

Henning Müller
Paul Clough
Thomas Deselaers
Barbara Caputo (Eds.)

ImageCLEF

Experimental Evaluation
in Visual Information Retrieval

 Springer

INRE

Series Editor

W. Bruce Croft

Editorial Board

ChengXiang Zhai

Maarten de Rijke

Nicholas J. Belkin

Charles Clarke

Henning Müller • Paul Clough • Thomas Deselaers •
Barbara Caputo
Editors

ImageCLEF

Experimental Evaluation
in Visual Information Retrieval

 Springer

Editors

Henning Müller
Business Information Systems
University of Applied Sciences
Western Switzerland (HES–SO)
TechnoArk 3
3960 Sierre
Switzerland
henning.mueller@hevs.ch

Paul Clough
Information School
University of Sheffield
Regent Court
Sheffield S1 4DP
England
p.d.clough@sheffield.ac.uk

Thomas Deselaers
ETH Zürich
Computer Vision Lab/ETF-C 113.2
Zürich
Switzerland
deselaers@vision.ee.ethz.ch

Barbara Caputo
Idiap Research Institute
rue Marconi 19
1920 Martigny
Switzerland
bcaputo@idiap.ch

ISSN 1387-5264 The Information Retrieval Series
ISBN 978-3-642-15180-4 e-ISBN 978-3-642-15181-1
DOI 10.1007/978-3-642-15181-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010934959

ACM Classification (1998): H.3, I.4, C.4, J.3

© Springer-Verlag Berlin Heidelberg 2010

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: KunkelLopka GmbH, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*This book is dedicated to our families and the
love, support and encouragement they have
given us.*

Foreword

The pervasive creation and consumption of content, especially visual content, is ingrained into our modern world. We're constantly consuming visual media content, in printed form and in digital form, in work and in leisure pursuits. Like our cave-man forefathers, we use pictures to record things which are of importance to us as memory cues for the future, but nowadays we also use pictures and images to document processes; we use them in engineering, in art, in science, in medicine, in entertainment and we also use images in advertising. Moreover, when images are in digital format, either scanned from an analogue format or more often than not born digital, we can use the power of our computing and networking to exploit images to great effect.

Most of the technical problems associated with creating, compressing, storing, transmitting, rendering and protecting image data are already solved. We use accepted standards and have tremendous infrastructure and the only outstanding challenges, apart from managing the scale issues associated with growth, are to do with locating images. That involves analysing them to determine their content, classifying them into related groupings, and searching for images. To overcome these challenges we currently rely on image metadata, the description of the images, either captured automatically at creation time or manually added afterwards. Meanwhile we push for developments in the area of *content-based* analysis, indexing and searching of visual media and this is where most of the research in image management is concentrated.

Automatic analysis of the content of images, which in turn would open the door to content-based indexing, classification and retrieval, is an inherently tough problem and because of the difficulty, progress is slow. Like all good science it cannot be rushed yet there is a frustration with the pace of its development because the rollout and development of other related components of image management, components such as capture, storage, transmission, rendering, etc., has been so rapid. We seem to be stuck on the problems of how to effectively find images when we are looking for them. While this is partly caused by the sheer number of images available to us, it is mostly caused by the scientific difficulty of the challenge and so it requires a

basic scientific approach to exploring the problem and finding solutions. As in all science, a fundamental aspect is measurement and benchmarking.

In any science, each new development, each approach, algorithm, model, idea or theory has to be measured in order to determine its worth and validity. That is how progress is made, and how fields advance. A theory is put forward and experiments to assess and measure the theory are carried out which may or may not support the theory and we advance the field, either by learning more about what works, or equally important we learn about what does not work. In the technology sector and in the information management area in particular, measuring the validity and worth of new ideas, approaches, etc., now takes place as part of organised benchmarking activities and there are many established examples. The Pascal Visual Object Class recognition challenge addresses recognising objects in images, TREC Vid addresses content analysis, retrieval and summarization from video, the KDD competition addresses data mining, and there have been others in machine learning for stock market prediction, shape retrieval, coin classification, text detection and reading, face verification, fingerprint verification, signature verification and of course the well-known NetFlix data mining and recommender competition. All these and many others take place against a backdrop of exploring new ideas, new approaches, and measuring their efficacy in a controlled environment. Which takes us to the present volume which covers ImageCLEF.

Cross-language image retrieval is a niche application domain within the broader area of managing image/visual media. Its importance is huge, though given the directions in which Internet growth is heading with multi-linguality and cross-language resources and processes growing ever more important. Henning Müller and Paul Clough have put together an impressive collection of contributions describing the formation, the growth, the resources, the various tasks and achievements of the ImageCLEF benchmarking activity, covering seven years of development in an annual cycle and involving contributions from hundreds of researchers from across the globe. This book could be described as a capstone volume which brings together all the contributions into one place, but a capstone is a finishing stone or a final achievement, and ImageCLEF continues today, as active as ever. With four parts which address the settings and logistics of ImageCLEF, the various track reports, some reports from participants and finally some external views, the volume is balanced and presents a comprehensive view of the importance and achievements of ImageCLEF towards advancing the field of cross-lingual image retrieval. It will remain an essential reference for anybody interested in how to start up and run a sizeable benchmarking activity, as well as an invaluable source of information on image retrieval in a cross-lingual setting.

Alan F. Smeaton

CLARITY: Centre for Sensor Web Technologies

Dublin City University

Dublin, Ireland, May 2010

Preface

This book contains a collection of texts centred on the evaluation of image retrieval systems. Evaluation, whether it be system-oriented or user-oriented, is an important part of developing effective retrieval systems that meet the actual needs of their end users. To enable reproducible evaluation requires creating standardised benchmarks and evaluation methodologies. This book highlights some of the issues and challenges in evaluating image retrieval systems and describes various initiatives that have sought to provide researchers with the necessary evaluation resources.

In particular the book summarises activities within ImageCLEF, an initiative to evaluate cross-language image retrieval systems that has been running as part of the Cross Language Evaluation Forum (CLEF) since 2003. ImageCLEF has provided resources, such as benchmarks, for evaluating image retrieval systems and complements a number of initiatives within the image retrieval research community, such as TRECvid for video retrieval, PASCAL for object recognition and detection and the many other smaller benchmarks, databases and tools available to researchers.

In addition to providing evaluation resources, ImageCLEF has also run within an annual evaluation cycle culminating in a workshop where participants have been able to present and discuss their ideas and techniques, forming a community with common interests and goals. Over the years ImageCLEF has seen participation from researchers within academic and commercial research groups worldwide, including those from Cross-Language Information Retrieval (CLIR), medical informatics, Content-Based Image Retrieval (CBIR), computer vision and user interaction.

This book comprises contributions from a range of people: those involved directly with ImageCLEF, such as the organisers of specific image retrieval or annotation tasks; participants who have developed techniques to tackle the challenges set forth by the organisers; and people from industry and academia involved with image retrieval and evaluation in general and beyond ImageCLEF. The book is structured into four parts:

- **Part I.** This section describes the context of ImageCLEF and the issues involved with developing evaluation resources, such as test collections and selecting evaluation measures. A focal point throughout ImageCLEF and across many of the

tasks has been to investigate how best to combine textual and visualisation information to improve information retrieval. Within the first section we summarise approaches explored within ImageCLEF over the years for this critical step in the retrieval process.

- **Part II.** This section includes seven chapters summarising the activities of each of the main tasks that have run within within ImageCLEF over the years. The track reports are written by those involved in co-ordinating ImageCLEF tasks and provide summaries of individual tasks, describe the participants and their approaches, and discuss some of the findings.
- **Part III.** This section is a selection of chapters by groups participating in various tasks within ImageCLEF 2009. Summaries of the techniques used for various domains such as retrieving diverse sets of photos from a collection of news photographs, multi-modal retrieval from online resources, such as Wikipedia, and retrieval and automatic annotation of medical images are presented. The chapters in this section show the variety and novelty of state-of-the-art techniques used to tackle various ImageCLEF tasks.
- **Part IV.** The final section provides an external perspective on the activities of ImageCLEF. These help to offer insights into the current and emerging needs for image retrieval and evaluation from both a commercial and research perspective. The final chapter helps to put ImageCLEF into the context of existing activities on evaluating multimedia retrieval techniques, providing thoughts on the future directions for evaluation over the coming years.

Sierre, Zürich, Martigny, Switzerland
Sheffield, UK
July 2010

Henning Müller
Paul Clough
Thomas Deselaers
Barbara Caputo

Acknowledgements

It is hard to know where to begin when considering who to thank because ImageCLEF has been very much a collaborative effort and involved a large number of people over the past seven years. We thank the UK Engineering and Physical Sciences Research Council (EPSRC) for funding the Eurovision project (grant number GR/R56778/01): a proposal written by Mark Sanderson that included running a large-scale benchmarking event for cross-language image retrieval that was to become ImageCLEF. In addition to the EPSRC, we thank all of the national and international funding bodies who have supported ImageCLEF in one way or another over the years. In particular we thank the European Union for support through the funding of various EU projects that have supported task organisers. We would like to acknowledge the following projects for supporting ImageCLEF activities:

- UK projects: Eurovision (EPSRC, grant GR/R56778/01)
- EU projects: MUSCLE, Multimatch, SemanticMining, DebutIT, Promise, TrebleCLEF, DELOS;
- Swiss projects: FNS (205321–109304/1), HES–SO for the BeMeVIS project;
- German projects: Theseus (grant 01MQ07017);
- companies such as Google (medical task) and LTU (for the object recognition task);
- American projects: NSF (grant ITR–0325160)

From the start ImageCLEF has been a collaborative activity and involved many people, many bringing new ideas, which is necessary in evaluation as it needs to advance as technology advances. People have given up their time and put tremendous effort into helping co-ordinate and run tasks. This has enabled us to produce reusable resources for various image retrieval communities. The following is hopefully a complete list of those involved in the organisation of ImageCLEF: Thomas Arni, Peter Dunker, Thomas M. Deserno, Julio Gonzalo, Michael Grubinger, Allan Hanbury, William Hersch, Mark Huiskes, Patric Hensfelt, Charles Kahn, Jayashree Kalpathy–Cramer, Jussi Karlgren, Jana Kludas, Monica Lestari–Paramita, Stefanie Nowak, Adrian Popescu, Andrzej Pronobis, Mark Sanderson, Tatiana Tommasi and Theodora Tsikrika.

Besides those involved in the organisation and actually performing the work, we also need to thank all the data providers who have enabled us to distribute and use their content. In particular we acknowledge the help of Norman Reid from St. Andrews University Library in Scotland for providing us access to the historic set of photographs for the first ImageCLEF evaluation campaign. In addition we thank the following institutions: University of Basel (Pathopic), University of Geneva (Casimage), Mallinckrodt Institute of Radiology (MIR), RSNA, Flickr, viventura, Belga, Wikipedia, LTUtech, PASCAL, UCLA (HEAL, PEIR), OHSU (CORI), MyPACS, KTH, and the IRMA group.

Thanks go to all those involved in carrying out annotations and relevance assessments across the tasks without whom we would have no gold standard to benchmark against. Many people have given up their own time to generate relevance assessments and not been paid for their contributions. We also thank all the participants to ImageCLEF tasks. It is the participants who make an event such as ImageCLEF possible and without whose support and comments there would be no results to report. Over the seven years of ImageCLEF, approximately 200 groups have signed up for at least one for the tasks and over 100 groups have submitted results to one of the tasks in multi-lingual image retrieval.

We thank those who were involved from the start of ImageCLEF. In particular we thank Carol Peters who has provided CLEF with ten years of dedicated service and created a stable environment in which to run not just ImageCLEF, but many important evaluation tasks for comparing and improving multi-lingual information retrieval systems. We thank Carol for allowing us to include ImageCLEF within the CLEF activities and for her continual support and encouragement of our work. We also gratefully acknowledge the support of Donna Harman for her insightful comments and the discussions we had on evaluation methodologies and directions to follow.

Finally, in terms of this book we would like to thank Ralf Gerstner from Springer Verlag for his help and support, David Clough for proofreading the chapters and helping produce the final version of the book and all the contributing authors. We were perhaps a little naïve when fixing the publication deadlines and had to relieve some of the stress on our fellow writers. We apologise for this and thank everyone for staying with us despite the difficulties. We hope that you will enjoy reading your chapters now they are published.

Contents

Part I Introduction

1	Seven Years of Image Retrieval Evaluation	3
	Paul Clough, Henning Müller, and Mark Sanderson	
1.1	Introduction	3
1.2	Evaluation of IR Systems	5
1.2.1	IR Test Collections	6
1.2.2	Cross–Language Evaluation Forum (CLEF)	9
1.3	ImageCLEF	9
1.3.1	Aim and Objectives	9
1.3.2	Tasks and Participants	11
1.3.3	Data sets	12
1.3.4	Contributions	12
1.3.5	Organisational Challenges	14
1.4	Conclusions	15
	References	16
2	Data Sets Created in ImageCLEF	19
	Michael Grubinger, Stefanie Nowak, and Paul Clough	
2.1	Introduction	19
2.1.1	Collection Creation	20
2.1.2	Requirements and Specification	21
2.1.3	Collection Overview	23
2.2	Image Collections for Photographic Retrieval	24
2.2.1	The St. Andrews Collection of Historic Photographs	24
2.2.2	The IAPR TC–12 Database	26
2.2.3	The Belga News Agency Photographic Collection	28
2.3	Image Collections for Medical Retrieval	29
2.3.1	The ImageCLEFmed Teaching Files	30
2.3.2	The RSNA Database	34
2.4	Automatic Image Annotation and Object Recognition	35

2.4.1	The IRMA Database	35
2.4.2	The LookThatUp (LTU) Data set	36
2.4.3	The PASCAL Object Recognition Database	37
2.4.4	The MIR Flickr Image Data Set.....	38
2.5	Image Collections in Other Tasks	38
2.5.1	The INEX MM Wikipedia Collection.....	39
2.5.2	The KTH-IDOL2 Database	40
2.6	Conclusions	41
	References	42
3	Creating Realistic Topics for Image Retrieval Evaluation	45
	Henning Müller	
3.1	Introduction	45
3.2	User Models and Information Sources	48
3.2.1	Machine-Oriented Evaluation	48
3.2.2	User Models	49
3.2.3	Information Sources for Topic Creation	50
3.3	Concrete Examples for Generated Visual Topics in Several Domains	53
3.3.1	Photographic Retrieval	53
3.3.2	Medical Retrieval	54
3.4	The Influence of Topics on the Results of Evaluation.....	55
3.4.1	Classifying Topics Into Categories	56
3.4.2	Links Between Topics and the Relevance Judgments	57
3.4.3	What Can Be Evaluated and What Can Not?	57
3.5	Conclusions	58
	References	59
4	Relevance Judgments for Image Retrieval Evaluation	63
	Jayashree Kalpathy-Cramer, Steven Bedrick, and William Hersh	
4.1	Introduction	63
4.2	Overview of Relevance Judgments in Information Retrieval	64
4.2.1	Test Collections	64
4.2.2	Relevance Judgments	65
4.3	Relevance Judging for the ImageCLEF Medical Retrieval Task ...	72
4.3.1	Topics and Collection	72
4.3.2	Judges	73
4.3.3	Relevance Judgment Systems and the Process of Judging	74
4.4	Conclusions and Future Work	78
	References	79
5	Performance Measures Used in Image Information Retrieval	81
	Mark Sanderson	
5.1	Evaluation Measures Used in ImageCLEF	81
5.2	Measures for Retrieval	82
5.2.1	Measuring at Fixed Recall	83

5.2.2	Measuring at Fixed Rank	85
5.2.3	Measures for Diversity	87
5.2.4	Collating Two Measures Into One	88
5.2.5	Miscellaneous Measures	88
5.2.6	Considering Multiple Measures	89
5.2.7	Measures for Image Annotation and Concept Detection ..	90
5.3	Use of Measures in ImageCLEF	91
5.4	Conclusions	92
	References	92
6	Fusion Techniques for Combining Textual and Visual Information Retrieval	95
	Adrien Depeursinge and Henning Müller	
6.1	Introduction	95
6.1.1	Information Fusion and Orthogonality	97
6.2	Methods	98
6.3	Results	98
6.3.1	Early Fusion Approaches	98
6.3.2	Late Fusion Approaches	99
6.3.3	Inter-media Feedback with Query Expansion	104
6.3.4	Other Approaches	105
6.3.5	Overview of the Methods from 2004–2009	105
6.4	Justification for the Approaches and Generally Known Problems ..	105
6.5	Conclusions	108
	References	108
Part II Track Reports		
7	Interactive Image Retrieval	117
	Jussi Karlgren and Julio Gonzalo	
7.1	Interactive Studies in Information Retrieval	117
7.2	iCLEF Experiments on Interactive Image Retrieval	119
7.2.1	iCLEF Image Retrieval Experiments: The Latin Square Phase	120
7.2.2	iCLEF Experiments with Flickr	123
7.2.3	The Target Collection: Flickr	124
7.2.4	Annotations	124
7.2.5	The Task	125
7.2.6	Experiments	127
7.3	Task Space, Technology and Research Questions	134
7.3.1	Use Cases for Interactive Image Retrieval	134
7.3.2	Challenges: Technology and Interaction	135
	References	137

8	Photographic Image Retrieval	141
	Monica Lestari Paramita and Michael Grubinger	
8.1	Introduction	141
8.2	Ad hoc Retrieval of Historic Photographs: ImageCLEF 2003–2005	142
8.2.1	Test Collection and Distribution	143
8.2.2	Query Topics	144
8.2.3	Relevance Judgments and Performance Measures	147
8.2.4	Results and Analysis	147
8.3	Ad hoc Retrieval of Generic Photographs: ImageCLEFphoto 2006–2007	149
8.3.1	Test Collection and Distribution	150
8.3.2	Query Topics	151
8.3.3	Relevance Judgments and Performance Measures	152
8.3.4	Results and Analysis	153
8.3.5	Visual Sub-task	154
8.4	Ad hoc Retrieval and Result Diversity: ImageCLEFphoto 2008–2009	155
8.4.1	Test Collection and Distribution	155
8.4.2	Query Topics	156
8.4.3	Relevance Judgments and Performance Measures	158
8.4.4	Results and Analysis	158
8.5	Conclusion and Future Prospects	160
	References	161
9	The Wikipedia Image Retrieval Task	163
	Theodora Tsikrika and Jana Kludas	
9.1	Introduction	163
9.2	Task Overview	164
9.2.1	Evaluation Objectives	164
9.2.2	Wikipedia Image Collection	165
9.2.3	Additional Resources	165
9.2.4	Topics	166
9.2.5	Relevance Assessments	167
9.3	Evaluation	169
9.3.1	Participants	169
9.3.2	Approaches	170
9.3.3	Results	175
9.4	Discussion	179
9.4.1	Best Practices	179
9.4.2	Open Issues	180
9.5	Conclusions and the Future of the Task	181
	References	181

10	The Robot Vision Task	185
	Andrzej Pronobis and Barbara Caputo	
10.1	Introduction	185
10.2	The Robot Vision Task at ImageCLEF 2009: Objectives and Overview	187
10.2.1	The Robot Vision Task 2009	188
10.2.2	Robot Vision 2009: The Database	188
10.2.3	Robot Vision 2009: Performance Evaluation	189
10.2.4	Robot Vision 2009: Approaches and Results	192
10.3	Moving Forward: Robot Vision in 2010	194
10.3.1	The Robot Vision Task at ICPR2010	194
10.3.2	The Robot Vision Task at ImageCLEF2010	196
10.4	Conclusions	197
	References	197
11	Object and Concept Recognition for Image Retrieval	199
	Stefanie Nowak, Allan Hanbury, and Thomas Deselaers	
11.1	Introduction	199
11.2	History of the ImageCLEF Object and Concept Recognition Tasks	200
11.2.1	2006: Object Annotation Task	201
11.2.2	2007: Object Retrieval Task	202
11.2.3	2008: Visual Concept Detection Task	203
11.2.4	2009: Visual Concept Detection Task	204
11.3	Approaches to Object Recognition	204
11.3.1	Descriptors	206
11.3.2	Feature Post-processing and Codebook Generation	207
11.3.3	Classifier	207
11.3.4	Post-Processing	208
11.4	Results	208
11.4.1	2006: Object Annotation Task	209
11.4.2	2007: Object Retrieval Task	209
11.4.3	2008: Visual Concept Detection Task	210
11.4.4	2009: Visual Concept Detection Task	211
11.4.5	Evolution of Concept Detection Performance	213
11.4.6	Discussion	214
11.5	Combinations with the Photo Retrieval Task	215
11.6	Conclusion	215
	References	216
12	The Medical Image Classification Task	221
	Tatiana Tommasi and Thomas Deselaers	
12.1	Introduction	221
12.2	History of ImageCLEF Medical Annotation	222
12.2.1	The Aim of the Challenge	222
12.2.2	The Database	223
12.2.3	Error Evaluation	227

12.3	Approaches to Medical Image Annotation	229
12.3.1	Image Representation	230
12.3.2	Classification Methods	230
12.3.3	Hierarchy	231
12.3.4	Unbalanced Class Distribution	231
12.4	Results	231
12.5	Conclusion	235
	References	237
13	The Medical Image Retrieval Task	239
	Henning Müller and Jayashree Kalpathy–Cramer	
13.1	Introduction	239
13.2	Participation in the Medical Retrieval Task	240
13.3	Development of Databases and Tasks over the Years	242
13.3.1	2004	242
13.3.2	2005–2007	243
13.3.3	2008–2009	247
13.4	Evolution of Techniques Used by the Participants	249
13.4.1	Visual Retrieval	250
13.4.2	Textual Retrieval	250
13.4.3	Combining Visual and Textual Retrieval	251
13.4.4	Case–Based Retrieval Topics	251
13.5	Results	251
13.5.1	Visual Retrieval	252
13.5.2	Textual Retrieval	252
13.5.3	Mixed Retrieval	253
13.5.4	Relevance Feedback and Manual Query Reformulation	253
13.6	Main Lessons Learned	253
13.7	Conclusions	255
	References	255

Part III Participant reports

14	Expansion and Re–ranking Approaches for Multimodal Image Retrieval using Text–based Methods	261
	Adil Alpkocak, Deniz Kilinc, and Tolga Berber	
14.1	Introduction	262
14.2	Integrated Retrieval Model	263
14.2.1	Handling Multi–modality in the Vector Space Model	264
14.3	Document and Query Expansion	265
14.4	Re–ranking	267
14.4.1	Level 1: Narrowing-down and Re-indexing	269
14.4.2	Level 2: Cover Coefficient Based Re–ranking	269
14.5	Results	271
14.6	Conclusions	273
	References	274

15	Revisiting Sub–topic Retrieval in the ImageCLEF 2009 Photo Retrieval Task	277
	Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose	
15.1	Introduction	278
15.2	Background and Related Work	280
15.2.1	Sub–topic Retrieval	280
15.2.2	The Probability Ranking Principle	282
15.2.3	Beyond Independent Relevance	282
15.3	Document Clustering and Inter–Cluster Document Selection	284
15.3.1	Re–examining Document Clustering Techniques	284
15.3.2	Clustering for Sub–topic Retrieval	285
15.4	Empirical Study	287
15.5	Results	290
15.6	Conclusions	291
	References	293
16	Knowledge Integration using Textual Information for Improving ImageCLEF Collections	295
	Manuel Carlos Díaz–Galiano, Miguel Ángel García–Cumbreras, María Teresa Martín–Valdivia, and Arturo Montejo–Ráez	
16.1	Introduction	295
16.2	System Description	297
16.2.1	Photo Retrieval System	297
16.2.2	Medical Retrieval System	298
16.3	Photo Task	298
16.3.1	Using Several IR and a Voting System	301
16.3.2	Filtering	302
16.3.3	Clustering	305
16.4	The Medical Task	306
16.4.1	Metadata Selection using Information Gain	306
16.4.2	Expanding with Ontologies	308
16.4.3	Fusion of Visual and Textual Lists	311
16.5	Conclusion and Further Work	311
	References	313
17	Leveraging Image, Text and Cross–media Similarities for Diversity–focused Multimedia Retrieval	315
	Julien Ah-Pine, Stephane Clinchant, Gabriela Csurka, Florent Perronnin, and Jean-Michel Renders	
17.1	Introduction	315
17.2	Content–Based Image Retrieval	317
17.2.1	Fisher Vector Representation of Images	318
17.2.2	Image Retrieval at ImageCLEF Photo	320
17.3	Text Representation and Retrieval	321
17.3.1	Language Models	321
17.3.2	Text Enrichment at ImageCLEF Photo	322

17.4	Text–Image Information Fusion	326
17.4.1	Cross–Media Similarities	327
17.4.2	Cross–Media Retrieval at ImageCLEF Photo	329
17.5	Diversity–focused Multimedia Retrieval	332
17.5.1	Re–ranking Top–Listed Documents to Promote Diversity	333
17.5.2	Diversity–focused Retrieval at ImageCLEF Photo	336
17.6	Conclusion	339
	References	340
18	University of Amsterdam at the Visual Concept Detection and Annotation Tasks	343
	Koen E. A. van de Sande and Theo Gevers	
18.1	Introduction	343
18.2	Concept Detection Pipeline	344
18.2.1	Point Sampling Strategy	345
18.2.2	Color Descriptor Extraction	346
18.2.3	Bag–of–Words model	347
18.2.4	Machine Learning	348
18.3	Experiments	349
18.3.1	Spatial Pyramid Levels	349
18.3.2	Point Sampling Strategies and Color Descriptors	350
18.3.3	Combinations of Sampling Strategies and Descriptors ...	351
18.3.4	Discussion	353
18.4	ImageCLEF 2009	353
18.4.1	Evaluation Per Image	355
18.4.2	Conclusion	355
18.5	ImageCLEF@ICPR 2010	356
18.6	Conclusion	356
	References	357
19	Intermedia Conceptual Indexing	359
	Jean–Pierre Chevallet and Joo Hwee Lim	
19.1	Introduction	359
19.2	Conceptual Indexing	361
19.2.1	Concept Usage and Definition in IR	361
19.2.2	Concept Mapping to Text	362
19.2.3	Mapping Steps	363
19.2.4	IR Models Using Concepts	366
19.2.5	Experiments using the ImageCLEF Collection	367
19.3	Image Indexing using a Visual Ontology	369
19.3.1	Image Indexing Based on VisMed Terms	370
19.3.2	FlexiTile Matching	373
19.3.3	Medical Image Retrieval Using VisMed Terms	374
19.3.4	Spatial Visual Queries	375
19.4	Multimedia and Intermedia Indexing	376

19.5	Conclusions	378
	References	379
20	Conceptual Indexing Contribution to ImageCLEF Medical Retrieval Tasks	381
	Loïc Maisonnasse, Jean–Pierre Chevallet, and Eric Gaussier	
20.1	Introduction	382
20.2	Semantic Indexing Using Ontologies	382
20.3	Conceptual Indexing	383
20.3.1	Language Models for Concepts	383
20.3.2	Concept Detection	384
20.3.3	Concept Evaluation Using ImageCLEFmed 2005–07	385
20.4	From Concepts to Graphs	386
20.4.1	A Language Model for Graphs	386
20.4.2	Graph Detection	387
20.4.3	Graph Results on ImageCLEFmed 2005–07	388
20.5	Mixing Concept Sources	388
20.5.1	Query Fusion	389
20.5.2	Document Model Fusion	389
20.5.3	Joint Decomposition	390
20.5.4	Results on ImageCLEFmed 2005–07	392
20.6	Adding Pseudo–Feedback	393
20.6.1	Pseudo–Relevance Feedback Model	393
20.6.2	Results	394
20.7	Conclusions	395
	References	395
21	Improving Early Precision in the ImageCLEF Medical Retrieval Task	397
	Steven Bedrick, Saïd Radhouani, and Jayashree Kalpathy–Cramer	
21.1	Introduction	397
21.1.1	What is Early Precision?	398
21.1.2	Why Improve Early Precision?	399
21.2	ImageCLEF	399
21.3	Our System	400
21.3.1	User Interface	400
21.3.2	Image Database	401
21.3.3	Query Parsing and Indexing	402
21.4	Improving Precision	403
21.4.1	Modality Filtration	403
21.4.2	Using Modality Information for Retrieval	406
21.4.3	Using Interactive Retrieval	408
21.5	Conclusions	411
	References	412

22 Lung Nodule Detection	415
Luca Bogoni, Jinbo Bi, Charles Florin, Anna K. Jerebko, Arun Krishnan, Sangmin Park, Vikas Raykar, and Marcos Salganicoff	
22.1 Introduction	415
22.1.1 Lung Cancer — Clinical Motivation	416
22.1.2 Computer-Aided Detection of Lung Nodules	418
22.1.3 Ground Truth for Lesions	419
22.2 Review of Existing Techniques	420
22.2.1 Gray-Level Threshold	421
22.2.2 Template Matching	421
22.2.3 Spherical Enhancing Filters	422
22.3 Description of Siemens LungCAD System	423
22.3.1 Lung Segmentation	423
22.3.2 Candidate Generation	423
22.3.3 Feature Extraction	424
22.4 Classification	425
22.4.1 Multiple Instance Learning	425
22.4.2 Exploiting Domain Knowledge in Data-Driven Training-Gated Classifiers	426
22.4.3 Ground Truth Creation: Learning from Multiple Experts	427
22.5 ImageCLEF Challenge	428
22.5.1 Materials and Methods	428
22.5.2 Results	429
22.6 Discussion and Conclusions	430
22.6.1 Clinical Impact	430
22.6.2 Future Extensions of CAD	432
References	433
23 Medical Image Classification at Tel Aviv and Bar Ilan Universities	435
Uri Avni, Jacob Goldberger, and Hayit Greenspan	
23.1 Introduction	435
23.1.1 Visual Words in Medical Archives	436
23.2 The Proposed TAU-BIU Classification System Based on a Dictionary of Visual-Words	437
23.2.1 Patch Extraction	438
23.2.2 Feature Space Description	438
23.2.3 Quantization	439
23.2.4 From an Input Image to a Representative Histogram	440
23.2.5 Classification	441
23.3 Experiments and Results	442
23.3.1 Sensitivity Analysis	444
23.3.2 Optimizing the Classifier	446
23.3.3 Classification Results	449
23.4 Discussion	450
References	451

24	Idiap on Medical Image Classification	453
	Tatiana Tommasi and Francesco Orabona	
24.1	Introduction	453
24.2	Multiple Cues for Image Annotation	454
24.2.1	High-Level Integration	455
24.2.2	Mid-Level Integration	456
24.2.3	Low-Level Integration	456
24.3	Exploiting the Hierarchical Structure of Data: Confidence Based Opinion Fusion	457
24.4	Facing the Class Imbalance Problem: Virtual Examples	458
24.5	Experiments	458
24.5.1	Features	458
24.5.2	Classifier	461
24.5.3	Experimental Set-up and Results	462
24.6	Conclusions	463
	References	464

Part IV External views

25	Press Association Images — Image Retrieval Challenges	469
	Martin Stephens and Dhavalkumar Thakker	
25.1	Press Association Images — A Brief History	469
25.1.1	The Press Association	469
25.1.2	Images at the Press Association	471
25.2	User Search Behaviour	472
25.2.1	Types of Users	472
25.2.2	Types of Search	473
25.2.3	Challenges	474
25.3	Semantic Web for Multimedia Applications	475
25.3.1	Introduction to the Semantic Web	475
25.3.2	Success Stories and Research Areas	475
25.3.3	The Semantic Web Project at Press Association Images ..	477
25.4	Utilizing Semantic Web Technologies for Improving User Experience in Image Browsing	478
25.4.1	PA Data set: Linking to the Linked Data Cloud	478
25.4.2	Information Extraction and Semantic Annotation	480
25.5	Conclusions and Future Work	481
	References	481
26	Image Retrieval in a Commercial Setting	483
	Vanessa Murdock, Roelof van Zwol, Lluís Garcia, and Ximena Olivares	
26.1	Introduction	483
26.2	Evaluating Large Scale Image Search Systems	486
26.3	Query Logs and Click Data	487
26.4	Background Information on Image Search	490
26.5	Multilayer Perceptron	491

26.6	Click Data	493
26.7	Data Representation	495
26.7.1	Textual Features	495
26.7.2	Visual Features	497
26.8	Evaluation and Results	498
26.8.1	Analysis of Features	500
26.9	Discussion of Results	502
26.10	Looking Ahead	503
	References	504
27	An Overview of Evaluation Campaigns in Multimedia Retrieval	507
	Suzanne Little, Ainhua Llorente, and Stefan Rüger	
27.1	Introduction	507
27.2	ImageCLEF in Multimedia IR (MIR)	509
27.2.1	INEX XML Multimedia Track	510
27.2.2	MIREX	511
27.2.3	GeoCLEF	511
27.2.4	TRECVID	512
27.2.5	VideOlympics	514
27.2.6	PASCAL Visual Object Classes (VOC) Challenge	514
27.2.7	MediaEval and VideoCLEF	515
27.2.8	Past Benchmarking Evaluation Campaigns	516
27.2.9	Comparison with ImageCLEF	517
27.3	Utility of Evaluation Conferences	518
27.4	Impact and Evolution of Metrics	519
27.5	Conclusions	521
	References	522
	Glossary	527
	Index	533

List of Contributors

Julien Ah-Pine

Xerox Research Centre Europe, Meylan, France

Adil Alpkocak

Dokuz Eylul University, Department of Computer Engineering, Izmir, Turkey

Uri Avni

Tel Aviv University, Tel Aviv, Israel

Steven Bedrick

Oregon Health and Science University, Portland, OR, USA

Tolga Berber

Dokuz Eylul University, Department of Computer Engineering, Izmir, Turkey

Jinbo Bi

Siemens at Malvern, PA, USA

Luca Bogoni

Siemens at Malvern, PA, USA

Barbara Caputo

Idiap Research Institute, Martigny, Switzerland

Jean–Pierre Chevallet

University of Grenoble, Laboratoire d’Informatique de Grenoble, Grenoble, France

Stephane Clinchant,

Xerox Research Centre Europe, Meylan, France

Paul D. Clough

University of Sheffield, Sheffield, UK

Gabriela Csurka,

Xerox Research Centre Europe, Meylan, France

Adrien Depeursinge

University and University Hospitals of Geneva (HUG), Geneva 14, Switzerland

Thomas Deselaers

Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland

Manuel Carlos Díaz Galiano

SINAI Research group, University of Jaén, Jaén, Spain

Charles Florin

Siemens at Malvern, PA, USA

Lluís Garcia

Yahoo! Research, Barcelona, Spain

Miguel Ángel García Cumbreiras

SINAI Research group, University of Jaén, Jaén, Spain

Eric Gaussier

University of Grenoble, Laboratoire d'Informatique, Grenoble, France

Theo Gevers

University of Amsterdam, Amsterdam, The Netherlands

Jacob Goldberger

Bar Ilan University, Ramat-Gan, Israel

Julio Gonzalo

E.T.S.I. Informática de la UNED, Madrid, Spain

Hayit Greenspan

Tel Aviv University, Tel Aviv, Israel

Michael Grubinger

Medellín, Colombia

Allan Hanbury

Information Retrieval Facility, Vienna, Austria

William Hersh

Oregon Health and Science University, Portland, OR, USA

Joo Hwee Lim

Institute for Infocom Research, Singapore

Anna K. Jerebko

Siemens at Malvern, PA, USA

Joemon M. Jose

University of Glasgow, Glasgow, UK

Jayashree Kalpathy-Cramer

Oregon Health and Science University, Portland, OR, USA

Jussi Karlgren
SICS, Kista, Sweden

Deniz Kilinc
Dokuz Eylul University, Department of Computer Engineering, Izmir, Turkey

Jana Kludas
CUI, University of Geneva, Switzerland

Arun Krishnan
Siemens at Malvern, PA, USA

Teerapong Leelanupab
University of Glasgow, Glasgow, UK

Monica Lestari Paramita
University of Sheffield, Sheffield, UK

Suzanne Little
KMi, The Open University, UK

Ainhua Llorente
KMi, The Open University, UK

Loïc Maisonnasse
TecKnowMetrix, Voiron, France

María Teresa Martín Valdivia
SINAI Research group, University of Jaén, Jaén, Spain

Arturo Montejo Ráez
SINAI Research group, University of Jaén, Jaén, Spain

Henning Müller
University of Applied Sciences Western Switzerland (HES–SO), Sierre,
Switzerland

Vanessa Murdock
Yahoo! Research, Barcelona, Spain

Stefanie Nowak
Fraunhofer IDMT, Ilmenau, Germany

Ximena Olivares
Universitat Pompeu Fabra, Barcelona, Spain

Francesco Orabona
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano,
Milano, Italy

Sangmin Park
Siemens at Malvern, PA, USA

Florent Perronnin
Xerox Research Centre Europe, Meylan, France

Andrzej Pronobis
Department of Computer Science, Royal Institute of Technology, Stockholm,
Sweden

Saïd Radhouani
Koodya sàrl, Bou Salem, Tunisia

Vikas Raykar
Siemens at Malvern, PA, USA

Jean-Michel Renders
Xerox Research Centre Europe, Meylan, France

Stefan Rüger
KMi, The Open University, Milton Keynes, UK

Marcos Salganicoff
Siemens at Malvern, PA, USA

Koen E. A. van de Sande
University of Amsterdam, Amsterdam, The Netherlands

Mark Sanderson
University of Sheffield, Sheffield, UK

Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies, Dublin City University, Dublin,
Ireland

Martin Stephens
Press Association Images, Nottingham, UK

Dhavalkumar Thakker
Press Association Images, Nottingham, UK

Tatiana Tommasi
Idiap Research Institute, Martigny, Switzerland

Theodora Tsikrika
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Guido Zuccon
University of Glasgow, Glasgow, UK

Roelof van Zwol
Yahoo! Research, Barcelona, Spain

Chapter 16

Knowledge Integration using Textual Information for Improving ImageCLEF Collections

Manuel Carlos Díaz–Galiano, Miguel Ángel García–Cumbreras, María Teresa Martín–Valdivia, and Arturo Montejo–Ráez

Abstract In this chapter we explain our participation at ImageCLEF from 2005 to 2009. During these years we have mainly developed systems for the ad hoc and the medical retrieval tasks. Although the different proposed tasks include both visual and textual information, the diverse approaches applied by the participants also include the use of only one type of information. The SINAI group specializes in the management of textual collections. For this reason, our main goal has been to improve the general system by taking advantage of the textual information.

16.1 Introduction

The first participation of the SINAI research group at the Cross Language Evaluation Forum (CLEF) was in 2002 presenting a multi-lingual system for the ad hoc task. Since then, we have followed the developments of CLEF and have participated in different tasks (GeoCLEF, CL-SR, etc.). Our first contribution at ImageCLEF was in 2005. In consecutive years we have mainly developed systems for the ad hoc and the medical retrieval tasks. We have modified the different models in order to adapt them to the new tasks proposed in ImageCLEF (wiki, photo, etc.), the new collections (CasImage, Pathopic, IAPR TC-12, etc.) and our areas of interest (application of machine translation, filtering of information, usage of ontologies, and knowledge integration). The changes have been addressed using the results obtained by both our systems and the other techniques presented at ImageCLEF.

Although the corpora provided by the ImageCLEF organizers include both textual and visual information, we have principally managed the textual data contained

Manuel Carlos Díaz Galiano · Miguel Ángel García Cumbreras · María Teresa Martín Valdivia · Arturo Montejo Ráez
SINAI Research group, University of Jaén, Paraje Las Lagunillas, s/n, Jaén (SPAIN), e-mail: *{mcdiaz, magc, maite, amontejo}@ujaen.es*

in the different collections. In practice, our main goal is to improve the general system by taking advantage of the textual information.

In addition, we have developed separate systems for the two main retrieval tasks at ImageCLEF: ad hoc and medical retrieval, although the best and more interesting results have been achieved with medical retrieval systems.

Thus, for the ad hoc task we were mainly interested in the different translation schemes even though we have also applied several retrieval models, weighting functions and query expansion techniques. However, from 2008 the task took a different approach to evaluate the diversity of results. Each query contained textual information and some relevant clusters. From that moment, our main interest has been to develop a clustering methodology in order to improve the result obtained. We have expanded the cluster terms with WordNet¹ synonyms. We have also introduced a clustering module based on the k-means algorithm and the creation of the final topics using the information of the title and the cluster terms. Unfortunately, the application of clustering does not improve the results.

Regarding the medical retrieval, we have investigated several methods and techniques. In our first participation in 2005 we studied different fusion methods in order to merge the results obtained from the textual Information Retrieval (IR) and Content Based Information Retrieval (CBIR) systems. In 2006, we tried to filter some features in the collections by applying Information Gain (IG) techniques. We accomplished several experiments in order to determinate the set of data that introduces less noise in the corpus. However, the results were not very relevant. Thus, in 2007 our major efforts were oriented to knowledge integration. We expanded the terms in the queries using the Medical Subject Headings (MeSH²) ontology. The results obtained were very good using only textual information. For this reason, in 2008 we investigated the effect of using another ontology, the Unified Medical Language System (UMLS³) meta-thesaurus. Surprisingly the results were not as good as we thought. Our main conclusion was that it is necessary to address the expansion of terms in a controlled way. The integration of all the terms without any filter scheme can include more noise in the final model and the system performance can be affected. Finally, the last participation at ImageCLEFmed tried to investigate the effect of expanding not only the query but also the whole collection. Again the results were not successful.

The next sections describe in a more detailed way the different systems developed for the ad hoc and medical retrieval tasks during our consecutive participation

¹ <http://wordnet.princeton.edu/>

² MeSH is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. <http://www.ncbi.nlm.nih.gov/mesh>

³ The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records. <http://www.nlm.nih.gov/research/umls/>

at ImageCLEF. Finally, Section 16.5 concludes this chapter and discusses further work.

16.2 System Description

It is often said that an image is worth 1,000 words. Unfortunately, these 1,000 words may differ from one individual to another depending on their perspective and/or knowledge of the image context. Thus, even if a 1,000-word image description were available, it is not certain that the image could be retrieved by a user with a different description.

Since 2005 we have developed and improved two independent systems: a photo retrieval system and a medical retrieval system. These systems work without user interaction (fully automatic) and they are focused on the textual information of the collections. Our aim was to develop and test different methods to improve the retrieval results, working with the associated text of the images.

It is usual for IR systems to pre-process the collections and the queries. All our approaches run this step, applying stopword removal and stemming (Porter, M.F., 1980). In addition, each non-English query is translated into English with our translation module, called SINTRAM (García-Cumbreras, M.A., Urena-López, L.A., Martínez-Santiago, F. and Perea-Ortega, J.M., 2007).

16.2.1 Photo Retrieval System

For more than ten years the SINAI group has tested and developed techniques to improve mono and multi-lingual information retrieval systems. For the ad hoc task of ImageCLEF techniques included the following:

- **IR systems.** Some IR systems have been used, selecting the ones that obtained the best results in our IR experiments (mono and multi-lingual). Different parameters have been tested, such as weighting functions (TFIDF, Okapi, InQuery), Psedo-Relevance Feedback (PRF) (Salton, G. and Buckley, G., 1990) and Query Expansion (QE).
- **Translation techniques.** Our machine translation system works with different online machine translators and implements several heuristics to combine them.
- **Fusion techniques.** When using several systems, the results lists have to be combined into a single combined one.
- **Expansion vs. Filtering.** Some approaches have been tested in order to expand terms from the query and the document and, also, to filter them when they are not very informative.

Figure 16.1 shows a general schema of our photo retrieval system.

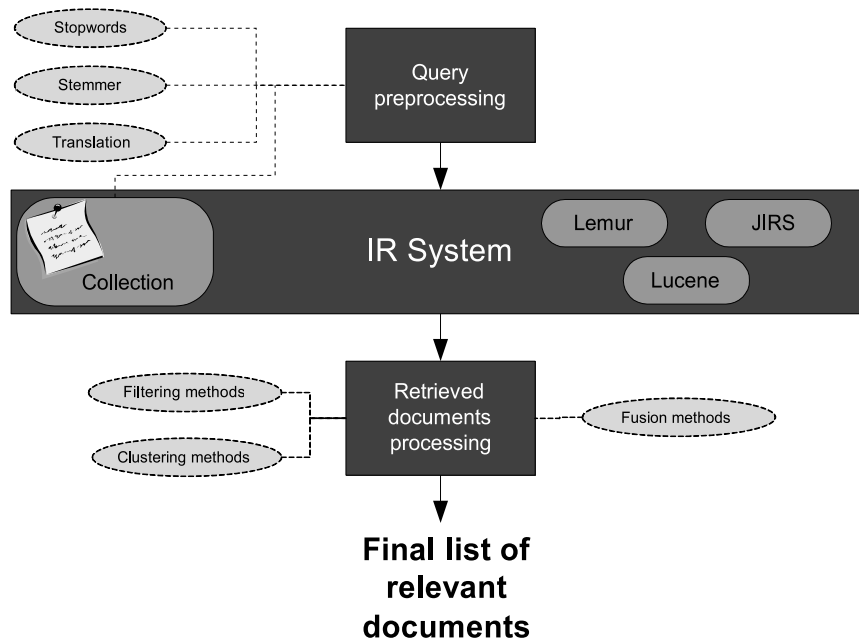


Fig. 16.1: General scheme of our photo retrieval system.

16.2.2 Medical Retrieval System

We only used textual techniques in the mixed IR system (visual and textual). For the medical retrieval task of ImageCLEF we have experimented in three aspects:

- Filtering textual information. We selected the best XML tags of the collection applying information gain (IG) metrics.
- Expanding the original query. We experimented using MeSH and UMLS ontologies.
- Combining relevant lists of textual and visual results. We applied several fusion techniques in order to merge visual and textual information.

Figure 16.2 shows a general scheme of our medical image retrieval system.

16.3 Photo task

The main aim of the photo retrieval task is to retrieve relevant photos given a photo query. The images have associated text, normally a few words, that describe them. Our photo retrieval system only works with the associated text to retrieve relevant

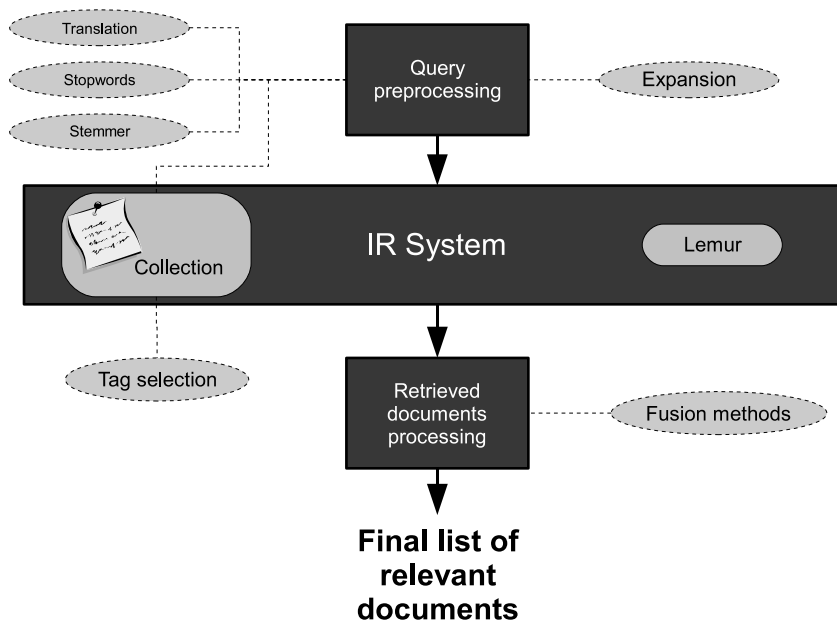


Fig. 16.2: General scheme of our medical retrieval system.

images and it only uses the text associated with each query, with or without context information.

In 2005 and 2006 the texts of the images (collection and queries) were independent phrases with a few words in the title field and a brief description in the narrative or description field. Other metadata was given for each query such as notes, dates and the location associated with the image. Some information was given in other languages than English, so it was necessary to use machine translation resources to translate them. In general, the results obtained with our system were good but it did not work well with the so-called *difficult queries*, queries with few relevant images in the collection or those with poor information.

To promote the diversity of results, with the aim of retrieving relevant images for all the queries, the query topics since 2007 included information about clusters. Each topic was clustered manually into sub-topics and the relevance judgements, to evaluate the results, included which cluster an image belonged to.

In the first developments of our system, the translation module SINTRAM was very important, because of multi-lingual queries used (English, Dutch, Italian, Spanish, French, German, Danish, Swedish, Portuguese and Russian). These first systems were composed of the following modules:

- a pre-processing module (normalization, stoword removal and stemming);

Table 16.1: Summary of results for the ad hoc task with multi-lingual queries.

Language	Initial Query	Expansion	MAP	%MONO	Rank
Dutch	title	with	0.3397	66.5%	2/15
Dutch	title	without	0.2727	53.4%	9/15
English	title + narr	with	0.3727	n/a	31/70
English	title	without	0.3207	n/a	44/70
French	title + narr	with	0.2864	56.1%	1/17
French	title + narr	without	0.2227	43.6%	12/17
German	title	with	0.3004	58.8%	4/29
German	title	without	0.2917	57.1%	6/29
Italian	title	without	0.1805	35.3%	12/19
Italian	title	with	0.1672	32.7%	13/19
Russian	title	with	0.2229	43.6%	11/15
Russian	title	without	0.2096	41.0%	12/15
Spanish	title	with	0.2416	47.3%	5/33
Spanish	title	without	0.2260	44.2%	8/33
Swedish	title	without	0.2074	40.6%	2/7
Swedish	title	with	0.2012	39.4%	3/7

- a translation module: based on the analysis of previous experiments, an automatic machine translator was defined by default for each pair of languages. For instance, Epals⁴ (German and Portuguese), Prompt⁵ (Spanish), Reverso⁶ (French) or Systran⁷ (Dutch and Italian);
- an IR module: the Lemur⁸ IR system was tuned up, and PRF with the Okapi weighting function was applied.

Table 16.1 shows the best result obtained for each language with the first development. These results are presented in terms of Mean Average Precision (MAP). The first column shows the language of the queries; the second one includes the fields used (*title*, *narr*, *description*); the third one shows if there was query expansion. The %MONO column shows the loss of precision of the multi-lingual queries according to the monolingual one (English MAP). The last column shows the ranking obtained with our experiment among the rest of the participants in the ImageCLEF photo task.

The results obtained show that, in general, the IR system Lemur works well with the Okapi weighting function, and the application of query expansion improves the results. Only one Italian experiment without query expansion gets a better result. In the case of the use of only *title* or *title + narrative*, the results are not conclusive, but the use of only *title* seems to produce better results. Multi-lingual queries produced a loss of precision of around a 25%. Figure 16.3 shows the loss of MAP with multi-lingual queries.

⁴ <http://www.epals.com/>

⁵ <http://www.online-translator.com/>

⁶ <http://www.reverso.net/>

⁷ <http://www.systran.co.uk/>

⁸ <http://www.lemurproject.org/>

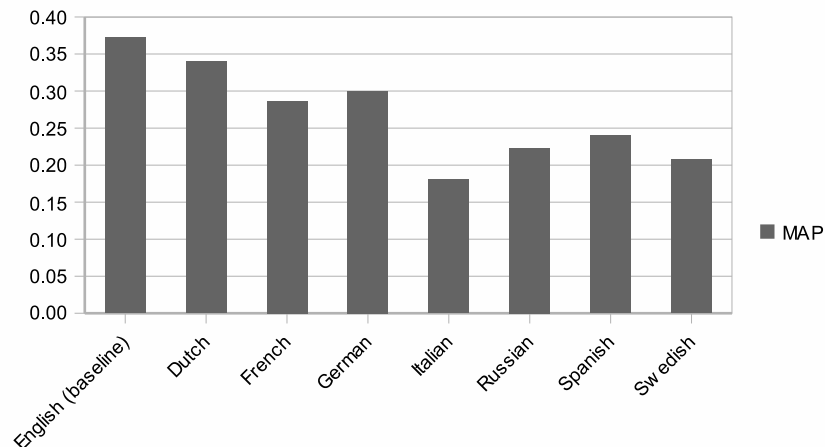


Fig. 16.3: Loss of MAP between the English queries and the multi-lingual ones.

16.3.1 Using Several IR and a Voting System

Later development of our photo retrieval system used several IR systems and a voting scheme to combine the results. Lemur and JIRS (Java Information Retrieval System) (Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., and Rosso, P., 2005) were adapted for our system. Lemur is a toolkit that supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or subcollections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models. JIRS is a passage retrieval system oriented to Question Answering (QA) tasks although it can be applied as an IR system. The complete architecture of the voting system is described in Figure 16.4.

Baseline cases with only Lemur and JIRS were run, so a final result was generated from a simple voting system with both IR systems that normalizes the scores and combines them with weights for each IR system (based on previous experiments and their evaluations). Table 16.2 shows the results obtained with the voting system (monolingual and bilingual runs).

In general, the results were poor because the set of queries was composed by only a few words. Nevertheless, our results were good in comparison with the other participants. After the analysis of these experiments, the English runs have obtained a loss of MAP of around 25%, being the worst results. Our best Spanish experiment was similar to the best one in the competition. For Portuguese we obtained the best one, and for French and Italian our runs were a bit worse: only a loss of MAP of around 8%. From these results we conclude that the Lemur IR system works better than JIRS, although the difference is not significant.

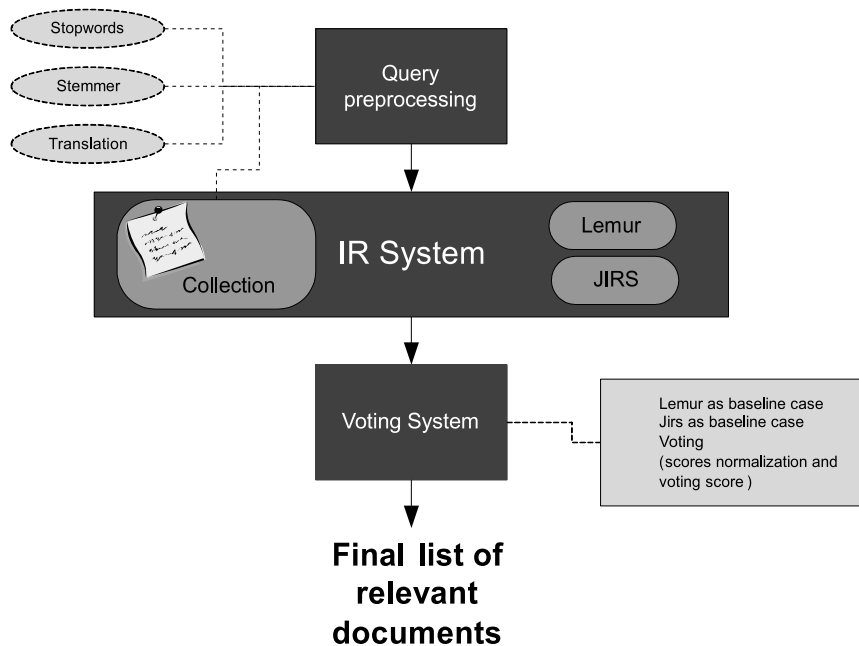


Fig. 16.4: Complete architecture of the voting system.

Fusion techniques have not improved the single ones. Lower MAP values decreased when we combined relevance lists. Other techniques must be used when the queries have few words.

16.3.2 Filtering

In the later evolution of our photo retrieval system we applied a filtering method over the results. In a first step the cluster term is expanded with its WordNet synonyms (the first sense). Then, the list of relevant documents generated by the IR system is filtered. If the relevant document contains the cluster term or a synonym, its *doc_id* (the identifier of the document) is written in another list. Finally, the new list with the filtered documents is combined with the original ones (Lemur and JIRS) in order to improve them. A simple method to do this was to double the score value of the documents in the filtered list and to add them to the original ones. The general architecture of the filtering system is shown in Figure 16.5.

The experiments carried out with the filtering system are as follows:

Table 16.2: Summary of results with the voting system (monolingual and bilingual runs).

Language	IR	MAP	Best MAP
English	Lemur	0.1591	0.2075
English	JIRS	0.1473	0.2075
English	Voting	0.0786	0.2075
Spanish	Lemur	0.1498	0.1558
Spanish	JIRS	0.1555	0.1558
Spanish	Voting	0.0559	0.1558
Portuguese	Lemur	0.1490	0.1490
Portuguese	JIRS	0.1350	0.1490
Portuguese	Voting	0.0423	0.1490
French	Lemur	0.1264	0.1362
French	JIRS	0.1195	0.1362
French	Voting	0.0323	0.1362
Italian	Lemur	0.1198	0.1341
Italian	JIRS	0.1231	0.1341
Italian	Voting	0.0492	0.1341

1. **Exp1: baseline case.** As baseline, Lemur was used as the IR system with automatic feedback and Okapi as weighting function. There was no combination of results, nor filtering method with the cluster term.
2. **Exp2: LemurJirs.** We combined the IR lists of relevant documents. Lemur with Okapi as weighting function and PRF. Before the combination of results Lemur and JIRS lists are filtered, only with the cluster term.
3. **Exp3: Lemur fb okapi.** The Lemur list of relevant documents (with Okapi and PRF) is filtered with the cluster term and its WordNet synonyms.
4. **Exp4: Lemur fb tfidf.** It is the same experiment as before, but in this case the weighting function used was TFIDF.
5. **Exp5: Lemur simple okapi.** The Lemur IR system has been run with Okapi as weighting function but without feedback. The list of relevant documents has been filtered with the cluster term and its WordNet synonyms.
6. **Exp6: Lemur simple tfidf.** The Lemur IR system has been used with TFIDF as weighting function but without feedback. The list of relevant documents has not been filtered.

The results are shown in Table 16.3. The last column represents the best F_1 score obtained in the 2008 competition (complete automatic systems with only text).

The results show that a simple filtering method is not useful if the cluster term or related words are used to filter the IR retrieved documents. It happens because some good documents are deleted and new relevant documents are not included in the second step. In general, the results in terms of MAP or other precision values are not very different. Between the best MAP and the worse one the difference is less than 8%. Filtering methods have not improved the baseline case.

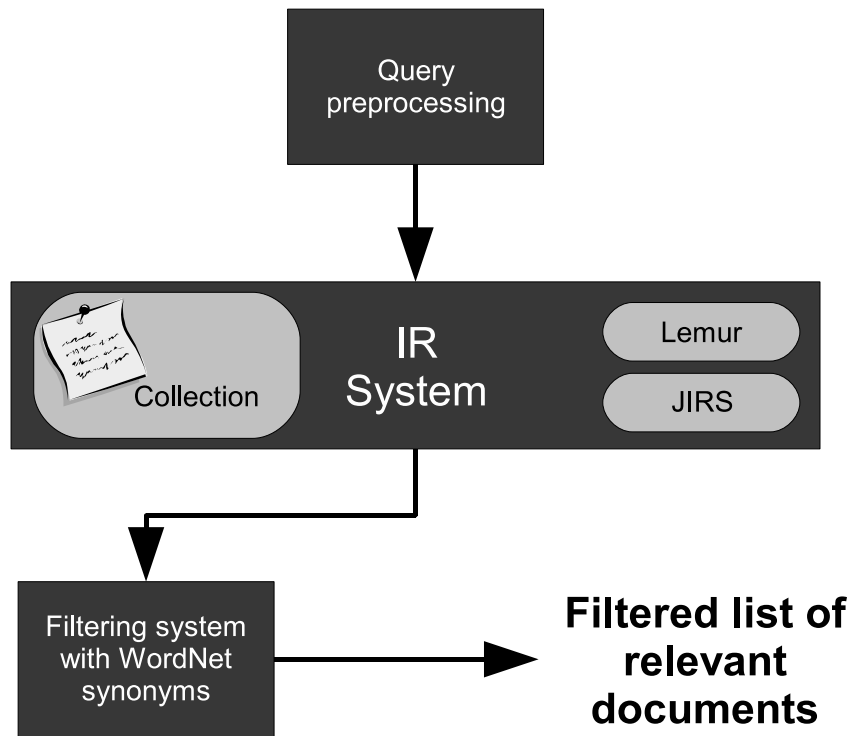


Fig. 16.5: Filtering scheme of the SINAI system.

Table 16.3: Results obtained with the filtering system.

Id	Filtering	FB	Expansion	MAP	P@5	P@10	Best F_1
Exp1	No	Yes	No	0.2125	0.3744	0.3308	0.2957
Exp6	No	No	No	0.2016	0.3077	0.2872	0.2957
Exp2	Yes	Yes	No	0.2063	0.3385	0.2949	0.2957
Exp3	Yes	Yes	No	0.2089	0.3538	0.3128	0.2957
Exp4	Yes	Yes	No	0.2043	0.2872	0.2949	0.2957
Exp5	Yes	No	No	0.1972	0.3385	0.3179	0.2957

After an analysis of the performance of filtering we can infer some reasons for this:

- Some relevant documents that appear in the first retrieval phase have been deleted because they do not contain the cluster term, so the cluster term is not useful in a filtering process.
- Other documents retrieved by the IR system that are not relevant, contain synonyms of the cluster term, so they are not deleted and the precision decreases.

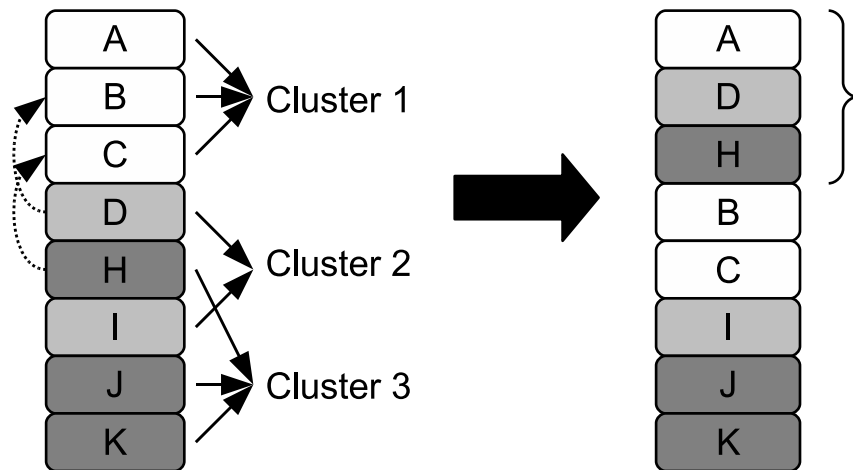


Fig. 16.6: Reordering of top results to increase variability according to clusters found.

16.3.3 Clustering

It was found that when increasing the variability of the top results in a list of documents retrieved as an answer to a query, the performance of the retrieval system increases too. Thus, in some cases it is more desirable to have less but more varied items in the results list (Chen and Karger, 2006). In order to increase variability, a clustering system has been applied. This was also used in other systems with the same aim (Ah-Pine et al, 2009). The idea behind it is rather simple: re-arrange the most relevant documents so that documents belonging to different clusters are promoted to the top of the list.

The K-means algorithm was applied on each of the lists returned by the Lemur IR system. For this, the Rapid Miner tool was used⁹. The clustering algorithm tried to group these results into four different groups, without any concern about ranking. The number of groups was established on this value as documents in the training set have this average number of clusters specified in their metadata.

Once each of the documents in the list was labeled to its computed cluster index, the list was reordered according to the described principle: we fill the list by alternating documents from different clusters. In Figure 16.6 a graphical example of this approach is given.

The list obtained with the base case was reordered according to the method described. The aim of this experiment is to increment the diversity of the retrieved results using a clustering algorithm. Results were discouraging: when no reordering

⁹ Available at <http://rapid-i.com/>

of documents in the list is performed a MAP of 0.4454 was reached, whereas a MAP of 0.2233 resulted from applying our clustering based approach.

16.4 The Medical Task

The main aim of the medical task is to retrieve medical images relevant to a given query. The query has several image examples and an associated text. The collection used to search relevant images has changed since 2005. Until 2007 the collection was very heterogeneous, with several subcollections. The subcollections without XML tags were processed to mark the structure of documents using XML. We used the SINTRAM tool to translate non-English text. Each subcollection is divided up into *cases* where a case is made up of one or various images (depending on the collection), along with an associated set of textual annotations. All the collections were processed to generate one textual document per image (Díaz-Galiano et al, 2006).

In 2007 a new collection was introduced, a subset of the Goldminer¹⁰ collection. This collection contains images from articles published in *Radiology and Radiographics* including the text of the captions and a link to the Web page of the full text article. To create the different textual collections, first we have obtained the textual information by downloading all the articles from the Web. Then, we have filtered the articles to extract different sections (title, authors, abstract, introduction, etc.). Our experiments were conducted with the LEMUR retrieval information system, applying the KL-divergence weighting scheme (Ogilvie and Callan, 2001) and PRF.

16.4.1 Metadata Selection using Information Gain

The collection used until 2007 includes a large number of XML tags. The main problem was to choose the most useful data, discarding anything that might add non-relevant information (noise) to our system. In order to automate the tag selection process we have pre-processed the collections using Information Gain (IG) (Cover and Thomas, 2006). The XML tags were selected according to the amount of information supplied. For this reason, we have used the IG measure to select the best tags in the collection, using the following formula:

$$IG(C|E) = H(C) - H(C|E) \quad (16.1)$$

¹⁰ <http://goldminer.arrs.org/>

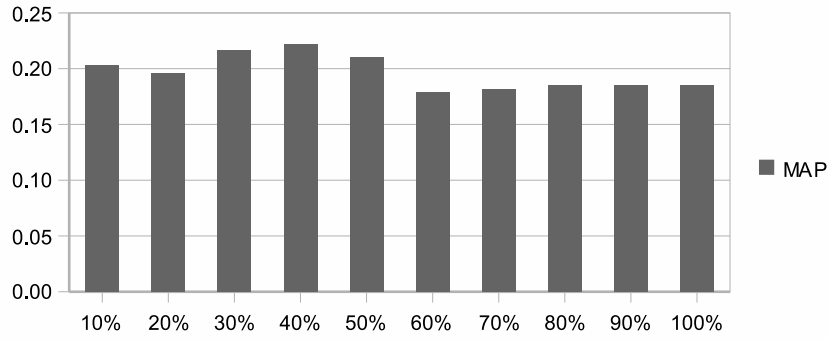


Fig. 16.7: Performance for Medical Image Retrieval in 2006.

where

C is the set of cases,

E is the value set for the E tag,

$IG(C|E)$ is the information gain for the E tag,

$H(C)$ is the entropy and of the set of cases C

$H(C|E)$ is the relative entropy of the set of cases C conditioned by the E tag

Both, $H(C)$ and $H(C|E)$ are calculated based on the frequencies of occurrence of tags according to the combination of words which they represent. The final equation for the computation of the information gain supplied by a given tag E over the set of cases C is defined as follows:

$$IG(C|E) = -\log_2 \frac{1}{|C|} + \sum_{j=1}^{|E|} \frac{|C_{e_j}|}{|C|} \log_2 \frac{1}{|C_{e_j}|} \quad (16.2)$$

where

C_{e_j} is the subset of cases in C having the tag E set to the value e_j (this value is a combination of words where order does not matter).

Since each subcollection has a different set of tags, the information gain was calculated for each subcollection individually. Then, the tags selected to compose the final collection are those showing high values of IG. We have accomplished several experiments preserving 10%, 20%...100% of tags. Figure 16.7 shows the values of MAP obtained for the Medical Image Retrieval task from 2006 using only textual information.

The results show that the collections with a low percentage of labels (between 30% and 50%) obtain the best performance, with a MAP value between 0.21 and 0.22. Therefore, this method reduces the size of the collections used and allows us to select the most significant labels within the corpus or, at least, those that provide better information. This selection system does not require external training or

knowledge; it simply studies the importance of each label with regard to all the documents. Furthermore, this method is independent from the corpus as a whole since in our experiments the IG calculation has been done separately in each subcollection.

16.4.2 Expanding with Ontologies

We have experimented with two ontologies: MeSH and UMLS, performing several experiments with different expansion types. The best results have been obtained using synonyms and related terms.

To expand with the MeSH ontology we have used the *record* structure. Each record contains a representative term and a bag of synonyms and related terms. We consider that a term is a set of words (no word sequence order):

$$t = \{w_1, \dots, w_{|t|}\} \quad (16.3)$$

where w is a word.

We have used the bag of terms to expand the queries. A bag of terms is defined as:

$$b = \{t_1, \dots, t_{|b|}\} \quad (16.4)$$

Moreover, a term t exists in the query q ($t \in q$) if:

$$\forall w_i \in t, \exists w_j \in q / w_i = w_j \quad (16.5)$$

Therefore, if all the words of a term are in the query, we generate a new expanded query by adding all its bag of terms.

$$q \text{ is expanded with } b \text{ if } \exists t \in b / t \in q \quad (16.6)$$

In order to compare the words of a particular term to those of the query, all the words are put in lowercase and no stopword removal is applied. To reduce the number of terms that could expand the query, we have only used those that are in A, C or E categories of MeSH (A: Anatomy, C: Diseases, E: Analytical, Diagnostic and Therapeutic Techniques and Equipment): (Chevallet et al, 2006). Figure 16.8 shows an example of query expansion, with two query terms found in MeSH and their respective bags of terms.

On the other hand, to expand the queries with the UMLS metathesaurus, we have used the MetaMap program (Aronson, 2001) that was originally developed for information retrieval. MetaMap uses the UMLS meta-thesaurus for mapping concepts from an input text. For query expansion with MetaMap, we have mapped the terms from the query. As carried out with MeSH, in order to restrict the categories of terms that could expand the query, we have restricted the semantic types in the mapped terms (Chevallet et al, 2006) as follows:

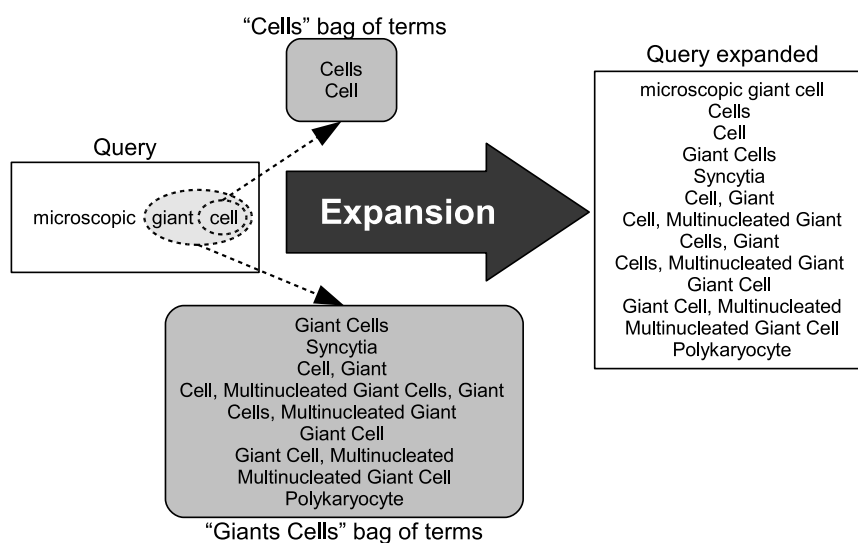


Fig. 16.8: Example of query expansion with MeSH ontology.

- bpoc: Body Part, Organ, or Organ Component;
- diap: Diagnostic Procedure;
- dsyn: Disease or Syndrome;
- neop: Neoplastic Process.

MetaMap gives two types of mapped terms: *Meta Candidates* and *Meta Mapping*. The difference between both mapped terms is that the second are the Meta Candidate with best score. For our expansion we have used the Meta Candidate terms, because these provide similar terms with differences in the words (Díaz-Galiano et al, 2006).

Prior to the inclusion of Meta Candidates terms in the queries, the words of the terms are added to a set where repeated words are deleted. All words in the set are included in the query. Figure 16.9 shows a example of query expansion using UMLS.

The organizers of the ImageCLEF medical task provided the *ImageCLEF Consolidated Test Collection* (Hersh et al, 2009). This collection combines all the collections, queries and relevance judgements used in ImageCLEFmed from 2005 to 2007. We have used this new collection to experiment with MeSH and UMLS query expansion. On the other hand, to experiment with the 2008 collection we have generated three different collections. In these collections each document contains information about each image from the original collection. The information is different for each collection. These collections are defined as follows:

- **CT**: contains *caption* of image and *title* of the article.
- **CTS**: contains *caption*, *title* and text of the *section* where the image appears.

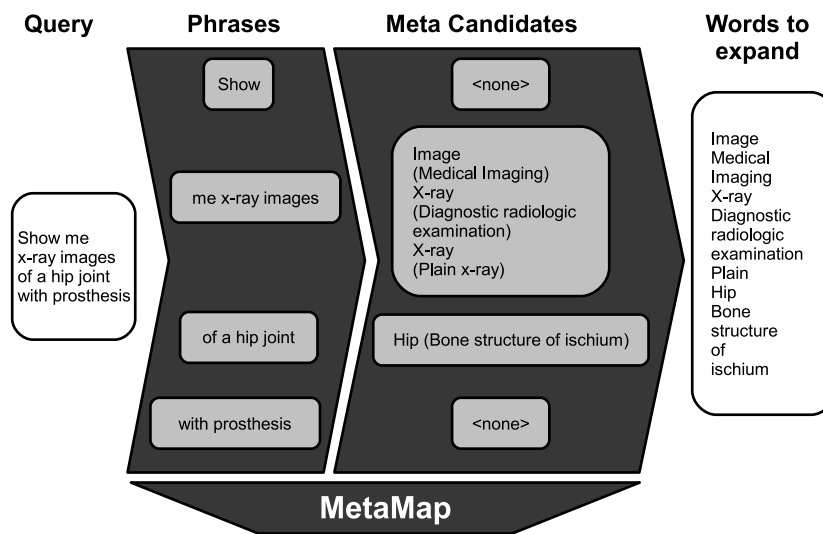


Fig. 16.9: Example of query expansion with UMLS ontology.

Table 16.4: MAP values of query expansion experiments.

Expansion	CT	CTS	CTA	Consolidated
Base	0.2480	0.1784	0.1982	0.2039
MeSH	0.2792	0.1582	0.2057	0.2202
UMLS	0.2275	0.1429	0.1781	0.1985

- **CTA**: contains *caption*, *title* and text of the full *article*.

Table 16.4 shows the results obtained in experiments on these collections.

The MeSH expansion obtained better results than no expansion or UMLS expansion. In 2008 the University of Alicante group obtained the best results (Navarro et al, 2008) in the textual task using a similar MeSH expansion and negative feedback. The Miracle group performed a MeSH expansion in documents and topics using the hyponyms of UMLS entities (Lana-Serrano et al, 2008) but the results obtained are worse than the baseline results. In short, the use of UMLS expansion obtained worse results than the baseline. Although the UMLS meta-thesaurus includes the MeSH ontology in the source vocabularies, MetaMap adds, in general, more terms in the queries. The MetaMap mapping was different from MeSH mapping, therefore the terms selected to expand were not the same.

One conclusion is that it is better to have less but more specific textual information. Also, including the whole section where the image appears was not a good approach. Sometimes a section contains several images, therefore the same information references different images.

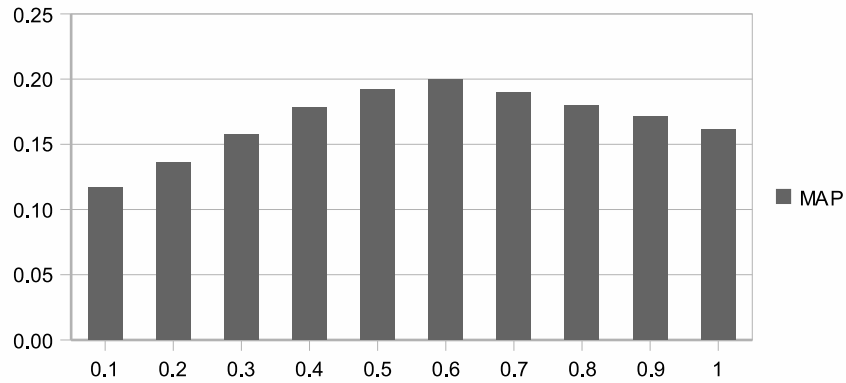


Fig. 16.10: Performance of experiments in 2005 with visual and textual fusion.

16.4.3 Fusion of Visual and Textual Lists

The fusion experiments merge the ranked lists from both systems (visual and textual) in order to obtain one final list (FL) with relevant images ranked by relevance. The merging process was accomplished giving different importance to the visual (VL) and textual lists (TL):

$$FL = TL * \alpha + VL * (1 - \alpha) \quad (16.7)$$

In order to adjust these parameters some experiments were accomplished varying α in the range $[0,1]$ with step 0.1 (i.e.: 0.1, 0.2,...,0.9 and 1).

The next figures show the results obtained on different collections used in the ImageCLEF medical task. Figure 16.10 shows experiment results with the 2005 collection. Results with the 2007 collection are presented in Figure 16.11.

The results obtained show that the combination of heterogeneous information sources (textual and visual) improves the use of a single source. Although textual retrieval on its own overcomes visual retrieval, when used jointly the results are better than those obtained from independent retrievals.

16.5 Conclusion and Further Work

In this chapter, we have described our participation in ImageCLEF from 2005 to present. We have presented a summary of different systems developed for the photo and medical retrieval tasks.

For the photo retrieval system we have tested multiple resources and techniques: different IR systems, weighting schemes, pseudo relevance feedback, ma-

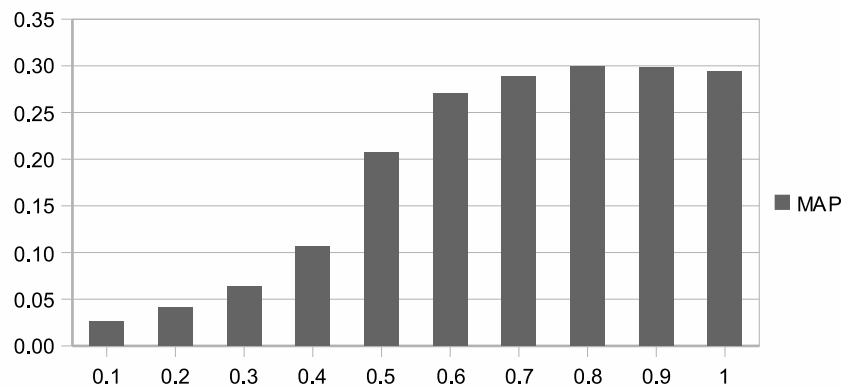


Fig. 16.11: Performance of experiment in 2007 with visual and textual fusion.

chine translators, filtering methods and clustering to increase diversity in the results. The experiments show that the translation of non-English queries introduces a loss of MAP that depends on the source language, although those multi-lingual runs achieved almost the best results in the competition. Our IR system works well, in general, with the Okapi weighting function. In addition, the application of query expansion and PRF improved the results. The applied filtering method shows that the cluster terms given in the query are not useful to filter the relevant list of images, and the applied clustering method obtained poor results in terms of MAP. However, the diversity of the relevant images was increased, so further research should be conducted on this issue.

In our future work with the photo retrieval system, we will improve the machine translation subsystem, including a new translator and heuristics to combine the results. New filtering methods are ruled out for the time being, because we are developing a new clustering module that introduces diversity in the results but taking into account the score and position of the documents in original the ranked list.

Regarding the medical task, we have applied Information Gain in order to filter tags in the collections. The best results have been obtained using around 30%-50% of the tags. In addition, it has been found that the application of fusion techniques to combine textual and visual information improves the system. Finally, several query expansion techniques have been tested using two medical resources: MeSH and UMLS. The experiments show that the expansion with less and more specific terms improves the results.

As future work we will study which resources from UMLS are more convenient for term expansion. In addition, we are interested in detecting when the query expansion is useful to improve the final results.

Acknowledgements This work has been partially supported by a grant from the Spanish Government, project TEXT-COOL 2.0 (TIN2009-13391-C04-02), a grant from the Andalusian Gov-

ernment, project GeOasis (P08-TIC-41999), and two grants from the University of Jaén, project RFC/PP2008/UJA-08-16-14 and project UJA2009/12/14.

References

- Ah-Pine J, Bressan M, Clinchant S, Csurka G, Hoppenot Y, Renders JM (2009) Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications* 42(1):31–56
- Aronson AR (2001) Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In: *Proceedings of the AMIA Symposium*, pp 17–21
- Chen H, Karger DR (2006) Less is more: probabilistic models for retrieving fewer relevant documents. In: Efthimiadis EN, Dumais ST, Hawking D, Järvelin K (eds) *Proceedings of the SIGIR conference*, ACM press, pp 429–436
- Chevallet JP, Lim JH, Radhouani S (2006) A structured visual learning approach mixed with ontology dimensions for medical queries. In: *Accessing Multilingual Information Repositories*, Springer, *Lecture Notes in Computer Science (LNCS)*, pp 642–651
- Cover T, Thomas J (2006) *Elements of information theory*. Wiley–Interscience
- Díaz-Galiano M, García-Cumbreras M, Martín-Valdivia M, Montejo-Raez A, , Ureña López L (2006) SINAI at ImageCLEF 2006. In: *Working Notes of CLEF 2006*
- García-Cumbreras, MA, Ureña-López, LA, Martínez-Santiago, F and Perea-Ortega, JM (2007) BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. In: *Lecture Notes in Computer Science (LNCS)*, Springer, vol 4730, pp 328–338
- Gómez-Soriano, JM, Montes-y-Gómez, M, Sanchis-Arnal, E, and Rosso, P (2005) A Passage Retrieval System for Multilingual Question Answering. In: *8th International Conference of Text, Speech and Dialogue 2005 (TSD 2005)*, Springer, *Lecture Notes in Artificial Intelligence (LNCS)*, pp 443–450
- Hersh WR, Müller H, Kalpathy-Cramer J (2009) The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging* 22(6):648–655
- Lana-Serrano S, Villena-Román J, González-Cristóbal J (2008) MIRACLE at ImageCLEFmed 2008: Evaluating Strategies for Automatic Topic Expansion. In: *Working Notes of CLEF 2008*
- Navarro S, Llopis F, Muñoz R (2008) Different Multimodal Approaches using IR-n in ImageCLEFphoto 2008. In: *Working Notes of CLEF 2008*
- Ogilvie P, Callan JP (2001) Experiments using the lemur toolkit. In: *Proceedings of TREC*
- Porter, MF (1980) An algorithm for suffix stripping. In: *Program 14*, pp 130–137
- Salton, G and Buckley, G (1990) Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences* 21:288–297