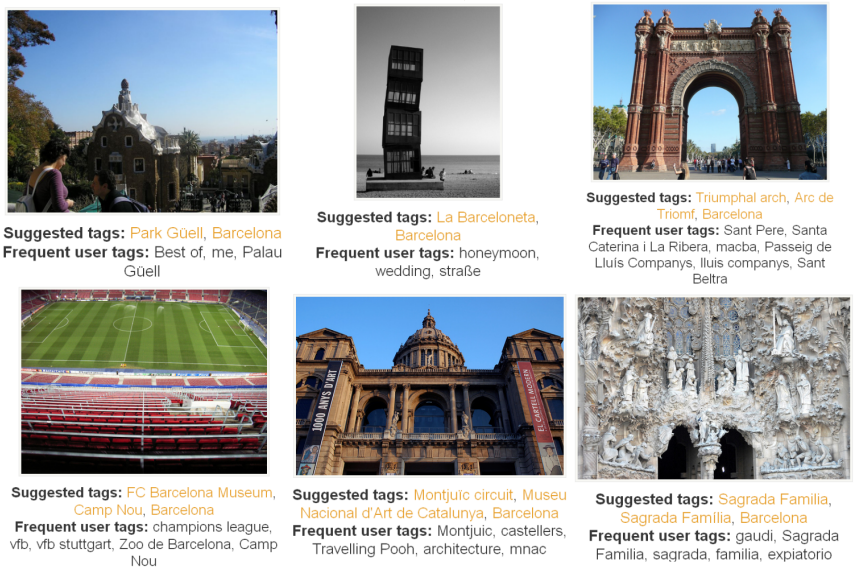

Figure 21: Samples of query images with suggested and frequent tags. Landmarks are recognized successfully and the corresponding Wikipedia links are provided.

as shown in Figure 23.

Figure 22: Sample user images taken during a single day tour around London.

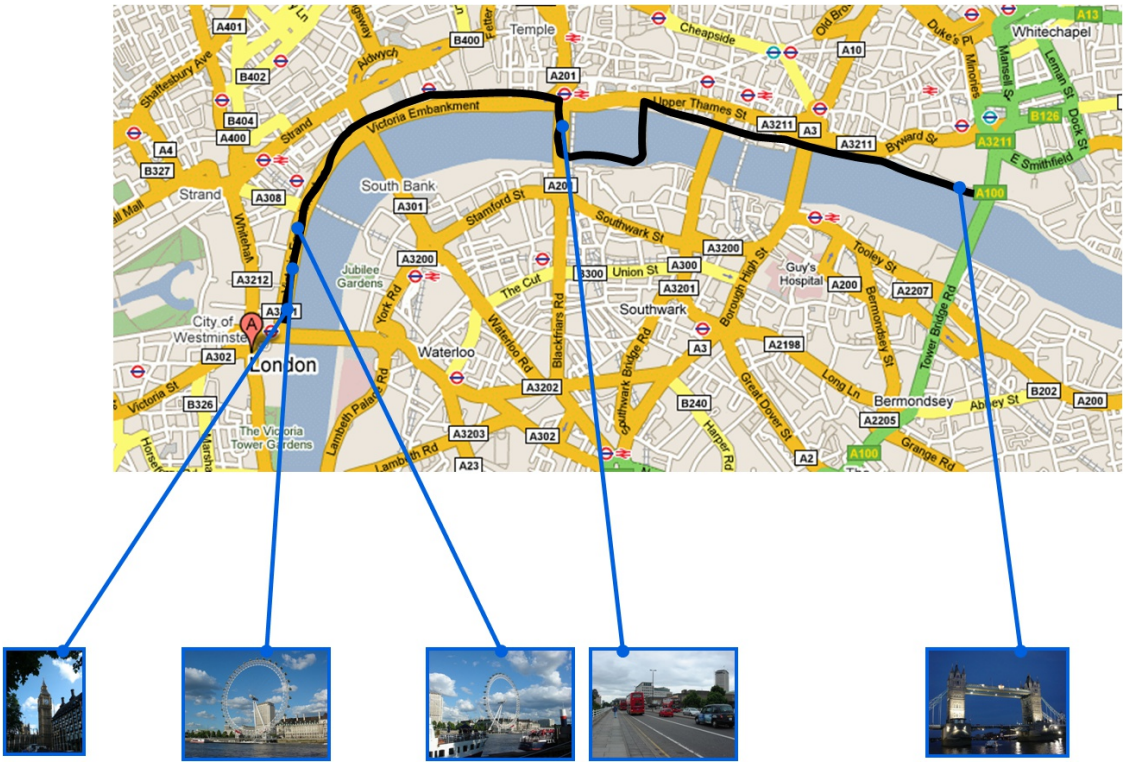

Figure 23: Initial set of images localized on the map after issuing queries to the ViRaL tool.

From visual words back to descriptors. The ViRaL tool integrates methods for large scale image retrieval. This is done in a fast way, and in a some cases similar images might be missed in favor of speed. From the image set depicted in Figure 23, only a few are initially localized with the VIRaL tool, although some of the images left without an estimated location do indeed contain a landmark. Since this image set is relatively small, these images could be identified with a more detailed method, that is now computationally feasible. We geometrically match all images from $\hat{L}$ with each image successfully localized. If an image $p_i \in \hat{L}$ is geometrically verified against an image $p_j \in L$, then p_i is also localized in the same position. We set $g(p_j) = g(p_i) = [lat_i, lon_i]$ and $\ell(p_j) = 1$, where $g(p_i)$ are the geo-coordinates (latitude: lat_i, longitude: lon_i) assigned to p_i. Function g is only defined for images localized successfully ($\ell(p_i) = 1$).

The visual word representation allows images to be organized in an inverted file structure and also tentative correspondences to be found really fast. This is used in the ViRaL tool for speedup and sub-linear access to a large database of images. Now, we go back to the descriptors of local

features and use them to find tentative correspondences. We use SURF local feature and descriptors [7], which are already extracted while issuing queries to ViRaL. We find initial assignments between images with approximate nearest neighbor search using FLANN [57]. We also employ the ratio-test used in [53] in order to keep assignments forming our tentative correspondences. Finally geometric matching is performed with Fast Spatial Matching [64], as in the ViRaL tool.

Using time from EXIF data. We now exploit EXIF data, and particularly time at which each photo was taken. We utilize this information in order to define a possible position on the map for all images that are still not localized. If $p_i,\ p_j \in L$ are two consecutive images in time (excluding images in $\hat{L}$), then we define a part of the travel route v_{ij}, as a path between $g(p_i)$ and $g(p_j)$. All images $p_k \in \hat{L}$ for which $t(p_i) < t(p_k) < t(p_j)$ are placed on the path v_{ij}. In this way the total travel route is constructed by the union of all those paths (Figure 24).

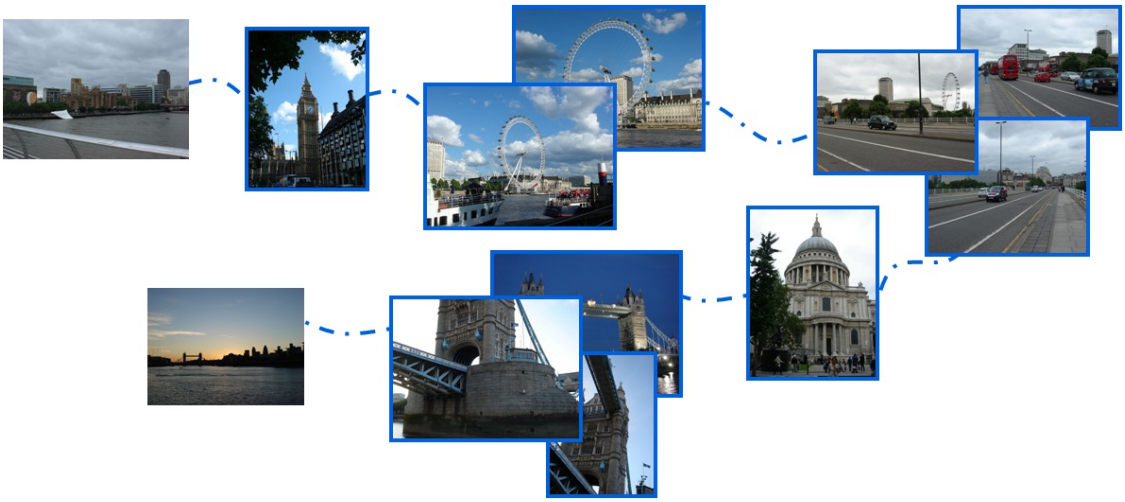

Figure 24: Virtual route after initial localization, pairwise matching and grouping of verified images and localization of all other images using time from EXIF data.

Identifying landmarks. When an image is localized with the ViRaL tool, landmarks might also be identified. Those images and all other verified with them are related with the specific landmark. Finally, a route through well know landmarks is also formed, as show in Figure 25.

4.4 Point-of-interest exploration

The first part of our recently proposed scene maps technique[5][40] provides a clustering of the images given an (arbitrarily large) geo-tagged image collection. This *view clustering* proposes a summarization of the collection and can evidently guide an interface for visualization or exploration. In this *point-of-interest exploration* task we start from view clusters and form an exploration interface that uses Google Maps API to pin the clusters on the map and also displays information regarding a selected cluster. The information displayed are the images contained in the cluster and, if the image depicts a landmark, the landmark name and a link to the corresponding Wikipedia article. In order to make the exploration interface appealing to the user, we exploited both the importance of each individual cluster and spatial distribution of the clusters on the map to select the visible clusters per zoom level.

Following the technique described in [5] and [40], we form view clusters using geographical and visual clustering on a large image corpus. Each of the view clusters has by default a *reference*

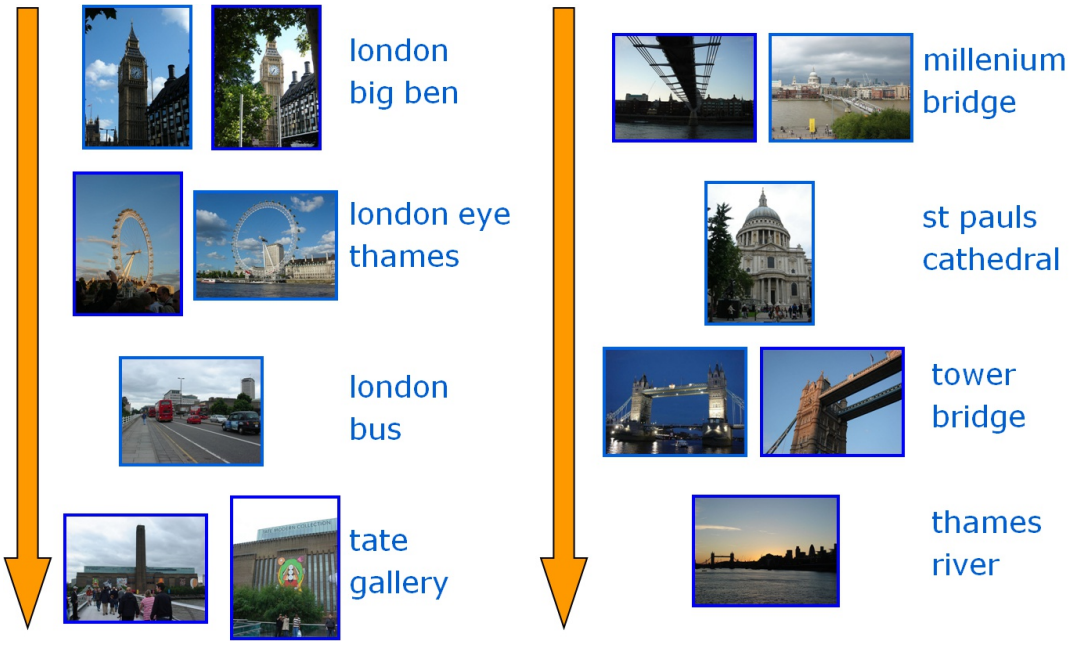

Figure 25: User images associated with well known landmarks.

image, i.e. an image that all other cluster images share visual information with. What is more, given the cluster images, we process the geographical and textual metadata available and extract both a robust estimation of the cluster's geographical location and also the name of any landmark or point-of-interest that might be depicted in the cluster images. In the exploration interface, each cluster is represented on the map by the reference image, pinned at the estimated location of the cluster.

To pick the best clusters to be displayed at each zoom level, we exploit the size and spatial distribution of the clusters. The size of each cluster, i.e. the number of images that it contains, can be seen as a measure of importance. Spatial distribution is also important for visualization, since we prefer the map to be covered with clusters to explore at all visible locations.

For each zoom level a greedy algorithms selects the most important images to cover the map, given a radius parameter. As the zoom levels grow, and the visible area of the map decreases, the radius parameter also decreases, enabling more and more clusters to be displayed. At the final zoom level we set no restrictions and present any available cluster in the visible area.

The exploration interface is shown in Figure 26. Most area is covered by the map and there is some space on the right, reserved to display information of a selected cluster. The user can select a cluster by clicking on its reference image. This will trigger the cluster information to appear on the right frame, i.e. the cluster images and the landmark(s) present, if available. In Figure 26 the selected cluster is the one marked with a red border on the map.

Figure 26: The point-of-interest exploration interface. On the left frame there is a fully functional Google Map with image clusters depicted by their reference image, and on the right the user can see the images and landmark of the cluster that has a red border on the map.

5 Evaluation of retrieval techniques on long-tail POIs

This section discusses the problem of matching the photos of a tagged photo collection to a list of "long-tail" Points Of Interest (POIs), that is POIs that are not very popular and thus not well represented in the photo collection. Despite the significance of improving "long-tail" POI photo retrieval for travel applications, most landmark detection methods to date have been evaluated on very popular landmarks. In this section, we present an empirical analysis comparing four baseline matching methods that rely on photo metadata, three variants of an approach that uses cluster analysis in order to discover POI-related photo clusters, and a real-world retrieval mechanism (Flickr search) on a set of less popular POIs. The results of this study were published in [82].

A user-based evaluation of the aforementioned methods is conducted on a Flickr photo collection of over 100,000 photos from 10 well-known touristic destinations in Greece. A set of 104 "long-tail" POIs is collected for these destinations from Wikipedia, Wikimapia and OpenStreetMap. The results demonstrate that two of the baseline methods outperform Flickr search in terms of precision and F-measure, whereas two of the cluster-based methods outperform it in terms of recall and POI coverage. The results of this study are valuable for enhancing the indexing of pictorial content and the user experience in the CSG use case.

5.1 Motivation

The massive amounts of user contributed content in social media sites has provided valuable input for a series of mining applications and for numerous intelligent services built on top of the mined knowledge. An important problem that has recently attracted considerable interest is the detection of Points of Interest (POIs) in large user-contributed photo collections [42, 69, 23]. It has been demonstrated that photos of popular landmarks around the world can be success-

fully mined in large sets of tagged photos. In addition, different representative photos for each detected landmark can be identified [42, 70] leading to more diverse pictorial descriptions for landmarks. The results of automatic POI detection are valuable in the context of tourist applications both for presenting the interesting attractions of a place to the potential visitor [70, 23] as well as for planning efficient tourist itineraries based on the available POIs [17, 37, 67].

Despite the great interest in this problem, most existing works are limited to the discovery of prominent landmarks that are well represented in the photo collections. Less known POIs are usually disregarded due to the fact that most existing methods require a large number of photos per POI in order to reliably detect it. The objective of this task is to evaluate the effectiveness of different schemes for identifying photos that depict "long-tail" POIs. Starting from three geographical sources of information, namely Wikipedia[11], Wikimapia[12], and OpenStreetMap[13], we compile lists of "long-tail" POIs. Then, we devise several matching functions in order to associate individual photos with POIs by use of textual (title, description, tags) and geo-location metadata. In addition, we leverage the collective intelligence (WP2-WP3) hybrid photo clustering service of WeKnowIt [62, 61] to discover photo clusters related to the target POIs. Three variants of the method are tested that rely on different types of photo similarity graphs (visual, tag-based, and hybrid). Finally, we make use of the Flickr search service by posting queries constructed from the POI name and place. The latter implementation is a widely used real-world retrieval mechanism, and hence, constitutes a suitable and high performance competitor. Our empirical study demonstrates that two of our baseline matching schemes outperform Flickr search in terms of F-measure, while the graph-based methods perform best in terms of recall and coverage, thus being more suitable for the problem of "long-tail" POIs, where conventional matching schemes may yield no matches.

The rest of this section is structured as follows. Subsection 5.2 briefly discusses several related works. In subsection 5.3, we present background information on the problem, namely the necessary notation, the POI list compilation process, and the employed graph-based photo clustering method. Subsection 5.4 presents the proposed matching methods for addressing the problem. Subsection 5.5 describes the experimental setup and the obtained results.

5.2 Related Work

The application of social media mining in tourist scenarios has recently attracted significant interest due to its potential for the automatic production of high-quality tourist-related multimedia content. In particular, there is a wealth of research activity in the area of landmark recognition in large tagged photo collections. For instance, Kennedy and Naaman [42] make use of tags, location information and visual features of photos in order to identify clusters of photos that correspond to different views of popular landmarks. However, they make no use of external knowledge sources (e.g. Wikipedia) and they rely on a large number of geotagged and representative photos for each landmark. Thus, their approach is limited to landmarks that are well covered (in terms of pictorial content and metadata) within a photo collection. Similar limitations hold for the work in [70] that relies on the "interestingness" property of photos (provided by Flickr), which limits its utility to highly voted photos.

[11] Wikipedia, http://www.wikipedia.org/
[12] Wikimapia, http://wikimapia.org/
[13] OpenStreetMap, http://www.openstreetmap.org/

Quack et al. [69] mine landmarks and events from a large set of photos by clustering them based on their visual and textual similarity, classifying the photo clusters into landmarks or events and mapping the clusters to Wikipedia articles by use of query formulation and visual matching. Despite the high precision reported by the authors, their method also suffers from low recall, i.e. it does not detect places with few photos in Flickr.

Other travel-oriented applications of social media mining are presented in [16, 17, 37, 67]. Crandall et al. [16] attempt to estimate the geographic position of tagged photos by use of visual and textual features, which is complementary to our work, since the more geotagged pictures are available the better the performance of the POI detection will be. The works in [17, 37, 67] deal with the automatic travel itinerary creation from tagged photo collections of cities. Such applications can greatly benefit from POI detection, since they rely on POIs to create itineraries passing through them. The work in [67] makes use of Wikipedia articles and categories in order to identify POI names and locations.

Within WeKnowIt, work has been conducted to mine structured POI information from Wikipedia articles [38] and clusters of photos corresponding to POIs [71]. The first (Wikipedia POI extraction) is used as one of the POI sources for this experiment (OpenStreetmap and Wikimapia being the other two), and the latter is used as the basis of the cluster-based matching schemes benchmarked in this task.

5.3 Background - Notation

5.3.1 Notation

Table 10 contains all necessary notation used throughout the section. Our starting point is the collection (set) $\mathbb{R} = \{r\}$ of tagged photos, where each photo r is a tuple $(t_r, d_r, X_r, \lambda_r)$, comprising a title t_r, a description d_r, a set of tags X_r and the capture location of the photo λ_r, expressed as a latitude-longitude pair of values. Furthermore, we consider the set $\mathbb{P} = \{p\}$ of POIs, where each POI p is a tuple (pl_p, T_p, λ_p) containing the place of the POI pl_p, a set T_p of alternative titles for the POI and the location λ_p of the POI. In this study, a place may refer to a city (e.g. Herakleion) or an island (e.g. Santorini). The problem we address is the evaluation of different mappings $f : \mathbb{R} \rightarrow \{\mathbb{P}, nil\}$, which map each photo of the collection to one of the available POIs or to no POI at all. In the case of cluster-based POI-photo matching (described in subsection 5.4.2), we achieve the POI-photo mapping f through an intermediate mapping $f : CL_{type} \rightarrow \{\mathbb{P}, nil\}$, which associates each photo cluster of the cluster set CL_{type} to one of the available POIs or to no POI at all. The photo clusters are derived based on the graph-based photo clustering method described in subsection 5.3.3. Three types of clusterings are considered depending on the underlying image similarity graph; more specifically $type \in \{VIS, TAG, HYB\}$ corresponding to visual, tag-based, and hybrid similarity graph respectively. A cluster $c \in CL_{type}$ is a tuple (M_c, RT_c), where M_c is a set of cluster members (i.e. $M_c \subset \mathbb{R}$) and RT_c is the set of representative titles for the given cluster, derived from the process described in subsection 5.4.2.

5.3.2 POI list creation

We use three sources for compiling the list of POIs: (a) Wikipedia, (b) Wikimapia and (c) OpenStreetMap. The advantage of Wikipedia is that it contains rich additional information

Table 10: Notation used in this section

Symbol	Definition
$\mathbb{R} = \{r\}$	Collection of photos
$\mathbb{P} = \{p\}$	Set of POIs
$CL_{type} = \{c\}$	Set of clusters
$type$	*VIS, TAG or HYB*
$r = (t_r, d_r, X_r, \lambda_r)$	Photo metadata
t_r	title
d_r	description
X_r	tags
λ_r	geolocation
$p = (pl_p, T_p, \lambda_p)$	POI information
pl_p	place
T_p	representative titles
λ_p	geolocation
$c = (M_c, RT_c)$	Cluster metadata
M_c	members $M_c \subset \mathbb{R}$
RT_c	titles $RT_c = \{rt_{c_1}, rt_{c_2}, \ldots\}$
$tok(s)$	set of tokens for string s
$place(p)$	alternate names for place of p

for each POI contained in it. However, it misses several less important POIs. Wikimapia and OpenStreetMap, on the other hand, contain only basic information for each POI, but they have much higher coverage of POIs, especially in smaller places and cities. We populate our final list having in mind that we are more interested in "long-tail" POIs than popular ones. Starting the selection from a list of touristic destinations, we end up with a number of POIs ready to be used by the POI-photo matching methods.

5.3.3 Graph-based clustering

The cluster-based POI-photo matching method of subsection 5.4.2 relies on the creation of three types of photo similarity graphs representing three kinds of similarities between photos of the collection, namely visual, tag and hybrid. In the visual graph, the edge weights represent the pairwise similarities in terms of visual content (SIFT descriptors [52] are extracted for each photo and a bag-of-visual-worlds feature vector is computed based on the software implementation of [79]). On the other hand, the tag graph is built by use of tag co-occurrences between photos. Each edge on this graph is weighted by the number of tags shared between the two photos. Both popular tags and weak edges are discarded to increase noise resilience and reduce the computational needs of the clustering algorithm. The hybrid graph is formed by considering the union of the visual and tag graph. Then, a community detection procedure is applied on each graph with the goal of identifying sets of nodes (i.e. photo clusters) that are more densely connected to each other than to the rest of the network. This graph-based clustering framework is described in detail in [62, 61] and exposed as a WeKnowIt collective intelligence service.

5.4 POI-Photo Matching

5.4.1 Baseline matching

In order to match Flickr photos with the extracted POIs, we make use of the following metadata: (a) title, (b) tags, (c) description, and (d) capture position. Due to the unrestricted nature of photo sharing applications, there are numerous cases, in which one or more of the aforementioned metadata fields are missing. In such cases, it is not possible to match the photo to any POI by use of the baseline methods proposed here. We consider matching functions of the form $s : \mathbb{R} \times \mathbb{P} \rightarrow [0,1]$ for estimating how well a particular photo r matches a given POI p. In particular, the matching functions can be expressed as:

$$s_k(r,p) = \begin{cases} A_k & \text{if matching criterion } \phi_k \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

where ϕ_T, ϕ_D, ϕ_L and ϕ_H are the criteria for title-tag, description, location and hybrid matchings respectively.

In order to simplify the function specification, we also define the tokenization function $tok(s)$, which, given a string s, produces a set of tokens (strings) by splitting s around matches of white spaces. We further define the function $place(p)$, which, given a POI, returns the name pl_p of the place containing the POI, and some alternate names for it, e.g. $place(\text{“Acropolis”}) = \{\text{“Athens”}, \text{“Athina”}\}$. Finally, we slightly abuse the subset operator to denote string containment, i.e. $a \subseteq b$ denotes that string a is contained in string b. Depending on the kind of metadata we rely on, we obtain a different matching criterion ϕ and a different score A. In the end, we get the following matching function variants:

- Title+tag token match (s_T):
  $A_T = \alpha$, $\phi_T \equiv \{\exists t \in T_p :$
  $(tok(t) - place(p)) \subseteq (tok(t_r) \cup X_r)\}$
- Location match (s_L): $A_L = \beta$, $\phi_L \equiv \{d(\lambda_r, \lambda_p) < l\}$ where $d(\lambda_r, \lambda_p)$ is the geodesic distance between λ_r and λ_p, and l is a predefined threshold.
- Description match (s_D):
  $A_D = \gamma$, $\phi_D \equiv \{\exists t \in T_p : t \subseteq d_r\}$
- Hybrid match (s_H): $A_H = 1$, $\phi_H = \phi_T \cap \phi_L$

The positive scores returned by the aforementioned variants obey the constraints $\gamma < \beta < \alpha < 1$ in order to reflect the confidence we have in each criterion. In the end, for each criterion k, given POI p, we obtain the set $POI_k(p) = r : s_k(r,p) > 0$ of photos matched with p by use of matching criterion k.

5.4.2 Cluster-based matching

Here, we attempt to match a POI p with a cluster c by using the POI and cluster titles, T_p and RT_c respectively. The latter is a result of a process that finds the most frequent word sequences within the titles of photos of the cluster c. More specifically, we define the function $seq(t_r)$ that, for each title t_r of a photo, returns the set of all possible term sequences up to length 6. We then aggregate over all term sequence sets for the photos of the cluster maintaining a count for each

one of them. In the end, we select the top five term sequences as the representative titles RT_c of the cluster, by ranking them with a function that takes into account the term sequence frequency, the sequence length (we prefer longer to shorter titles), and the lexical diversity of titles (we prefer to have diverse titles in order to capture alternative names).

After extracting RT_c, we apply the cluster-based matching procedure that is implemented as a cascade of three consecutive criteria, namely full title match (ϕ_F), relaxed title match (ϕ_{Rel}), and partial lexical match (ϕ_{Par}). We consider matching functions of the form $s_{CL} : \mathbb{CL} \times \mathbb{P} \rightarrow [0,1]$ for estimating how well a particular cluster c matches a given POI p. The matching s_{CL} is expressed by Equation 5.2.

$$s_{CL}(c,p) = \begin{cases} \mu & \text{if matching criterion } \phi_F \\ \nu & \text{if matching criterion } \phi_{Rel} \\ \xi & \text{if matching criterion } \phi_{Par} \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

where ϕ_F, ϕ_{Rel}, and ϕ_{Par} are defined as:

- $\psi_F \equiv \{\exists t \in T_p : t \subseteq RT_c\}$
- $\psi_{Rel} \equiv \{\exists t \in T_p : (tok(t) - place(p)) \subseteq tok(RT_c)\}$
- $\psi_{Par} \equiv \{\exists t_1 \in T_p, \exists t_2 \in RT_c :$
  $|\{(pt_i, ct_j) | LV(pt_i, ct_j) > \sigma_1\}| > \sigma_2 * |PT|\}$

  where $LV(s_1,s_2) \in [0,1]$ denotes the Levenshtein similarity [47] between the strings s_1,s_2, $pt_i \in PT = \{(tok(t_1) - place(p))\}$ and $ct_i \in \{(tok(t_2) - place(p))\}$ and $\sigma_1 = 0.6, \sigma_2 = 0.5$.

The full title match criterion (ϕ_F) searches for identical titles, while the relaxed criterion (ϕ_{Rel}) tries to match all tokens of a POI title, regardless of their sequence, with tokens from a cluster title taking into account the function $place(p)$. The third criterion (ϕ_{Par}) tries to match titles, of which a certain number of tokens is lexically quite similar. With the partial lexical match we are trying to solve both the problems of misspelling and the identification of slightly different written words with the same meaning (e.g. "Saint Anna" or "St. Anna"). We set the values of the constants in Equation 5.2, such that the following condition holds: $\mu > \nu > \xi > \sigma_2$. In the end, given POI p, we obtain the set $POI_{CL}(p) = \bigcup M_c, c : s_{CL}(c,p) > \sigma_2$ of photos matched with p by use of cluster-based matching.

5.4.3 Flickr search matching

For a POI p, we form a query q_p by concatenating the POI name (most characteristic name from the set T_p) and the place name pl_p (in order to avoid ambiguity for landmarks with the same name in different places). We post the query to the Flickr search API (flickr.photos.search) and then obtain the set $POI_{FL}(p) = FLICKR(q_p)$ of the photos matched with p by use of Flickr search matching.

5.5 Experiments

5.5.1 Dataset

Starting from 10 touristic places of Greece [14], we compiled a list of 104 "long-tail" POIs contained in them. Furthermore, we downloaded a number of place-focused photos (129.023 in total, 46.181 geotagged) by issuing appropriate queries for each place to the Flickr search service (e.g. "Chania Greece" to retrieve photos from "Chania"). This collection will be referred to as the "original" photo collection and be denoted as $\mathbb{R}_0$. The use of the Flickr search matching method (FL) on the list of POIs resulted in an extension of the "original" dataset, since Flickr search is conducted over the whole Flickr collection. Adding these additional photos to our "original" collection resulted in a total of $\backsim$ 148.000, which will be referred to as the "extended" photo collection, and denoted as $\mathbb{R}_{ext}$.

The application of the proposed matching methods took place separately for each place, which means that POIs of a place were matched with photos (or clusters) of the same place. For instance, for each POI of Santorini, for which 20,086 photos were downloaded from Flickr, the candidate photos for matching were limited to the set of 20,086 photos, or to the 187 photo clusters in the case of the hybrid graph cluster-based matching (CL_{HYB}).

5.5.2 User study

We conducted a study involving 10 users. Each user was assigned approximately 20 POIs, and for each POI, she was presented with the union of the results of all matching methods for that POI, the so-called "POI photo pool". The user was instructed to decide for each photo whether it was relevant or irrelevant to the POI in question. If the user could not decide, they were given the option of not filling in the annotation for the photo (*don't know*). In order to help the users make the decision, for each POI we provided a link to a site containing a characteristic photo and some description of the POI. In total, a set of $\backsim$ 34.000 photos (the union of all POI photo pools for the 104 POIs of the study) were evaluated. Each photo was evaluated by two users in order to be able to estimate inter-annotator agreement (κ-statistic). Although the users were oblivious to the matching method that produced the photos for a given POI, this association was maintained in the back-end, thus enabling us to compute precision (P), recall (R), and F-measure (F) for each one of the methods. Furthermore, since methods could not identify relevant photos for each of the 104 POIs, we computed the POI coverage (C) for each method, i.e. the percentage of POIs, for which at least one relevant photo could be retrieved by the method.

5.5.3 Results and discussion

Depending on whether we considered the "don't know" response as relevant or irrelevant, we computed two different retrieval performance sets, which we will refer to as *relaxed* ("don't know" counts as relevant) and *strict* ("don't know" counts as irrelevant). In addition, depending on the photo set used to compute the retrieval performance scores, we obtained two different results, one computed with reference to $\mathbb{R}_0$ and the other with reference to $\mathbb{R}_{ext}$. The combination of the aforementioned variants of computing performance resulted in four different result sets,

[14] Chania, Corfu, Heraklion, Ioannina, Nafplion, Naxos, Paros, Santorini, Thessaloniki, Zakynthos

(a) Red Beach (Santorini)

(b) Nea Kameni (Santorini)

(c) Perama Caves (Ioannina)

(d) Patsides (Herakleion)

Figure 27: Examples of long tail POIs.

namely relaxed in $\mathbb{R}_0$ (Table 11), strict in $\mathbb{R}_0$ (Table 12), relaxed in $\mathbb{R}_{ext}$ (Table 13) and, strict in $\mathbb{R}_{ext}$ (Table 14).

Tables 11 and 12 present valid comparative results, since all matching methods are evaluated with reference to the same photo collection. One can see that the highest precision was achieved by use of the hybrid (H) matching method (75.3% in relaxed, and 68% in strict mode). Inter-annotator agreement was also high (0.498). However, this method gave the lowest recall values due to the added criterion of location in the matching procedure. In our dataset, most of the photos do not include geographic information, thus, this method is not applicable for them.

Considering recall values, we see that the tag cluster-based method CL_{TAG} achieved the best score (51.5% in relaxed, 51.4% in strict mode), which indicates that tag-based photo clustering can be beneficial for extending the POI search results. However, this comes at a cost in precision, since the CL_{TAG} method presents the worst performance in terms of precision. The highest F-measure was achieved by the title+tag (T) method (55.3% in relaxed, and 53.3% in strict mode) making it the best candidate for balanced results in terms of precision and recall. An additional noteworthy observation pertains to the very low κ-statistic scores achieved by methods (with the exception of the highly selective Hybrid (H) matching method). This indicates that associating photos with POIs is a hard task even for human annotators. For instance, some users may find relevant only the characteristic views of a POI while others may also find relevant some alternative views of it (e.g. indoor views). In addition, some POIs are intrinsically difficult to evalutate (e.g. neighborhoods, beaches, old cities).

Table 11: IR performance results (relaxed, $\mathbb{R}_0$)

Method	P	R	F	κ	C
T	0.680	0.466	**0.553**	0.252	0.894
D	0.613	0.176	0.273	0.160	0.760
L	0.533	0.271	0.359	-0.478	0.490
H	**0.753**	0.069	0.126	**0.498**	0.270
CL_{HYB}	0.525	0.429	0.472	-0.464	**0.952**
CL_{TAG}	0.518	**0.515**	0.516	-0.419	**0.952**
CL_{VIS}	0.626	0.107	0.183	-0.020	0.450
FL	0.677	0.437	0.531	0.263	0.800

Table 12: IR performance results (strict, $\mathbb{R}_0$)

Method	P	R	F	κ	C
T	0.591	0.485	**0.533**	0.027	0.894
D	0.534	0.189	0.279	-0.176	0.760
L	0.439	0.262	0.328	-2.003	0.490
H	**0.676**	0.083	0.147	**0.400**	0.270
CL_{HYB}	0.442	0.435	0.438	-0.942	**0.952**
CL_{TAG}	0.432	**0.514**	0.470	-0.710	**0.952**
CL_{VIS}	0.553	0.112	0.186	-0.493	0.450
FL	0.600	0.458	0.520	-0.016	0.800

Finally, the best POI coverage was achieved by two of the cluster-based methods, CL_{HYB} and CL_{TAG}, since they managed to find relevant photos for 99 out of the 104 POIs of the list. Figure 27 presents four examples of POIs, for which Flickr could not return any relevant photo, while the proposed methods could match several related ones. We also computed for each method the number of POIs, for which the method returned the maximum number of relevant photos. In the case of the "strict" annotation mode, the title+tag matching (T) returned the most relevant photos for 34 POIs, description (D) for 1, location (L) for 9, hybrid graph clustering (CL_{HYB}) for 12, tag graph clustering (CL_{TAG}) for 26, and Flickr search matching (FL) for 22.

Although Tables 13 and 14 are not valid for comparison, they can be used to draw the following interesting conclusion: even when the Flickr search matching uses the whole Flickr dataset, the performance of the method does not improve in terms of coverage. This marks the importance of cluster-based matching methods in enriching "long-tail" POIs with photos, and thus in improving the search experience in the CSG use case.

Table 13: IR performance results (relaxed, $\mathbb{R}_{ext}$)

Method	P	R	F	κ	C
T	0.680	0.345	0.458	0.252	0.894
D	0.613	0.128	0.212	0.160	0.760
L	0.533	0.214	0.305	-0.478	0.490
H	**0.753**	0.044	0.083	**0.498**	0.270
CL_{HYB}	0.525	0.349	0.420	-0.464	**0.952**
CL_{TAG}	0.518	0.410	0.458	-0.419	**0.952**
CL_{VIS}	0.626	0.084	0.147	-0.020	0.450
FL	0.651	**0.526**	**0.582**	0.201	0.800

Table 14: IR performance results (strict, $\mathbb{R}_{ext}$)

Method	P	R	F	κ	C
T	0.591	0.358	0.446	0.027	0.894
D	0.534	0.140	0.222	-0.176	0.760
L	0.439	0.203	0.277	-2.003	0.490
H	**0.676**	0.051	0.094	**0.400**	0.270
CL_{HYB}	0.442	0.353	0.392	-0.942	**0.952**
CL_{TAG}	0.432	0.407	0.419	-0.710	**0.952**
CL_{VIS}	0.553	0.087	0.151	-0.493	0.450
FL	0.573	**0.544**	**0.558**	-0.071	0.800

6 Evaluation of textual methodologies

6.1 Introduction

Throughout the WeKnowIt project the development of the Text Analysis techniques, tool and services have attempted to address the requirements of real world collective intelligence applications as represented by the WP7 use cases. As was discussed in previous deliverables this principally involves considering the nature of collective intelligence (textual) data, which tends to be short and informal. In addition, due to its conversational nature, information within a single user generated document may require context outside of that document to enable its comprehension.

The following section evaluates the Text Analysis tools on a real-world dataset, showing the effectiveness of applying such automatic analysis techniques in comparison with human annotation. The evaluation data also offers an additional difficulty as whilst the text is generally in English it also contains a mixture of French and Haitian Creole colloquial terms. The results indicate that the techniques are effective even on such text.

6.2 Haiti Data

The Ushahidi data from the Haiti humanitarian crisis in 2010 provides comprehensive information which has been gathered in near real time from reports coming from inside Haiti via: SMS, Web, Email, Radio, Phone, Twitter, Facebook, Television, List-serves, Live streams and Situation Reports. Volunteers at Ushahidi's Situation Room at the Fletcher School, in Washington

DC, Geneva, London and Portland identify GPS coordinates and geo-tag the reports, and add relevant classification tags. This data is plotted on a webpage (Figure 28[15]) pursuing similar goals to the WKI project. Therefore the data provides an excellent gold-standard with which to evaluation the WeKnowIt text analysis technology. Unfortunately the data does not include any user information and therefore user and social network context cannot be considered during the analysis process, which has been shown to be effective in previous evaluations [19].

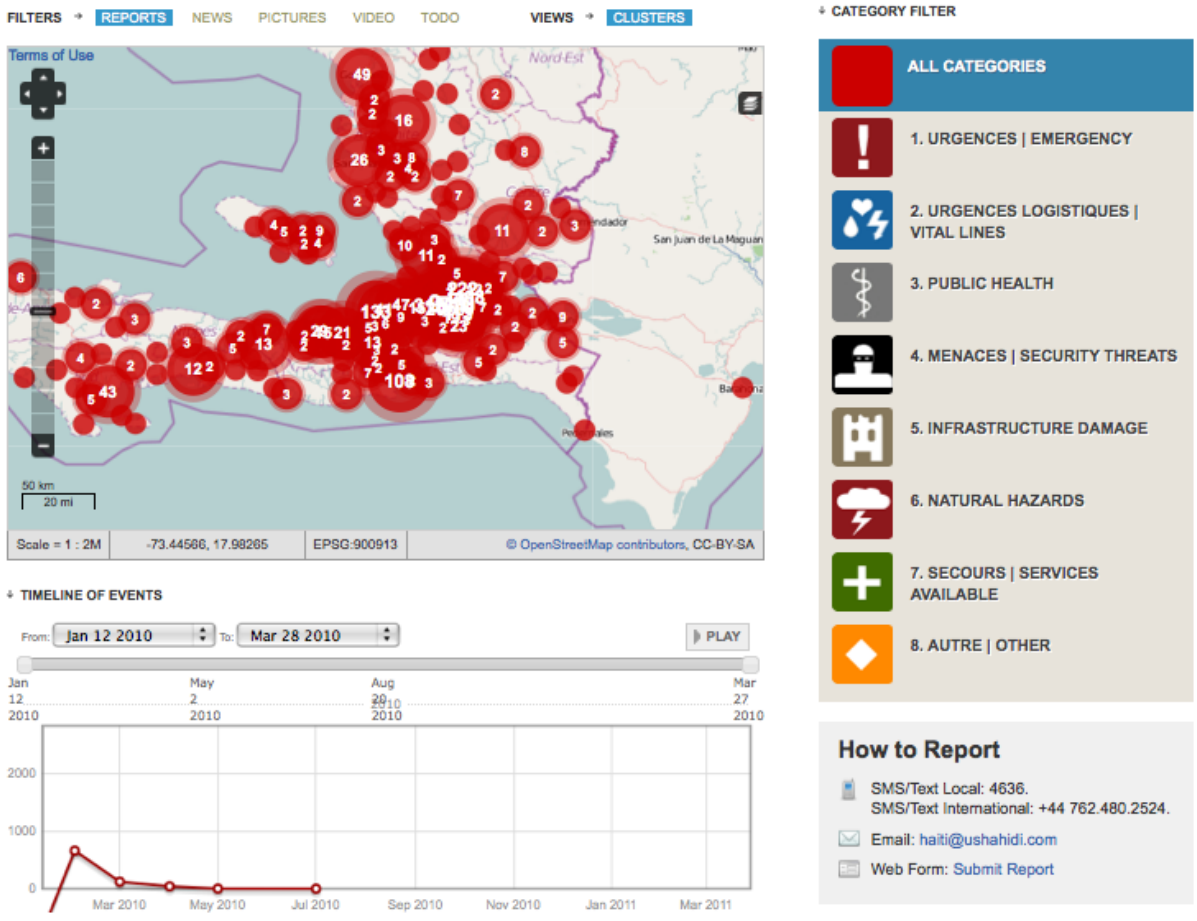

Figure 28: Ushahdidi Haiti Web Page

There was a total of 3,600 documents in the Haiti dataset, of these 13 were removed as they contained less than three words in the description (the majority of these were test messages). All the remaining 3,587 documents have been manually geocoding with a latitude and longitude values, however 6 of the documents were not classified with any tag, these documents are included in the geocoding evaluation but not the classification evaluation.

Table 15 shows the descriptive statistics of the data, both in terms of words and characters used. As can be seen the messages are generally brief; the average being about 60 words and 300 characters.

Table 16 shows the classification tags used to categorise the documents, and their observed frequency in the dataset. The tags are structured in a two level hierarchy (similar in structure to that devised by the Sheffield City Council in WP7). Aside from the 6 untagged documents,

[15] http://haiti.ushahidi.com

Table 15: Ushahdidi Haiti Data Descriptive Statistics

Measure	Words	Characters
Min	3	12
Max	762	4423
Mean	63.3	332.3
Median	51	274

each of the other documents has been assigned between 1 and 6 tags. In order to standardise the tagging, the documents were pre-processed so that if a document is assigned a leaf tag then it is also tagged with its parent root tag. However root tags can be assigned to documents without further differentiating with leaf tags.

There is a considerably skewed distribution of tags with almost 63% of documents having the "2. Vital Lines" tag, while just over 1% have the "6. Natural Hazards" tag

Table 16: Ushahdidi Haiti Data Classification Tags

Tag	Name	Frequency
1.	Emergency	581
1a.	Highly vulnerable	2
1b.	Medical Emergency	209
1c.	People trapped	166
1d.	Fire	7
2.	Vital Lines	2252
2a.	Food Shortage	1597
2b.	Water shortage	1333
2c.	Contaminated water	21
2d.	Shelter needed	477
2e.	Fuel shortage	21
2f.	Power Outage	35
3.	Public Health	330
3a.	Infectious human disease	10
3b.	Chronic care needs	2
3c.	Medical equipment and supply needs	305
3d.	OBGYN/Women's Health	7
3e.	Psychiatric need	4
4.	Security Threats	320
4a.	Looting	25
4c.	Group violence	1
4e.	Water sanitation and hygiene promotion	240
5.	Infrastructure Damage	193
5a.	Collapsed structure	142
5b.	Unstable Structure	32
5c.	Road blocked	29
5d.	Compromised bridge	1
5e.	Communication lines down	1
6.	Natural Hazards	44
6a.	Deaths	2
6b.	Missing Persons	17
6c.	Earthquake and aftershocks	22
7.	Services Available	813
7a.	Food distribution point	333
7b.	Water distribution point	5
7c.	Non-food aid distribution point	77
7d.	Hospital/Clinics Operating	256
7g.	Human remains management	37
7h.	Rubble removal	7
8.	Other	490
8a.	IDP concentration	20
8c.	Price gouging	1
8d.	Search and Rescue	49
8e.	Persons News	291
8f.	Other	4

6.3 Text Geocoding

The first experiment concerns the geocoding of the documents. The presumption being that the human geo-coding of the documents provides a Gold-Standard, although it should be recognised that there will be annotation errors, as there will be misinterpretations. Also it is not clear whether any validation of the geo-coding was undertaken. The Yahoo geo-tagging web service is used for the comparative evaluation.

6.3.1 Methodology

The geocoding technique follows from that described in the previous deliverable [19], whereby a semantically structure geographical resource is constructed by fusing two freely available and extensive resources, namely Yahoo! GeoPlanet[16] (for this experiment an update version was used[17]) and OpenStreetMap[18]. Yahoo! GeoPlanet provides the overall structure linking locations in a hierarchy and OpenStreetMap provides the detailed, fine-grained locations.

The processing of the data remains largely as described previously, excepting data specific augmentations and improvements to data pre-processing. Whilst the Haitian textual data is largely English, French and Creole nouns are often used. It was therefore necessary to add French name stopword list to prevent person names being incorrectly geocoded. In addition the ambiguity of locations is considered by measuring the co-occurrence of locations in the training data documents. If some location term, e.g. “Main Road”, co-occurs with a variety of disparate other locations then it is deemed too ambiguous and removed from consideration. The ambiguity measure is taken as the mean distance to co-occurring locations, which do not enclose the location being measured. Locations with ambiguity measures above some threshold (in the experiment 10 km) are removed from consideration. This measure of ambiguity provides data specific pre-processing.

When a document is geocoded a number of possible geo-locations can be identified. These are resolved down to a single geotagging value by selecting the most likely alternative (described in [19]).

A three-fold cross-validation experiment was performed with two-thirds of the data used for training, i.e. assessing location ambiguity. The results of the experiment are present below.

6.3.2 Results

Figure 29 shows the results of the geocoding. The WKI geocoding process provided locations for 2,045 (57%) of the 3,587 documents. Given the set of potential locations discovered for each document the figure shows the distances between the actual document location, the location selected to geocode the document and the optimal (i.e. nearest) location. The difference between the two curves indicates the degree to which the context was not available to properly disambiguate the locations. The mean values are 22.48 km for the selected locations and 12.51 km for the optimal location. As a comparison the mean values for geocoding without removing the ambiguous locations are 31 km from the selected locations and 8.26 km for the optimal location.

[16] http://developer.yahoo.com/geo/geoplanet/
[17] Version 7.6.0 released 2010-10-22
[18] http://www.openstreetmap.org

As can be seen a significant improvement is made by removing the ambiguous choices from the location selection process.

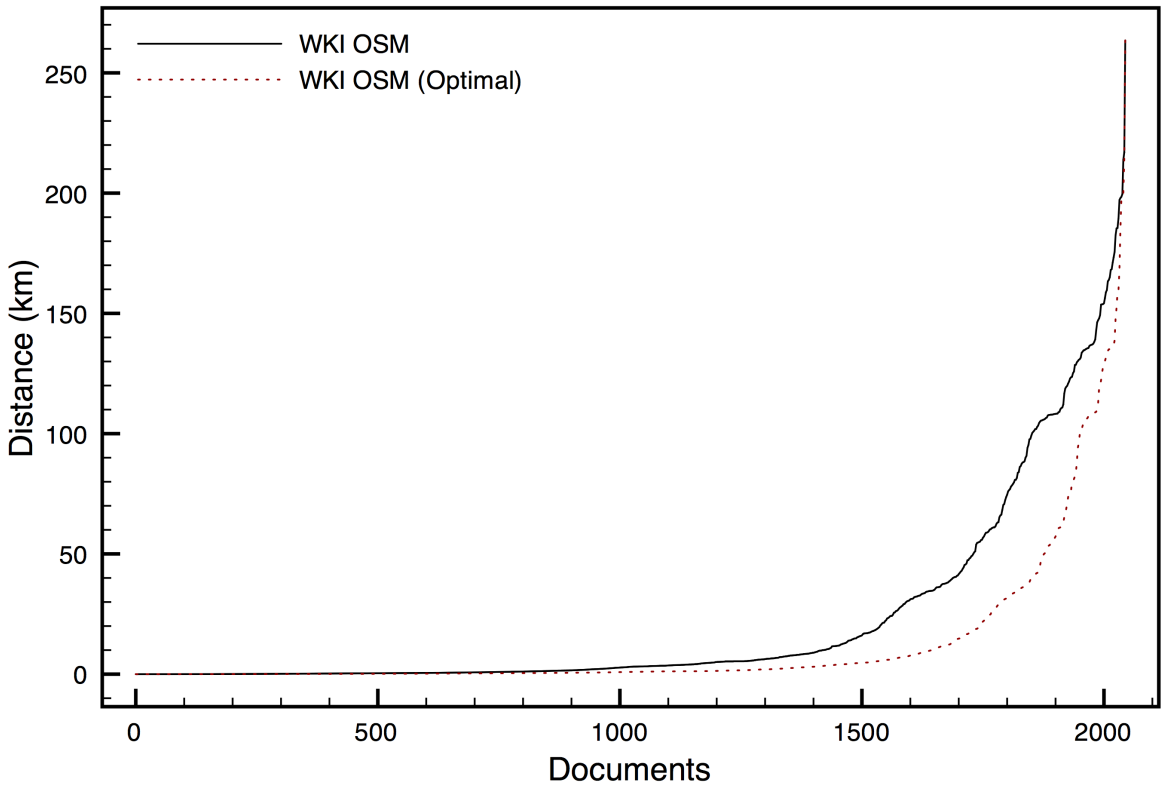

Figure 29: WeKnowIt Text Geocoding Performance on Haiti Data

In order to provide a comparative evaluation both the Google[19] and Yahoo[20] Geocoding Web Services were employed. The Google service is not intended to geoparse text, rather to provide geocoding of addresses, therefore it provided virtually no valid responses. However the Yahoo Placefinder service is intended for parsing text and provided a reasonable number (650) of responses. The overall comparison of the two services can be seen in Figure 30, which shows the distances from the actual to predicted geo-location. In addition, Table 17 provides a breakdown of the distances into a number of ranges. From this it can be seen that the WKI OSM system provides considerably more responses and in particular where the selected distances are proximate to the actual distances, e.g. WKI OSM provides nearly 16 times the located documents which are within 100 metres of the actual locations (254 and 16 documents, respectively).

Figure 31 considers solely the 650 documents for which Yahoo Placefinder has provided a location. As can be seen there is very little difference between the two systems for these documents. With a slightly higher mean value of 22.32 km for the WKI OSM systems compared to 22.1 km for Yahoo. A paired t-test shows that the results do not significantly differ (p-value: 0.91).

[19] http://code.google.com/apis/maps/documentation/geocoding/
[20] http://developer.yahoo.com/geo/placefinder/

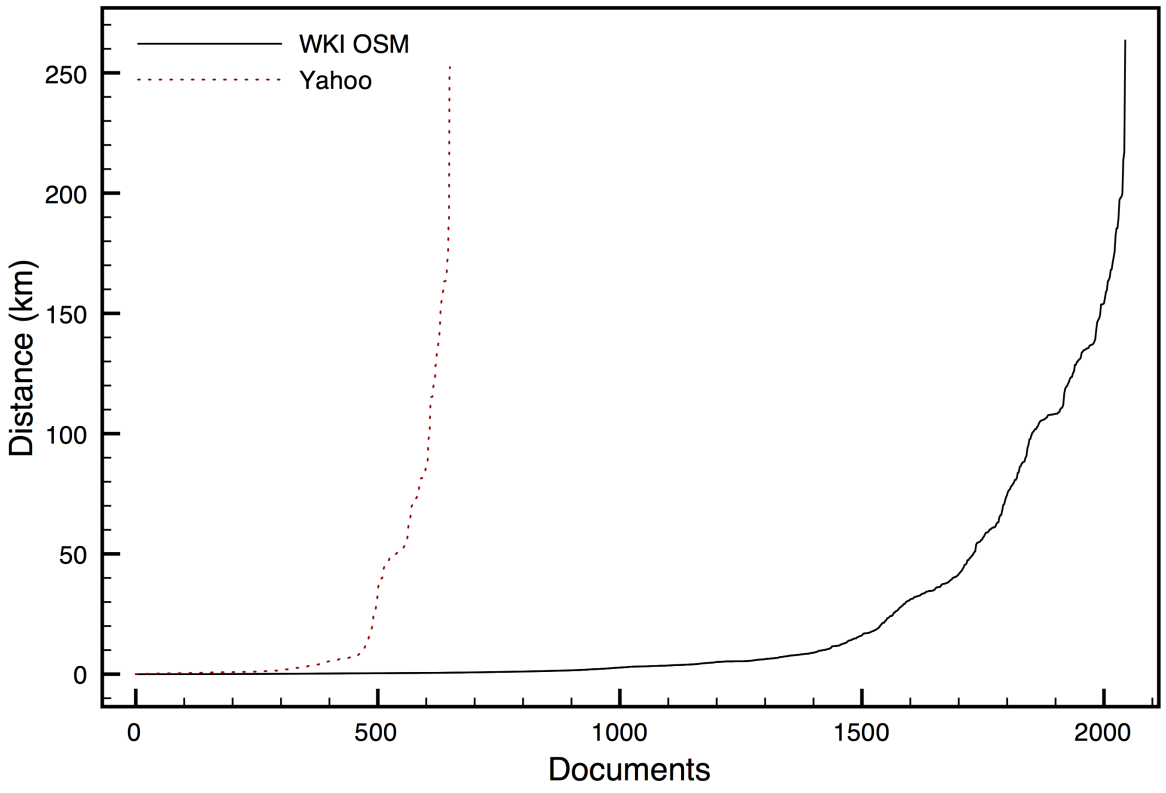

Figure 30: WeKnowIt/Yahoo Text Geocoding Performance on Haiti Data

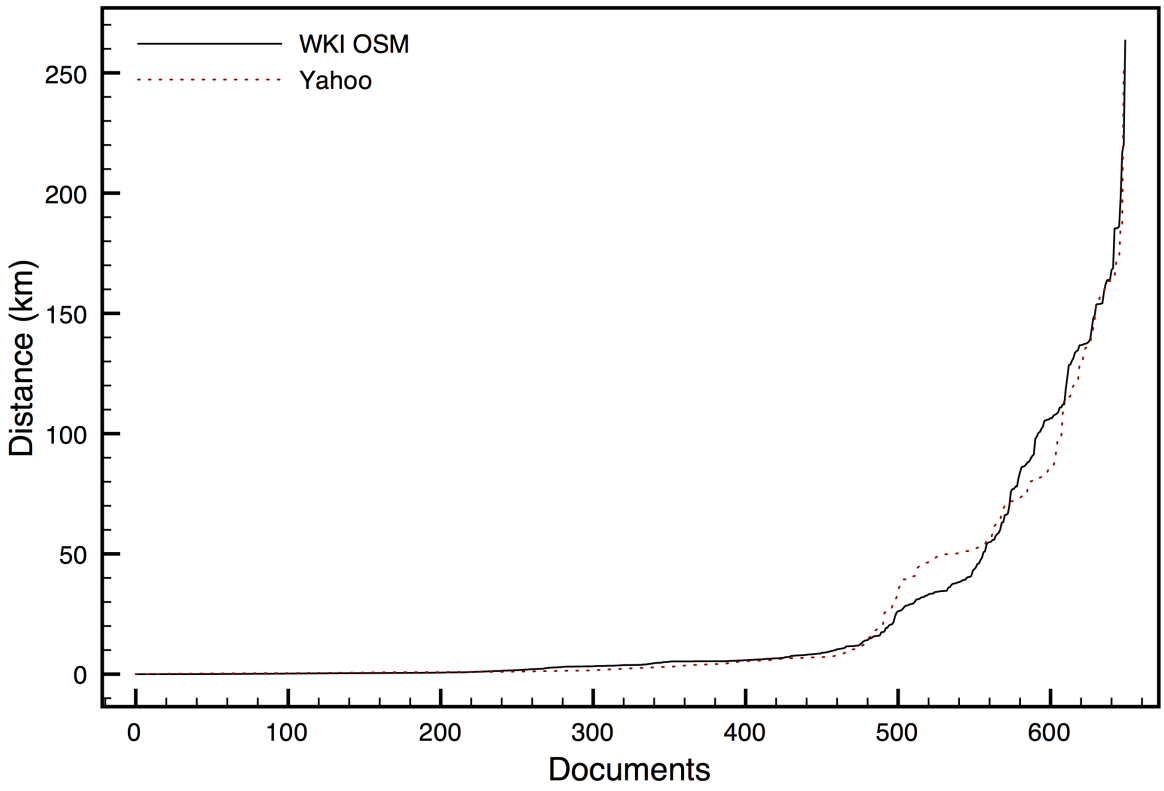

Figure 31: WeKnowIt/Yahoo Text Geocoding Paired Performance on Haiti Data

Table 17: WeKnowIt/Yahoo Text Geocoding Distance Histogram on Haiti Data

Distance (km)	WKI OSM	Yahoo
$0.0 \geq x < 0.01$	80	14
$0.01 \geq x < 0.1$	174	2
$0.1 \geq x < 0.25$	130	26
$0.25 \geq x < 0.5$	203	75
$0.5 \geq x < 1.0$	189	124
$1.0 \geq x < 2.0$	164	70
$2.0 \geq x < 5.0$	258	84
$5.0 \geq x < 10.0$	223	73
$10.0 \geq x < 20.0$	117	21
$20.0 \geq x$	507	161
Total	2045	650

6.3.3 Discussion

It can be argued that the Yahoo Placefinder service is being conservative in it's geocoding, i.e. high precision, low recall. However the WKI geocoding system shows that it is possible to geocode such informal text and that the performance is at least comparable with the current state of the art services. It is unfortunate that for this data set user information is not available as previous work has shown that user context can further improve performance.

6.4 Text Classification

The second experiment is concerned with facilitating the classification of the documents. As above the it is presumed that the human classification of the documents provides a Gold-Standard, although there is no validation of the classification (i.e. measures for inter-annotator agreement). However the experiment evaluates the degree to which the automatic classification techniques can model the human classifiers.

6.4.1 Methodology

The experiment is a standard text classification experiment looking at the application of the state of the art supervised machine learning algorithm to generate a classification model on a training data set and apply that model to the unseen data.

In application terms the classification process is seen as a facilitation services for determining the relevant tags for a document, rather than a fully automated system. This is particular important for complex classification scenarios, such as the ones required by WP7, where a document can be classified with a number of potential tags, which are not completely mutually exclusive, rather structured in a hierarchy.

In order to classify such data it is necessary to employ a hierarchical multi-labelled classification technique, for an overview see Tsoumakas et al. [78]. The approach adopted is called Hierarchical Binary Relevance (HBR) method, in which a multi-label classifier is used to train a binary classifier for each tag (label) of the hierarchy. The original dataset undergoes a transformation whereby for each tag a dataset is constructed where the selected tag is deemed positive and all other tags are deemed negative. For the classification of an unseen (test) instance the classifier outputs the union of the tags that are positively predicted by the all the separate tag classifiers. In the hierarchical classifier instances are classified in a top-down manner, with the sub-tag classifiers only being called if the parental tag receives a positive classification. For the generation of the binary relevance multi-labelled classifiers a Support Vector Machine (SVM), classifier was employed [13]. In order to optimise the parameters of the SVM function (cost and gamma) a grid search method was employed, again using a three-fold cross-validation sample on the training data set to determine the optimal values.

Note that as in the Haiti data a document may be classified with a parental tag without a child tag an additional child tag was constructed to account for this case. For a baseline comparison a classifier was employed which simply ranked the tags according to their frequency in the training data set.

6.4.2 Results

The section presents the results of the classification experiments which examine the generation of predictive models to automatically classify the Haiti data according to the categories (tags) presented above, in Table 16. The first set of results, shown in Figure 32 and Table 18, ignores the leaf tags and only consider the 8 root tags.

Figure 32 shows the performance measures as the number of ranks considered positive increases. In effect this simulates providing the user with more potential choices. If only the top rank is se-

lected as positive then precision is around 70%, while recall is only about 50%, as the number of positive ranks increases precision falls and recall increases. For small flat classification schemes this process may be seen as a ranking one, minimising the amount of time the user requires to scan down the list to find the correct tag. In this experiment the baseline performs better on all but the prediction of a single best tag, however this case accounts for 2477 (68.8%) documents for which there is a single root tag.

Table 18 shows the performance on each of the 8 tags (with a positive ranks for 3). As can be seen the performance on the tags is strongly correlated with the frequency of the tag.

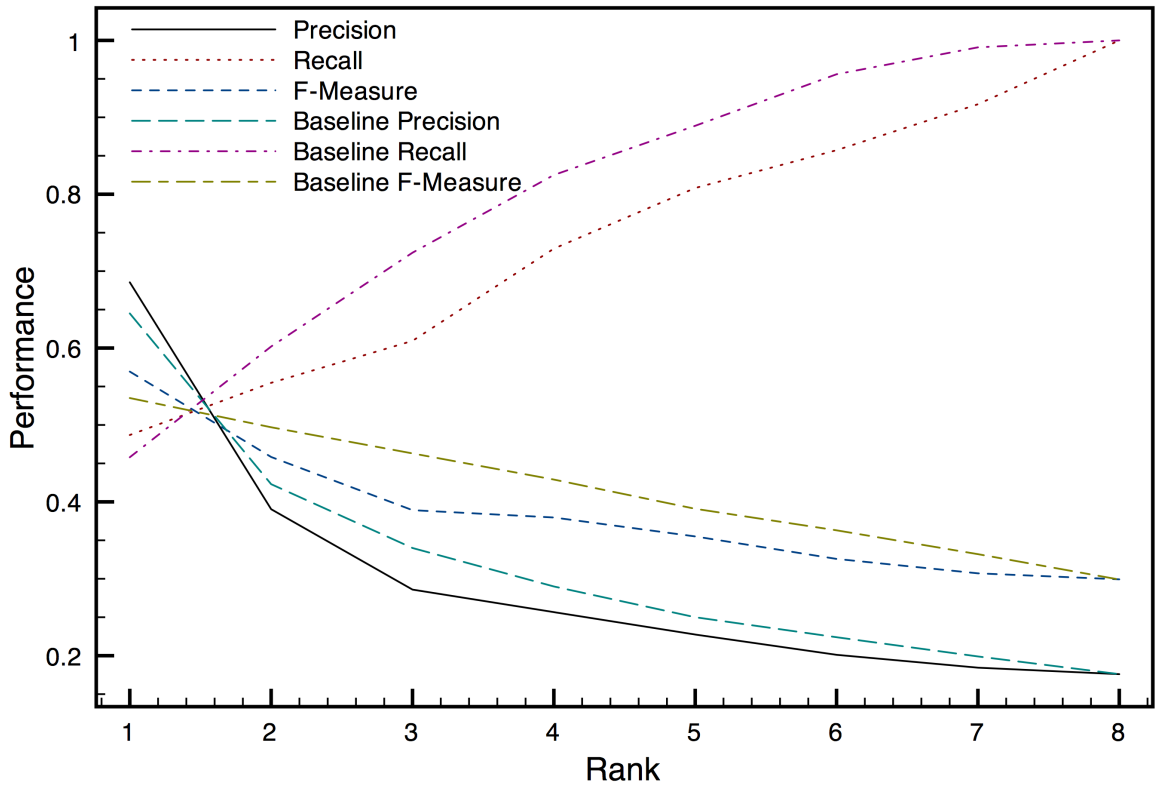

Figure 32: WeKnowIt Text Classification Performance on Haiti Data Root Tags

The next set of results, shown in Figure 33 and Table 19, consider all the categories (tags) in the structured hierarchy. Comparing the results with those above indicates that the performance does not substantially degrade when considering all the tags, which indicates that the methods employed in hierarchical classification are effective. In this experiment the WKI Classifier performs significantly better than the baseline throughout the whole graph.

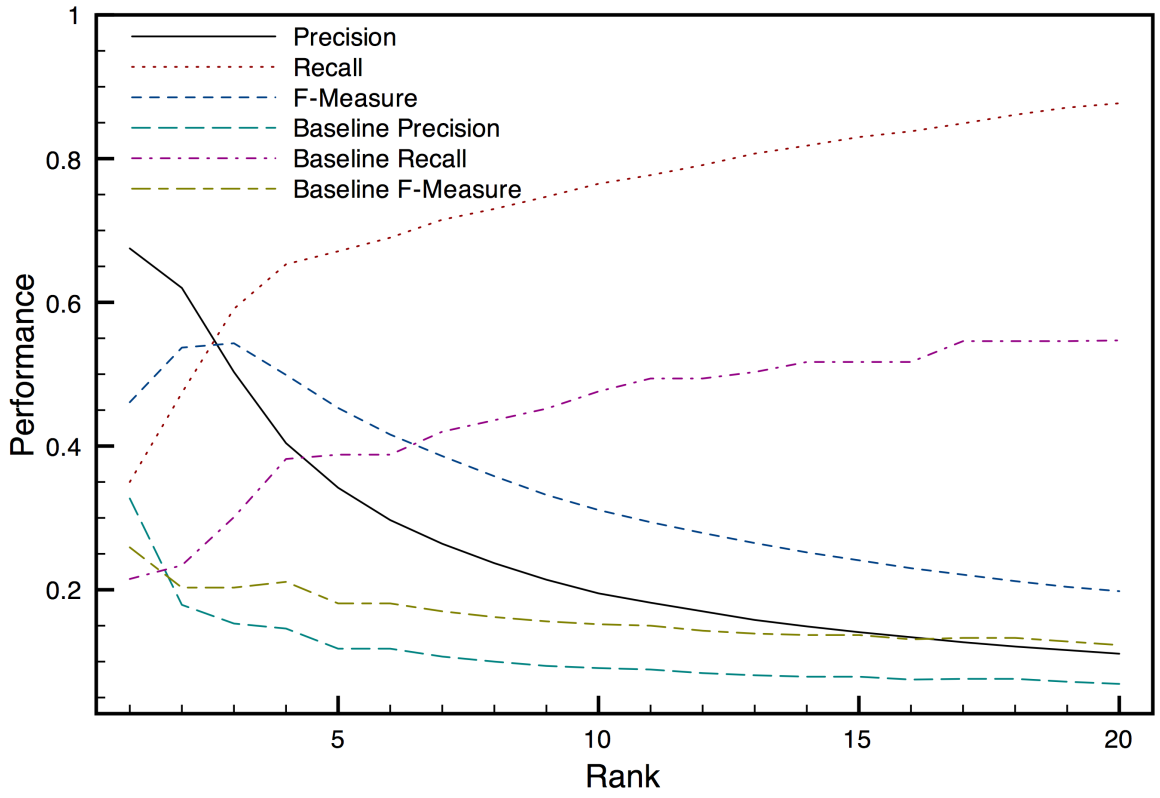

Figure 33: WeKnowIt Text Classification Performance on Haiti Data All Tags

Table 18: WeKnowIt Text Classification Performance on Haiti Data Root Tags

Tag	Frequency	Precision	Recall	F-Measure
1.	541	0.770	0.216	0.338
2.	2207	0.841	0.860	0.850
3.	325	0.609	0.086	0.151
4.	305	0.813	0.085	0.154
5.	173	0.743	0.150	0.250
6.	41	0.022	0.317	0.041
7.	777	0.762	0.140	0.237
8.	451	0.754	0.286	0.415

Table 19: WeKnowIt Text Classification Performance on Haiti Data All Tags

Tag	Frequency	Precision	Recall	F-Measure
1.	541	0.503	0.409	0.451
1a.	2	0.000	0.000	0.000
1b.	201	0.304	0.035	0.063
1c.	150	0.542	0.687	0.606
1d.	5	0.000	0.000	0.000
2.	2207	0.770	0.956	0.853
2a.	1576	0.642	0.876	0.741
2b.	1311	0.632	0.699	0.664
2c.	20	0.019	0.200	0.035
2d.	470	0.668	0.330	0.442
2e.	20	0.000	0.000	0.000
2f.	34	0.037	0.412	0.068
3.	325	0.631	0.311	0.416
3a.	10	0.500	0.100	0.167
3b.	2	0.000	0.000	0.000
3c.	300	0.604	0.310	0.410
3d.	7	0.714	0.714	0.714
3e.	4	0.000	0.000	0.000
4.	305	0.494	0.125	0.199
4a.	24	0.286	0.083	0.129
4c.	1	0.000	0.000	0.000
4e.	233	0.387	0.052	0.091
5.	173	0.542	0.225	0.318
5a.	123	0.596	0.228	0.329
5b.	32	0.000	0.000	0.000
5c.	28	0.091	0.107	0.098
5d.	1	0.000	0.000	0.000
5e.	1	0.000	0.000	0.000
6.	41	0.000	0.000	0.000
6a.	2	0.000	0.000	0.000
6b.	17	0.000	0.000	0.000
6c.	19	0.000	0.000	0.000
7.	777	0.349	0.456	0.396
7a.	323	0.117	0.300	0.169
7b.	5	0.000	0.000	0.000
7c.	74	0.026	0.054	0.035
7d.	233	0.642	0.300	0.409
7g.	37	0.667	0.108	0.186
7h.	7	0.000	0.000	0.000
8.	451	0.532	0.503	0.517
8a.	20	0.071	0.600	0.127
8c.	1	0.000	0.000	0.000
8d.	48	0.000	0.000	0.000
8e.	266	0.421	0.605	0.497
8f.	4	0.000	0.000	0.000

Two further experiments have been performed to examine factors which are relevant to the nature of the data.

6.4.2.1 Impact of Training Data Size on Text Classification

This experiment considers how the size of the training data influences the performance, as for real-time, ad-hoc data, such as that generated during emergencies, it is important to consider how much data is required to provide adequate performance. Figure 34 shows the improvement in performance measures as a greater proportion of the training data is used to generate the classification model. It can be seen that there is a fairly constant increase in performance, which indicates that for such a classification problem as undertaken in this evaluation there is little redundancy in the training data set.

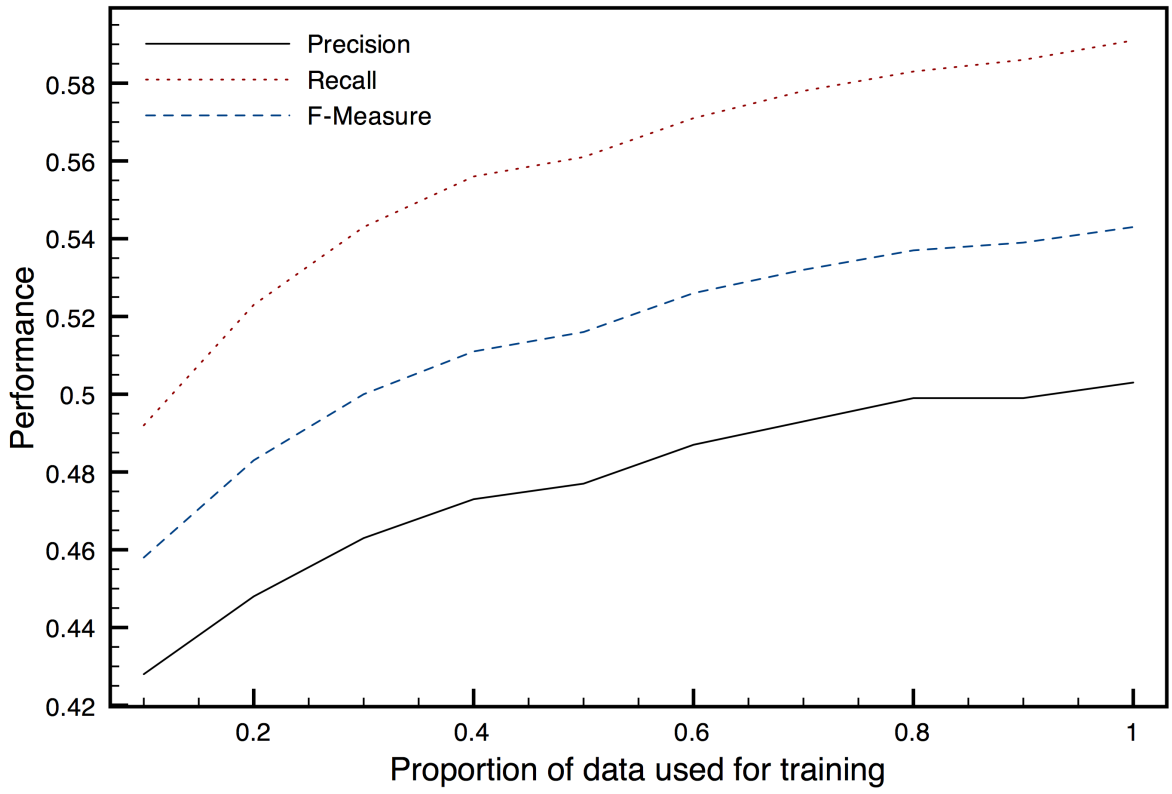

Figure 34: WeKnowIt Text Classification Performance on Haiti Data: Effect of Training Data Size

6.4.2.2 Using Spatial Context in Text Classification

The final experiment considers the use of the spatial information within the classification task. As there is no user context, it is pertinent to determine if location can provide useful context. Intuitively it is likely that a tag assigned to a document in a given locations is more likely to be relevant to another document in a proximate location. In order to perform the experiment

the locations from the documents in the training set are clustered, using a Simple K-means technique. The test documents are then assigned to these clusters according to their location and then the cluster is used as an input into the classification model.

Figure 35 shows the f-measure performance of this technique compared with the results from the experiment above, where the training data size is varied. As can be seen the use of location in classification does provide a small but significant increase in f-measure. The effect is particularly observable in the trials with low levels of training data. The fall performance at the large amounts of training data may be cause by a decrease in clustering performance as the number of data points increases and clusters become less distinct.

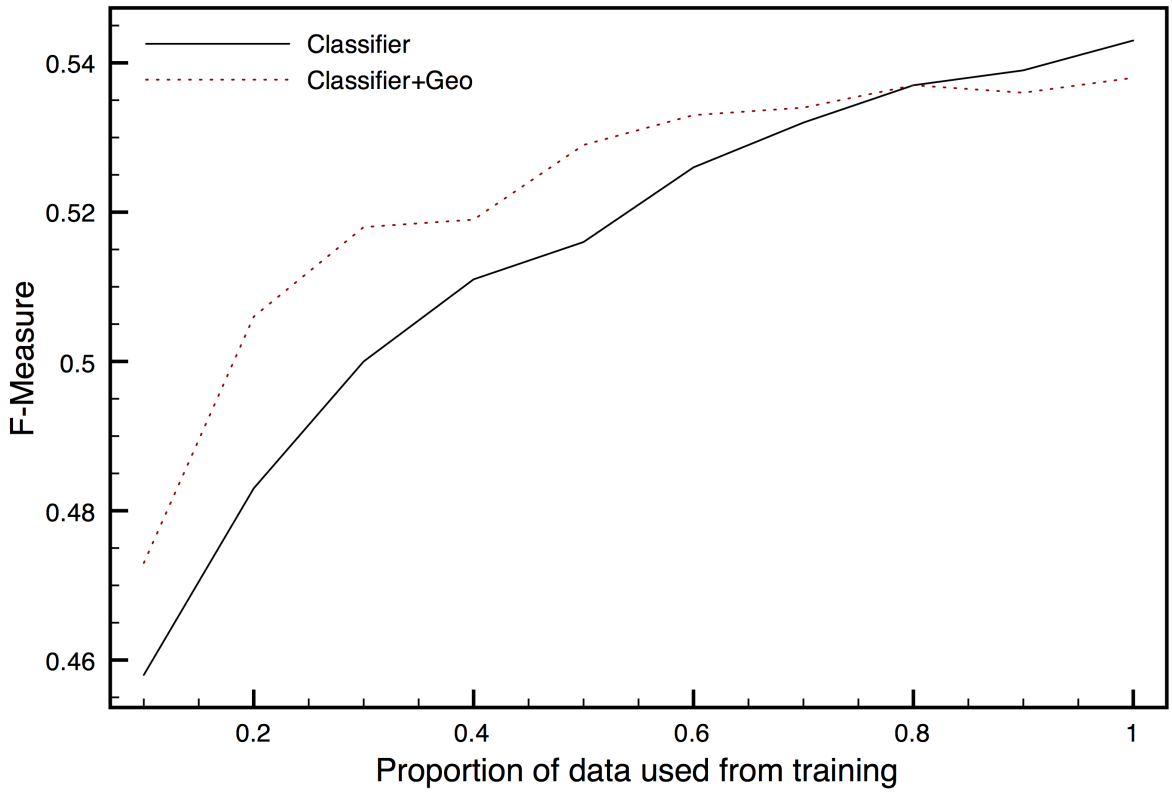

Figure 35: WeKnowIt Text Classification Performance on Haiti Data: With Geocoding Context

6.4.3 Discussion

In the classification evaluation solely considering the 8 root nodes, whilst the WKI classifier provided the best performance when selecting the single best tag the baseline classifier provided better performance when more tags are considered as positive. Therefore in cases where there are few, none hierarchical, unevenly distributed tags, simply presenting the user with a frequency ordered list may be the most effective means of selecting the correct tags. However the second experiment showed that in more complex classification scenarios such a simple approach is less effective. Figure 33 shows that when selecting the top five ranked tags as positive, recall for the WKI classifier is around 65% (compared to around 40% for the baseline), while this provides a significant improvement, it still means that in over a third of the cases the user will have to

consult the hierarchy to find the correct tag. It is therefore necessary for the user to understand the extent to which the tag suggestions should be "trusted". It would be hoped that as the users become more aware of the systems performance and the classification possibilities they develop an understanding of the sufficiency of the suggestions to fully classify the content.

The experiment concerning the size of the training set showed that maximising the amount of training data is important in terms of performance. This is an intuitive result given the complexity of the classification task. That is, to classify short, informal text with one or more of 45 possible tags. For simpler classification tasks less data should be required.

The final experiment shows that including relevant contextual information (in this case general location) can be beneficial for classification of such user-generate content, a result congruent with previous work relating to the inclusion of user contextual information [36].

6.5 Conclusion

The evaluation above shows that text analysis of user-generated content from an ad-hoc (emergency) event can provide potentially useful information, and in some cases performance comparable to human annotation. Although in applications where the performance provided by the automatic methods in insufficient they should be seen as facilitating humans (with suggestions) rather than replacing them. However, in using such an approach the users should be made aware of the danger of the system adversely affecting performance, as the user simply chooses from the potentially incorrect, imprecise or insufficient locations and tags from the selection which are offered by the system. This may be more likely to occur when the user has little vested interest in providing fully correct information, in which case some evaluation of the user would be useful, or where the user has limited time, in which case there is a potential trade-off between the user providing no information beyond the text versus them providing further information which is potentially imperfect.

7 Evaluation of speech methodologies

7.1 Evaluation of Speech Tagging

7.1.1 Setup of Experiments

Our recognizer developed for NIST Rich Transcription 2007 evaluations within the AMI/AMIDA project [29] served as a baseline for reporting word accuracies. It used a 50k language model, fast speaker adaptations (VTLN, CMLLR) and one-pass bi-gram lattice decoding. Decoding is done on PLP and posterior features processed using HLDA and CVN+CMN. The bi-gram lattices were expanded to 4-grams.

The hybrid recognizer used for partial OOV word detection was derived from this baseline system. The setup was reduced to the first pass (bi-gram decoding only) and the word recognition network was replaced by a hybrid word/sub-word recognition network.

The multigram sub-word language model [18] consisted of 3977 phone and multiphone units trained on the AMI RT06 50k dictionary. The word language model was an open-set Katz-backoff language model trained on a total of 60M words from conversational telephone speech (CTS) and broadcast news data (BBC).

Table 20 shows the data sets used for training the word language model. The vocabulary size

Data	BBC	fisher1	fisher2	AMI	AMI05	h5	icsi	m4
#Words	33M	11M	11M	800k	15k	3M	650k	26k

Table 20: Data for word LM training

was fixed to 36k words by frequency cut-off 2 on meeting and CTS data. The hybrid language model consisting of a word and a sub-word model was combined in form of weighted finite state transducers by the use of the OpenFST toolkit[21].

After mapping the reference transcript to unified UK spellings, we ran a forced-alignment to obtain a precise timing of all OOV words.

7.1.2 Recognition Results and Their Analysis

The baseline recognizer obtained a word accuracy of 69% on all the data. We investigated how the use of the hybrid word/sub-word model excels the recognition of words compared to a word-only recognizer. Therefore, we examine all reference words in the transcripts depending on their estimated and real frequency. We measured a per-word correctness. A recognized word has been considered correct iff there was an overlap in time with the reference word covering all phonemes of the reference word, and both reference and recognized words showed identical spelling. With the fixed operating point provided by the one-best recognition output of the hybrid recognizer, where sub-word sequences were interpreted as partial detections, we obtained partial OOV detections. Figure 36 shows the precision, recall and f-score of all partial OOV detection tokens sorted independently by score. A high number of false alarms are shown in the right with scores of 0. To the left, a fair number of partial detection tokens overlap almost perfectly.

[21] http://www.openfst.org

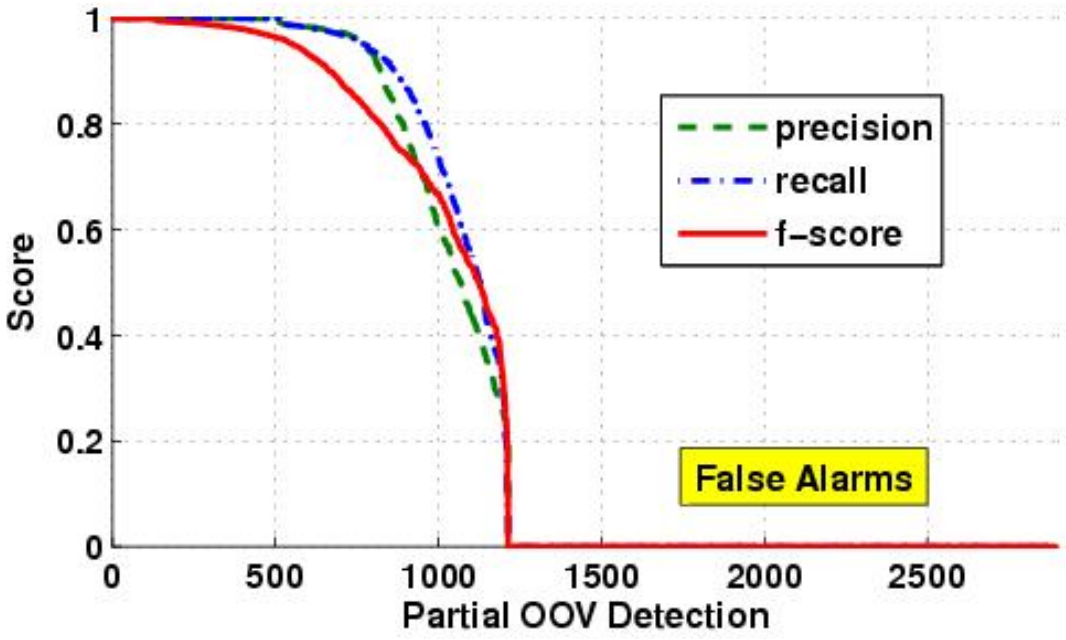

Figure 36: Score distributions over partial OOV detections sorted by descending score.

Partial OOV detections with high f-scores qualify for word recovery. Using Phoneme-to-Grapheme (P2G) conversion, we successfully recovered the correct spelling of detections from the phonetic description inherent in the corresponding sub-word sequences. We trained a joint multigram P2G model up to 8-grams on the 36k decoding dictionary using Sequitur[22]. Co-alignments of length one and zero between phonemes and characters have been used. and divided those into two sets according to whether the obtained spelling was an OOV or not. Approximately 24% of the hits were fully recovered to the correct spelling.

The new approach brings improvements in terms of word accuracy and OOV detection performance. The hybrid word/sub-word recognition as a mean to improve the speech recognition accuracy especially on rare, information-rich words showed to be successful. As shown experimentally here as well as in [30], the hybrid word/sub-word recognition in combination with phoneme-to-grapheme conversion is also able to recover rare words. The recovery of partial, potentially reoccurring OOV words, which get detected with a low f-score only, remains an interesting issue for future research.

7.2 Additional Value of the Speech Technology for Media Fusion Task

This section discusses the media fusion experiment aimed at showing the additional value of the speech recognition components to the overall performance of automatic tagging tasks based on the WeKnowIt media intelligence services.

First, it should be noted that the presented evaluation results focus on a specific aspect of the media fusion mechanisms applied in the project. To be able to quantify the exact contribution of a particular media analysis method (speech, in this case) in contrast to the other means of the

[22] http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

analysis (image/video and text), a real data containing each of the above-mentioned modality input for each individual case needs to be collected. This setting is somehow artificial – in reality, it is often the case that just one modality is available, e.g., there is only a picture without any tags, there is only a phone call and no video from the scene etc. The fair evaluation of the individual contributions of the media analysis methods is specific for the particular use cases so that the evaluations in WP7 should be seen as an extension to the experiments described here.

7.2.1 Media Fusion Setup

Two key requirements were considered when defining the setup for the experiment:

- to define a common basis for all three modalities involved (video and image taken as one here);
- to focus on realistic evaluation represented by ad-hoc queries rather than specific ones that the analysis methods could be tuned on.

The first requirement led to the definition of a shared classification setting in which all three media systems are employed to decide whether (to what extent) a particular case corresponds to a specific category. It is then easy to compare the results of the individual classifiers and to produce the fusion system combining the results of the particular media components.

As a result of the second requirement, data for a specific task was collected. It corresponds to the ER scenario situation in which the responsible person searches for a piece of information on "car in water" (e.g., to estimate the water level in the particular place). 172 images and video files were collected together with the related texts from the web pages. Related speech recordings were available for 97 of these cases. The rest was randomly distributed to 5 persons (3 males and 2 females) that were asked to describe the situation on the picture/in the video. The actual recording employed the speech tagging facility of the WKI mobile SearchInSpeech demo application.

Out of the 172 cases, 83 were manually selected as positive examples of the category, the rest as negative ones. Most of the pictures came from disaster reports so that the data realistically model what could be sent to an ER team and searched in actual situation. Further analysis of the data showed that the classification is very hard for all the classifiers – the level of noise in the voice recordings is high, the texts mentioning relevant terms are often used for images that do not show "car in water", there are videos of damaged cars not being in water and flooded objects that are not cars. Figures 7.2.1 and 7.2.1 demonstrate two positive examples from the Sheffield 2007 flooding subset of the dataset, Figures 7.2.1, 7.2.1 and 7.2.1 show the negative ones.

If there are enough training examples, all the classifiers could be trained for the particular task in hand. However, it is unrealistic to expect such a setting for most of the situation the media intelligence should be applied in. That is why we built the actual classifiers for image/video and speech from generic modules trained on non-specific data. The text classification was allowed to train on 12 texts describing additional positive examples (not included in the evaluation set) and on 19 texts for in-domain negative cases. Apart from that, it used the automatically derived list of paraphrases based on wordsketches [44] generated on the English GigaWord Corpus [28].

The image classification consisted in a combination of a general classifier of objects developed as a part of our TrecVid2010 activity [35] (object(car)) and the region classifier described in [32] (in region(water)).

Figure 37: A positive example for 'car in water' – http://www.bbc.co.uk/southyorkshire/content/image_galleries/flooding_june_07_gallery_two.shtml?59

Figure 38: A positive example for 'car in water' – http://www.bbc.co.uk/southyorkshire/content/image_galleries/flooding_june_07_gallery_two.shtml?56

Figure 39: A negative example for 'car in water' – http://www.bbc.co.uk/southyorkshire/content/image_galleries/flooding_june_07_gallery_two.shtml?30

Figure 40: A negative example for 'car in water' – http://www.bbc.co.uk/southyorkshire/content/image_galleries/flooding_june_07_gallery_two.shtml?36

Figure 41: A negative example for 'car in water' – http://www.bbc.co.uk/southyorkshire/content/image_galleries/flooding_june_07_gallery_two.shtml?75

The speech classifier used 2-pass decoding with speaker adaptations (CMLLR, VTLN) and was derived from the AMIDA 2005 CTS recognizer as previously used in [11]. As acoustic features, we used posterior features using long temporal context. The acoustic models were trained speaker independently on 250 hours of Switchboard data. As recognition network, we used a hybrid word/sub-word language model (LM). The sub-word LM consists of 3977 phone and multiphone units trained on the RT06 dictionary ([77], 47k words). The word LM (bigram open-set Katz-backoff) was trained on $\approx$ 2250 hours of Switchboard (1+2) and Fisher.

Note finally, that due to the specific character of the data necessary for this kind of evaluation, the performance figures of the individual media detectors strongly depend on the actual data collected for the experiment. Rather than to the particular results of the individual methods, it is therefore relevant to pay attention to the complementarity of the individual components and the potential for the fusion.

7.2.2 Results and Discussion

Table 7.2.2 summarizes the results. Even though the resulting accuracy figures for the individual classifiers are closed to each other, the analysis of the confusion matrices showed a relative high level of complementarity. This is especially true for the pairs text-image and speech-image, the results of the text-based and speech-based classifiers were highly correlated.

To take advantage of the media fusion, we employed a simple SVM-based combination of the individual classification results. The probability estimates / certainty factors of the particular

Classifier	Accuracy
Text-based classifier	57.6%
Image-based classifier	58.7%
Speech-based classifier	61.6%

Table 21: Results of the individual classifiers on the media fusion dataset

classification cases were used together with the knowledge on the correct answer for the training. Ten-fold cross validation with 90% training and 10% testing was applied.

The resulting average accuracy reached 67.2 %. The fusion-based classifier significantly overcame the results of each individual method. As expected, the highest gains were brought by the combination of the speech-based and image-based techniques.

8 Conclusions

In this document we reported on the evaluation techniques and results developed within the WeKnowIt *Media Intelligence* layer and integrated within the WeKnowIt system. More specifically, we discussed the latter modules implemented to demonstrate and evaluate automated visual analysis and knowledge extraction from raw visual content and associated metadata functionalities, speech indexing and OOV-detection techniques, text categorization and clustering techniques, as well as multimedia fusion techniques.

More specifically, in this deliverable we started by discussing the importance of multi-modal analysis in social media, given the widespread adoption of social networks and web 2.0 and the major trend in the new digital landscape towards the interconnection of platforms, networks and most importantly data.

Furthermore, we discussed our latest developments and updates with respect to the ViRaL application, focusing on its latest functionalities. The initially presented approach on visual image retrieval and localization (see D2.1.1 and D.2.1.2) is extended herein towards new, unexplored paths, focusing on how multimedia content may be explored towards todays social user communities. In the work described within this deliverable, we successfully exploited and evaluated the performance of Feature Map Hashing, presented in D2.3 and our recent publication [6], against state of the art approaches for large scale image retrieval. WE also evaluated the performance of Scene Maps and View Clustering, as presented again in D2.3 and our recent publications[5][40]. We compared against state of the art approaches for large scale image retrieval. Moreover, we evaluated our location and landmark recognition techniques, also presented within D2.3.

Moreover, we presented an empirical analysis comparing four baseline matching methods that rely on photo metadata, three variants of an approach that uses cluster analysis in order to discover POI-related photo clusters, and a real-world retrieval mechanism (Flickr search) on a set of less popular POIs. A user-based evaluation of the aforementioned methods has been conducted on a Flickr photo collection of over $100,000$ photos from 10 well-known touristic destinations in Greece.

In the speech analysis domain, work has been conducted advancing and evaluating techniques presented in the previous deliverables of WP2 (e.g. D2.1.1, D2.1.2 and D2.3), in order to take advantage of the media fusion results of the workpackage. Last but not least, work has been performed and presented herein in the textual analysis domain, in order to decrease the danger of the developed text analysis modules to adversely affect their analysis performance, as the WKI user selects the incorrect or imprecise locations and tags which are offered by the system., a fact potentially more likely to occur in situations where the user has limited time.

9 References

[1] MPEG-7 Visual Experimentation Model (XM). Version 10.0, ISO/IEC/JTC1/SC29/WG11, Doc. N4062, Mar., 2001.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD Int'l Conference on Management of Data, Seattle, Washington*, pages 94–105. ACM Press, June 1998.

[3] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 11:5–33, 2005.

[4] M. Aurnhammer, P. Hanappe, and L. Steels. Augmenting navigation for collaborative tagging with emergent semantics. In *International Semantic Web Conference*, 2006.

[5] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *in Proceedings of ACM Multimedia (Full paper) (MM 2010)*, Firenze, Italy, October 2010.

[6] Y. Avrithis, G. Tolias, and Y. Kalantidis. Feature map hashing: Sub-linear indexing of appearance and global geometry. In *in Proceedings of ACM Multimedia (Full paper) (MM 2010)*, Firenze, Italy, October 2010.

[7] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *European Conference Computer Vision*. Springer, 2006.

[8] Hila Becker, Mor Naaman, and Luis Gravano. Event identification in social media. In *12th International Workshop on the Web and Databases, WebDB*, 2009.

[9] Hila Becker, Mor Naaman, and Luis Gravano. Learning similarity metrics for event identification in social media. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 291–300, New York, NY, USA, 2010. ACM.

[10] Christian Blum. Ant colony optimization: Introduction and recent trends. *Physics of Life Reviews*, 2:353–373, 2005.

[11] Lukas Burget, Petr Schwarz, Pavel Matejka, Mirko Hannemann, Ariya Rastrow, Christopher M. White, Sanjeev Khudanpur, Hynek Hermansky, and Jan Cernocky. Combination of strongly and weakly constrained recognizers for reliable detection of oovs. In *Proc. ICASSP*, pages 4081–4084, 2008.

[12] Gianni Di Caro, Frederick Ducatelle, and Luca Maria Gambardella. Anthocnet: an adaptive nature-inspired algorithm for routing in mobile ad hoc networks. *European Transactions on Telecommunications*, 16(5):443–455, 2005.

[13] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[14] Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 84–93, New York, NY, USA, 1999. ACM.

[15] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–9, New York, NY, USA, 2009. ACM.

[16] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *WWW '09*, pages 761–770, New York, NY, USA, 2009. ACM.

[17] Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu. Constructing travel itineraries from tagged geo-temporal breadcrumbs. In *WWW '10*, pages 1083–1084, New York, NY, USA, 2010. ACM.

[18] Sabine Deligne and Frederic Bimbot. Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In *Proc. ICASSP*, pages 169–172, 1995.

[19] WeKnowIt Project Deliverable. D2.2: Contextual media analysis and fusion techniques.

[20] Carlotta Domeniconi and Muna Al-Razgan. Weighted cluster ensembles: Methods and analysis. *ACM Transactions on Knowledge Discovery from Data*, 2:17:1–17:40, January 2009.

[21] M. Dorigo. *Optimization, Learning and Natural Algorithms*. PhD thesis, Politecnico di Milano, Italy, 1992.

[22] Marco Dorigo and Gianni Di Caro. The ant colony optimization meta-heuristic, 1999.

[23] Yan-Tao Zheng et al. Tour the world: building a web-scale landmark recognition engine. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June, 2009.

[24] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[25] Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab. Triplerank: Ranking semantic web data by tensor decomposition. In *ISWC '09: Proceedings of the 8th International Semantic Web Conference*, pages 213–228, Berlin, Heidelberg, 2009. Springer-Verlag.

[26] Eirini Giannakidou, Ioannis Kompatsiaris, and Athena Vakali. Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems. In *ICSC*, pages 128–135, 2008.

[27] Eirini Giannakidou, Vassiliki A. Koutsonikola, Athena Vakali, and Yiannis Kompatsiaris. Co-clustering tags and social data sources. In *WAIM*, pages 317–324, 2008.

[28] David Graff and Christopher Cieri. English GigaWord, 2003. Linguistic Data Consortium.

[29] Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, David Leeuwen, Mike Lincoln, and Vincent Wan. Multimodal technologies for perception of humans. pages 414–428. Springer-Verlag, 2008.

[30] Mirko Hannemann, Stefan Kombrink, Martin Karafiat, and Lukas Burget. Similarity scoring for recognizing repeated out-of-vocabulary words. In *Proc. Interspeech*, pages 897–900, 2010.

[31] Richard A. Harshman and Margaret E. Lundy. Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39 – 72, 1994.

[32] Adam Herout, Pavel Zemcik, Michal Hradis, Roman Juranek, Jiri Havel, Radovan Josth, and Martin Zadnik. *Low-Level Image Features for Real-Time Object Detection*, pages 111–136. IN-TECH Education and Publishing, 2010.

[33] Thomas Hofmann. Unsupervised learning from dyadic data. pages 466–472. MIT Press, 1998.

[34] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[35] Michal Hradis, Vitezslav Beran, Ivo Reznicek, Adam Herout, David Barina, Adam Vlcek, and Pavel Zemcik. Brno University of Technology at TRECVid 2010. In *TRECVid 2010: Participant Notebook Papers and Slides*. NIST, 2010.

[36] Neil Ireson and Fabio Ciravegna. Toponym resolution in social media. In P.F. Patel-Schneider et al., editor, *Proceedings of the 9th International Semantic Web Conference (ISWC2010), LNCS 6496, Part 1*, pages 370–385, Shanghai, China, 7-11 November 2010. Springer, Heidelberg.

[37] Saral Jain, Stephan Seufert, and Srikanta Bedathur. Antourage: mining distance-constrained trips from flickr. In *WWW '10*, pages 1121–1122, New York, NY, USA, 2010. ACM.

[38] Maciej Janik, Symeon Papadopoulos, and Börkur Sigurbjörnsson. D3.3 mass classification and clustering. Technical report, WeKnowIt, 2010. http://www.weknowit.eu/sites/default/files/D3.3.pdf.

[39] H. Jegou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, pages 1–21, 2010.

[40] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias. Viral: Visual image retrieval and localization. *Multimedia Tools and Applications*, 2011.

[41] Lyndon Kennedy and Mor Naaman. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 311–320, New York, NY, USA, 2009. ACM.

[42] Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *WWW '08*, pages 297–306, NY, USA, 2008. ACM.

[43] Lyndon S. Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM Multimedia*, pages 631–640, 2007.

[44] Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. The sketch engine. In T. Fontenelle, editor, *Practical Lexicography: A Reader*, pages 297–306. Oxford University Press, U.K., 2008.

[45] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009.

[46] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[47] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, 1966.

[48] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. Multimedia content processing through cross-modal association. In *MULTIMEDIA '03*, pages 604–611, New York, USA, 2003. ACM.

[49] Rainer Lienhart, Stefan Romberg, and Eva Hörster. Multilayer plsa for multimodal image retrieval. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, New York, NY, USA, 2009. ACM.

[50] Stefanie Lindstaedt, Viktoria Pammer, Roland Mörzinger, Roman Kern, Helmut Mülner, and Claudia Wagner. Recommending tags for pictures based on text, visual content and user context. In *Proceedings of the 2008 Third International Conference on Internet and Web Applications and Services*, pages 506–511, Washington, DC, USA, 2008. IEEE Computer Society.

[51] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

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

[53] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[54] Joao Magalhaes and Stefan Rüger. Information-theoretic semantic multimedia indexing. In *CIVR '07*, pages 619–626, New York, USA, 2007. ACM.

[55] B. S. Manjunath, J. R. Ohm, V. V. Vinod, and A. Yamada. Colour and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, 11(6):703–715, Jun 2001.

[56] V. Mihajlovic, M. Petkovic, W. Jonker, and H.M. Blanken. Multimodal content-based video retrieval. In H.M. Blanken, A.P. de Vries, H.E. Blok, and L. Feng, editors, *Multimedia Retrieval*, Data-Centric Systems and Applications, pages 271–294. Springer Verlag, Berlin, 2007.

[57] M. Muja and D.G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *ICCV*, 2009.

[58] Milind R. Naphade and Thomas S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.

[59] Spiros Nikolopoulos, Eirini Giannakidou, Ioannis Kompatsiaris, Ioannis Patras, and Athena Vakali. *Combining multi-modal features for social media analysis*, volume VIII of *Social Media Modeling and Computing*, page 276. Springer, 1st edition, 2011.

[60] Ximena Olivares, Massimiliano Ciaramita, and Roelof van Zwol. Boosting image retrieval through aggregating search results based on visual annotations. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 189–198, New York, NY, USA, 2008. ACM.

[61] Symeon Papadopoulos, Christos Zigkolis, Yiannis Kompatsiaris, and Athena Vakali. Cluster-based landmark and event detection on tagged photo collections. *Multimedia, IEEE*, 2010.

[62] Symeon Papadopoulos, Christos Zigkolis, Giorgos Tolias, Yannis Kalantidis, Phivos Mylonas, Yiannis Kompatsiaris, and Athena Vakali. Image clustering through community detection on hybrid image similarity graphs. In *ICIP*, pages 2353–2356, 2010.

[63] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl.*, 6:90–105, June 2004.

[64] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision Pattern Recognition*, 2007.

[65] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition*, 2008.

[66] T. Piatrik and E. Izquierdo. Subspace clustering of images using ant colony optimisation. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 229 –232, 2009.

[67] Adrian Popescu, Gregory Grefenstette, and Pierre-Alain Moëllic. Mining tourist information from user-supplied collections. In *CIKM '09*, pages 1713–1716, New York, NY, USA, 2009. ACM.

[68] Till Quack, Bastian Leibe, and Luc J. Van Gool. World-scale mining of objects and events from community photo collections. In *CIVR*, pages 47–56, 2008.

[69] Till Quack, Bastian Leibe, and Luc Van Gool. World-scale mining of objects and events from community photo collections. In *CIVR '08*, pages 47–56, New York, NY, USA, 2008. ACM.

[70] Yuheng Ren, Mo Yu, Xin-Jing Wang, Lei Zhang, and Wei-Ying Ma. Diversifying landmark image search results by learning interested views from community photos. In *WWW '10*, pages 1289–1292, New York, NY, USA, 2010. ACM.

[71] Ansgar Scherp, Symeon Papadopoulos, Börkur Sigurbjörnsson, Rabeeh Abbasi, Sergej Sizov, and Eirini Giannakidou. D3.4 report on tools and methods for mass evolution analysis. Technical report, WeKnowIt, 2011. http://www.weknowit.eu/sites/default/files/D3.4.pdf.

[72] Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 327–336, New York, NY, USA, 2008. ACM.

[73] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1470, Washington, DC, USA, 2003. IEEE Computer Society.

[74] Sergej Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 281–290, New York, NY, USA, 2010. ACM.

[75] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Multi-modal classifier fusion for video shot content retrieval. In *In Proceedings of the 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, April 2005.

[76] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. Tag recommendations based on tensor dimensionality reduction. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 43–50, New York, NY, USA, 2008. ACM.

[77] Igor Szoke, Michal Fapso, Lukas Burget, and Jan Cernocky. Hybrid word-subword decoding for spoken term detection. In *Proc. SSCS 2008, Speech search workshop at SIGIR*, 2008.

[78] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010. 10.1007/978-0-387-09823-4$_3$4.

[79] Koen E.A. van de Sande, Theo Gevers, and Cees G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. on Pattern Analysis and Mach. Intelligence*, 99, 2009.

[80] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. Optimal multimodal fusion for multimedia data analysis. In *MULTIMEDIA '04*, pages 572–579, New York, USA, 2004. ACM.

[81] Rui Xu and II Wunsch. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[82] Christos Zigkolis, Symeon Papadopoulos, Yiannis Kompatsiaris, and Athena Vakali. Detecting the long tail of points of interest in tagged photo collections. In *CBMI*, 2011.