Overview of the WikipediaMM Task at ImageCLEF 2009

Theodora Tsikrika[1] and Jana Kludas[2]

[1] CWI, Amsterdam, The Netherlands
Theodora.Tsikrika@cwi.nl
[2] CUI, University of Geneva, Switzerland
jana.kludas@unige.ch

Abstract. ImageCLEF's wikipediaMM task provides a testbed for the system-oriented evaluation of multimedia information retrieval from a collection of Wikipedia images. The aim is to investigate retrieval approaches in the context of a large and heterogeneous collection of images (similar to those encountered on the Web) that are searched for by users with diverse information needs. This paper presents an overview of the resources, topics, and assessments of the wikipediaMM task at ImageCLEF 2009, summarises the retrieval approaches employed by the participating groups, and provides an analysis of the main evaluation results.

1 Introduction

The wikipediaMM task is an ad-hoc image retrieval task. The evaluation scenario is thereby similar to the classic TREC ad-hoc retrieval task and the ImageCLEF photo retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (i.e., topics are not known to the system in advance). Given a multimedia query that consists of a title and one or more sample images describing a user's multimedia information need, the aim is to find as many relevant images as possible from the (INEX MM) wikipedia image collection. A multi-modal retrieval approach in that case should be able to combine the relevance of different media types into a single ranking that is presented to the user.

The wikipediaMM task differs from other benchmarks in multimedia information retrieval, like TRECVID, in the sense that the textual modality in the wikipedia image collection contains less noise than the speech transcripts in TRECVID. Maybe that is one of the reasons why, both in last year's task and in INEX Multimedia 2006-2007 (where this image collection was also used), it has proven challenging to outperform the text-only approaches. This year, the aim is to promote the investigation of multi-modal approaches to the forefront of this task by providing a number of resources to support the participants towards this research direction.

The paper is organised as follows. First, we introduce the task's resources: the wikipedia image collection and additional resources, the topics, and the assessments (Sections 2–4). Section 5 presents the approaches employed by the

participating groups and Section 6 summarises their main results. Section 7 concludes the paper.

2 Task resources

The resources used for the wikipediaMM task are based on Wikipedia data. The collection is the **(INEX MM) wikipedia image collection**, which consists of 151,519 JPEG and PNG Wikipedia images provided by Wikipedia users. Each image is associated with user-generated alphanumeric, unstructured metadata in English. These metadata usually contain a brief caption or description of the image, the Wikipedia user who uploaded the image, and the copyright information. These descriptions are highly heterogeneous and of varying length. Further information about the image collection can be found in [13].

Fig. 1: Wikipedia image+metadata example from the wikipedia image collection.

Additional resources were also provided to support the participants in their investigations of multi-modal approaches. These resources are:

Image similarity matrix: The similarity matrix for the images in the collection has been constructed by the IMEDIA group at INRIA. For each image in the collection, this matrix contains the list of the top $K = 1000$ most similar images in the collection together with their similarity scores. The same is given for each image in the topics. The similarity scores are based on the distance between images; therefore, the lower the score, the more similar

the images. Further details on the features and distance metric used can be found in [2].

Image classification scores: For each image, the classification scores for the 101 MediaMill concepts have been provided by UvA [11]. The UvA classifier is trained on manually annotated TRECVID video data for concepts selected for the broadcast news domain.

Image features: For each image, the set of the 120D feature vectors that has been used to derive the above image classification scores [3] has also been made available. Participants can use these feature vectors to custom-build a content-based image retrieval (CBIR) system, without having to pre-process the image collection.

The additional resources are beneficial to researchers who wish to exploit visual evidence without performing image analysis. Of course, participants could also extract their own image features.

3 Topics

The topics are descriptions of multimedia information needs that contain textual and visual hints.

3.1 Topic Format

These multimedia queries consist of a textual part, the query title, and a visual part, one or several example images.

<**title**> query by keywords
<**image**> query by image content (one or several)
<**narrative**> description of query in which the definitive definition of relevance and irrelevance are given

<title> The topic <title> simulates a user who does not have (or want to use) example images or other visual constraints. The query expressed in the topic <title> is therefore a text-only query. This profile is likely to fit most users searching digital libraries.

Upon discovering that a text-only query does not produce many relevant hits, a user might decide to add visual hints and formulate a multimedia query.

<image> The visual hints are example images, which can be taken from outside or inside the wikipedia image collection and can be of any common format. Each topic has at least one example image, but it can have several, e.g., to describe the visual diversity of the topic.

<narrative> A clear and precise description of the information need is required in order to unambiguously determine whether or not a given document fulfils the given information need. In a test collection this description is known as the narrative. It is the only true and accurate interpretation of a user's needs. Precise recording of the narrative is important for scientific repeatability - there must exist, somewhere, a definitive description of what is and is not relevant to the user. To aid this, the <narrative> should explain not only what information is being sought, but also the context and motivation of the information need, i.e., why the information is being sought and what work-task it might help to solve.

These different types of information sources (textual terms and visual examples) can be used in any combination. It is up to the systems how to use, combine or ignore this information; the relevance of a result does not directly depend on these constraints, but it is decided by manual assessments based on the <narrative>.

3.2 Topic Development

The topics in the ImageCLEF 2009 wikipediaMM task have been partly developed by the participants and partly by the organisers. This year the participation in the topic development process was not obligatory, so only 2 of the participating groups submitted a total of 11 candidate topics. The rest of the candidate topics were created by the organisers with the help of the log of an image search engine. After a selection process performed by the organisers, a final list of 45 topics was created.

These final topics are listed in Table 1 and range from simple, and thus relatively easy (e.g., "bikes"), to semantic, and hence highly difficult (e.g., "aerial photos of non-artificial landscapes"), with the latter forming the bulk of the topics. Semantic topics typically have a complex set of constraints, need world knowledge, and/or contain ambiguous terms, so they are expected to be challenging for current state-of-the-art retrieval algorithms. We encouraged the participants to use multi-modal approaches since they are more appropriate for dealing with semantic information needs. On average, the 45 topics contain 1.7 images and 2.7 words.

4 Assessments

The wikipediaMM task is an image retrieval task, where an image with its metadata is either relevant or not (binary relevance). We adopted TREC-style pooling of the retrieved images with a pool depth of 50, resulting in pools of between 299 and 802 images with a mean and median both around 545. The evaluation was performed by the participants of the task within a period of 4 weeks after the submission of runs. The 7 groups that participated in the evaluation process used the web-based interface that was used last year and which has also been previously employed in the INEX Multimedia and TREC Enterprise tracks.

Table 1: Topics for the ImageCLEF 2009 wikipediaMM task: IDs, titles, the number of image examples providing additional visual information, and the number of relevant images in the collection.

ID	Topic title	# image examples	# relevant images
76	shopping in a market	3	31
77	real rainbow	1	12
78	sculpture of an animal	3	32
79	stamp without human face	3	89
80	orthodox icons with Jesus	2	28
81	sculptures of Greek mythological figures	3	30
82	rider on horse	2	53
83	old advertisement for cars	2	31
84	advertisement on buses	2	41
85	aerial photos of non-artificial landscapes	2	37
86	situation after hurricane katrina	2	5
87	airplane crash	2	12
88	madonna portrait	2	29
89	people laughing	3	12
90	satellite image of river	1	60
91	landline telephone	1	13
92	bikes	1	30
93	close up of antenna	2	21
94	people with dogs	2	52
95	group of dogs	2	39
96	cartoon with a cat	1	53
97	woman in pink dress	2	12
98	close up of people doing sport	3	37
99	flowers on trees	2	32
100	flower painting	2	18
101	fire	2	74
102	building site	1	6
103	palm trees	1	41
104	street musician	2	20
105	snowy street	2	31
106	traffic signs	2	32
107	red fruit	2	38
108	bird nest	2	21
109	tennis player on court	2	29
110	desert landscape	2	35
111	political campaign poster	2	19
112	hot air balloons	1	13
113	baby	1	23
114	street view at night	2	95
115	notes on music sheet	1	112
116	illustration of engines	1	40
117	earth from space	2	35
118	coral reef underwater	2	24
119	harbor	2	63
120	yellow flower	2	62

5 Participants

A total of 8 groups submitted 57 runs: CEA (LIC2M-CEA, Centre CEA de Saclay, France), DCU (Dublin City University, School of Computing, Ireland), DEUCENG (Dokuz Eylul University, Department of Computer Engineering, Turkey), IIIT-Hyderabad (Search and Info Extraction Lab, India), LaHC (Laboratoire Hubert Curien, UMR CNRS, France), SZTAKI (Hungarian Academy of Science, Hungary), SINAI (Intelligent Systems, University of Jaen, Spain) and UALICANTE (Software and Computer Systems, University of Alicante, Spain).

Table 2: Types of the 57 submitted runs.

Run type	**# runs**
Text (TXT)	26
Visual (IMG)	2
Text/Visual (TXTIMG)	29
Query Expansion	18
Relevance Feedback	7

Table 2 gives an overview of the types of the submitted runs. This year more multi-modal (text/visual) than text-only runs were submitted. A short description of the participants' approaches follows.

CEA (12 runs) [**8**] They extended the approach they employed last year by refining the textual query expansion procedure and introducing of a k-NN based visual reranking procedure. Their main aim was to examine whether combining textual and content-based retrieval improves over purely textual search.

DCU (5 runs) [**6**] Their main effort concerned the expansion of the image metadata using the Wikipedia abstracts' collection DBpedia. Since the metadata is short for retrieval by query text, they expand the query and documents using the Rocchio algorithm. For retrieval, they used the LEMUR toolkit. They also submitted one visual run.

DEUCENG (6 runs) [**4**] Their research interests focussed on 1) the expansion of native documents and queries, term phrase selection based on WordNet, WSD and WordNet similarity functions, and 2) a new reranking approach with Boolean retrieval and C3M based clustering.

IIT-H (1 run) [**12**] Their system automatically ranks the most similar images to a given textual query using a combination of the Vector Space Model and the Boolean model. The system preprocesses the data set in order to remove the non-informative terms.

LaHC (13 runs) [**7**] In this second participation, they extended their approach (a multimedia document model defined as a vector of textual and visual terms weighted using tf.idf) by using 1) additional information for the textual part (legend and image bounding text extracted from the original docu-

ments), 2) different image detectors and descriptors, and 3) a new text/image combination approach.

SINAI (4 runs) [5] Their approach focussed on query and document expansion techniques based on WordNet. They used the LEMUR toolkit as their retrieval system.

SZTAKI (7 runs) [1] They used both textual and visual features and employed image segmentation, SIFT keypoints, Okapi BM25 based text retrieval, and query expansion by an online thesaurus. They preprocessed the annotation text to remove author and copyright information and biased retrieval towards images with filenames containing relevant terms.

UALICANTE (9 runs) [9] They used IR-n, a retrieval system based on passages and applied two different term selection strategies for query expansion: Probabilistic Relevance Feedback and Local Context Analysis, and their multi-modal versions. They also used the same technique for Camel Case decompounding of image filenames that they used in last year's participation.

6 Results

Table 3 presents the evaluation results for the 15 best performing runs ranked by Mean Average Precision (MAP). DEUCENG's text-only runs performed best. But as already seen last year, approaches that fuse several modalities can compete with the text-only ones. Furthermore, it is notable that all participants that used both mono-media and multi-modal algorithms achieved their best results with their multi-modal runs. The complete list of results can be found at the ImageCLEF website `http://www.imageclef.org/2009/wikiMM-results`.

Table 3: Results for the top 15 runs.

	Participant	Run	Modality	FB/QE	MAP	P@10	P@20	R-prec.
1	deuceng	deuwiki2009_205	TXT	QE	0.2397	0.4000	0.3133	0.2683
2	deuceng	deuwiki2009_204	TXT	QE	0.2375	0.4000	0.3111	0.2692
3	deuceng	deuwiki2009_202	TXT	QE	0.2358	0.3933	0.3189	0.2708
4	lahc	TXTIMG_100_3_1_5_meanstd	TXTIMG	NOFB	0.2178	0.3378	0.2811	0.2538
5	lahc	TXTIMG_50_3_1_5_meanstd	TXTIMG	NOFB	0.2148	0.3356	0.2867	0.2536
6	cea	cealateblock	TXTIMG	QE	0.2051	0.3622	0.2744	0.2388
7	cea	ceaearlyblock	TXTIMG	QE	0.2046	0.3556	0.2833	0.2439
8	cea	ceabofblock	TXTIMG	QE	0.1975	0.3689	0.2789	0.2342
9	cea	ceatlepblock	TXTIMG	QE	0.1959	0.3467	0.2733	0.2236
10	cea	ceabofblockres	TXTIMG	QE	0.1949	0.3689	0.2789	0.2357
11	cea	ceatlepblockres	TXTIMG	QE	0.1934	0.3467	0.2733	0.2236
12	lahc	TXTIMG_Siftdense_0.084	TXTIMG	NOFB	0.1903	0.3111	0.2700	0.2324
13	lahc	TXT_100_3_1_5	TXT	NOFB	0.1890	0.2956	0.2544	0.2179
14	lahc	TXT_50_3_1_5	TXT	NOFB	0.1880	0.3000	0.2489	0.2145
15	ualicante	Alicante-MMLCA	TXTIMG	FB	0.1878	0.2733	0.2478	0.2138

Next, we analyse the evaluation results. In our analysis, we use only the top 90% of the runs to exclude noisy and buggy results. Furthermore, we excluded 3 runs that we considered to be redundant, i.e., they were produced by the same group and achieved the exact same result, so as to reduce the bias of the analysis.

6.1 Performance per modality for all topics

Table 4 shows the average performance and standard deviation with respect to modality. On average, the multi-modal runs manage to outperform the mono-media runs with respect to all examined evaluation metrics (MAP, Precison at 20, and precision after R (= number of relevant) documents are retrieved).

Table 4: Results per modality over all topics.

Modality	**MAP**		**P@20**		**R-prec.**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
All top 90% runs (46 runs)	0.1751	0.0302	0.2356	0.0624	0.2076	0.0572
TXT in top 90% runs (23 runs)	0.1726	0.0326	0.2278	0.0427	0.2038	0.0328
TXTIMG in top 90% runs (23 runs)	0.1775	0.0281	0.2433	0.0364	0.2115	0.0307

6.2 Performance per topic and per modality

To analyse the average difficulty of the topics, we classify the topics based on the AP values per topic averaged over all runs as follows:

easy: $MAP > 0.3$
medium: $0.2 < MAP <= 0.3$
hard: $0.1 < MAP <= 0.2$
very hard: $MAP < 0.1$.

Table 5 presents the top 7 topics per class (i.e., easy, medium, hard, and very hard), together with the total number of topics per class. Most of the topics are considered to be hard. This was actually intended during the topic development process where we opted for highly semantic topics that are challenging for current retrieval approaches. Nonetheless, 10 out of 45 topics were of easy and medium difficulty. Only 7 topics were very hard to solve. Therein, topics #97 "woman in pink dress" and #98 "close up of people doing sport" can be considered as unsolvable, since their $MAP < 0.05$.

Table 5: Topics classified based on their difficulty. The top 7 topics are shown per class together with the total number of topics per class.

easy (6 topics)	**medium (4 topics)**	**hard (28 topics)**	**very hard (7 topics)**
112 hot air balloons	**118** coral reef underwater	**120** yellow flower	**105** snowy street
88 madonna portrait	**90** satellite image of river	**91** landline telephone	**78** sculpture of an animal
80 orthodox icons	**110** desert landscape	**99** flowers on trees	**117** earth from space
108 bird nest	**77** real rainbow	**79** stamp human face	**85** aerial ph. of landscapes
103 palm trees		**107** red fruit	**89** people laughing
93 close up antenna		**94** people with dogs	**97** woman in pink dress
			98 close up of people doing sport

We also analysed the performance of runs that use only text (TXT) versus runs that use both text and visual resources (TXTIMG). Figure 2 shows the average performance on each topic for all, text-only, and text-visual runs. The text-based runs outperform the text-visual ones in 22 out of the 45, indicating that slightly more than half of the topics benefit from a multi-modal approach.

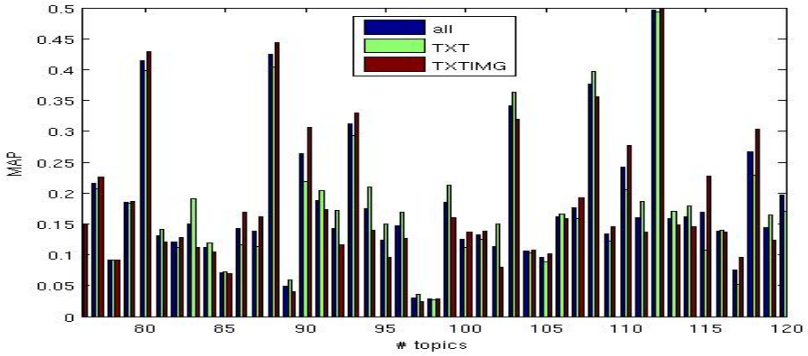

Fig. 2: Average topic performance over all, text-only, and text/visual runs.

6.3 Visuality of topics

The "visuality" of topics can be deduced from the performance of text-only and text-visual approaches that we presented in the last section. We consider that if, for a topic, the text-visual approaches improve significantly the MAP over all runs (i.e., by $diff(MAP) >= 0.01$), then we could consider that to be a visual topic. In the same way, we can define topics as textual, if the text-only approaches improve significantly the MAP over all runs of a topic. Based on this analysis, 15 of the topics can be characterised as textual and 14 as visual. The remaining 16 topics, where no clear improvements are observed, are considered to be neutral.

Table 6 presents the topics in each group, as well as some statistics on the topic, their relevant documents, and their distribution over the classes that indicate their difficulty. As expected, visual topics have more image examples per topic (#images/topic) than textual ones (1.66 vs. 1.85); however, the neutral topics have an even higher average of 2.06 images per topic. The same tendency is observed in the average number of words in the topics (#words/topic). Short titled topics are better solved with text-only approaches, topics with longer titles tend to be visual or neutral. Therefore, it appears that the latter two groups contain the more complex/semantic topics. The distribution of the textual, visual, and neutral topics over the classes expressing their difficulty shows that the visual topics are more likely to fall into the easy/medium class than the textual or

neutral ones. The neutral topics seem to contain in general very difficult topics, where neither the text-only approaches nor the text-visual ones could achieve good retrieval results.

Table 6: Best performing topics for textual and text-visual runs relative to the average over all runs.

	textual (15 topics)	**visual (14 topics)**	**neutral (16 topics)**
Topics	**83** advertisement for cars	**115** notes on music sheet	**76** shopping on a market
	102 building site	**90** satellite image of river	**77** real rainbow
	94 people with dogs	**118** coral reef underwater	**78** sculpture of an animal
	92 bikes	**110** desert landscape	**79** stamp without human face
	95 group of dogs	**120** yellow flower	**81** sculptures of Greek mythological figures
	99 flowers on trees	**86** situation after katrina	**82** rider on horse
	111 pol. campaign poster	**87** airplane crash	**84** advertisement on buses
	103 palm trees	**117** earth from space	**85** aerial photos of non-artificial landscapes
	96 cartoon with a cat	**88** madonna portrait	**97** woman in pink dress
	119 harbor	**93** close up of antenna	**98** close up of people doing sport
	108 bird nest	**107** red fruit	**101** fire
	114 street view at night	**80** orthodox icons with Jesus	**104** street musician
	91 landline telephone	**100** flower painting	**105** snowy street
	113 baby	**109** tennis player on court	**106** traffic signs
	89 people laughing		**112** hot air balloons
			116 illustration of engines
#images/topic	1.66	1.85	2.06
#words/topic	2.53	3.00	3.31
#reldocs	35.33	36.28	36.50
#words/reldocs	29.65	44.99	39.24
easy	2	3	1
medium	0	3	1
hard	12	7	9
very hard	1	1	5

6.4 Effect of Query Expansion and Relevance Feedback

Finally, we analyse the effect of the application of query expansion (QE) and relevance feedback (FB) techniques. Similarly to the analysis in the previous section, we consider the techniques to be useful for a topic, if they improved significantly the MAP over all runs. Table 7 presents the best performing topics for these techniques and some statistics. Query expansion is useful for 17 topics and relevance feedback for 11. The statistics show that these techniques can help improve the retrieval results for topics defined without too much detail, e.g., topics having a short title (#words/topic) and/or a small number of example images (#images/topic).

Table 7: Best performing topics for textual and text-visual runs relative to the average over all runs.

	QE (17 topics)	FB (11 topics)
Topics	**110** desert landscape	**88** madonna portrait
	118 coral reef underwater	**115** notes on music sheet
	120 yellow flower	**87** airplane crash
	109 tennis player on court	**93** close up of antenna
	92 bikes	**96** cartoon with a cat
	82 rider on horse	**79** stamp without human face
	101 fire	**116** illustration of engines
	115 notes on music sheet	**118** coral reef underwater
	117 earth from space	**95** group of dogs
	119 harbor	**104** street musician
	112 hot air balloons	**86** situation after hurricane katrina
	98 close up of people doing sport	
	113 baby	
	107 red fruit	
	79 stamp without human face	
	84 advertisement on buses	
	78 sculpture of an animal	
#images/topic	1.94	1.72
#words/topic	2.76	3.18
#reldocs	46.47	40.36
#words/reldocs	37.98	42.74
easy	1	2
medium	2	1
hard	11	8
very hard	3	0

7 Conclusions

This year (similarly to 2008), a text-based approach performed best in the wikipediaMM task, even though highly semantic multimedia topics were developed with the aim to encourage and show the potential of multi-modal approaches. It is worth noting though that all of the participants that submitted both mono-media and multi-modal runs achieved their best results with their multi-modal runs. Additionally, it is encouraging to see more than half of the submitted runs being multi-modal.

In 2010, a new collection of approximately 250,000 Wikipedia images will be provided with multi-lingual text annotations in English, French, and German.

8 Acknowledgements

Theodora Tsikrika was supported by the European Union via the European Commission project VITALAS (contract no. 045389). Jana Kludas was funded by the Swiss National Fund (SNF). The authors would also like to thank all the groups participating in the relevance assessment process.

References

1. Bálint Daróczy, István Petrás, András A. Benczúr, Zsolt Fekete, Dávid Nemeskey, Dávid Siklósi, and Zsuzsa Weiner. SZTAKI @ ImageCLEF 2009. In *CLEF 2009 working notes*, 2009.

2. Marin Ferecatu. Image retrieval with active relevance feedback using both visual and keyword-based descriptors. In *Ph.D. Thesis, Université de Versailles, France*, 2005.
3. Jan C. van Gemert, Jan-Mark Geusebroek, Cor J. Veenman, Cees G. M. Snoek, and Arnold W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 105, Washington, DC, USA, 2006. IEEE Computer Society.
4. Deniz Kilinc and Adil Alpkocak. DEU at ImageCLEF 2009 WikipediaMM task: Experiments with expansion and reranking approaches. In *CLEF 2009 working notes*, 2009.
5. M.T. Martín-Valdivia M.C. Díaz-Galiano, L.A. Urena-López, and J.M. Perea-Ortega. Using WordNet in multimedia information retrieval. In Peters et al. [10].
6. Jinming Min, Peter Wilkins, Johannes Leveling, and Gareth J. F. Jones. Document expansion for text-based image retrieval at CLEF 2009. In Peters et al. [10].
7. Christophe Moulin, Cécile Barat, Cédric Lemaître, Mathias Géry, Christophe Ducottet, and Christine Largeron. Combining text/image in WikipediaMM task 2009. In Peters et al. [10].
8. Débora Myoupo, Adrian Popescu, Hervé Le Borgne, and Pierre-Alain Moëllic. Multimodal image retrieval over a large database. In Peters et al. [10].
9. Sergio Navarro, Rafael Munoz, and Fernando Llopis. Evaluating fusion techniques at different domains at ImageCLEF subtasks. In *CLEF 2009 working notes*, 2009.
10. Carol Peters, Theodora Tsikrika, Henning Müller, Jayashree Kalpathy-Cramer, Gareth J.F.Jones, Julio Gonzalo, and Barbara Caputo, editors. *Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009)*, Lecture Notes in Computer Science. Springer, 2010.
11. Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM Press.
12. Srinivasarao Vundavalli. IIIT-H at ImageCLEF Wikipedia MM 2009. In *CLEF 2009 working notes*, 2009.
13. Thijs Westerveld and Roelof van Zwol. The INEX 2006 multimedia track. In Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors, *Advances in XML Information Retrieval: 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Revised Selected Papers*, volume 4518, pages 331–344. Springer, 2007.