Universidade Estadual de Campinas
Instituto de Computação

Rodrigo Tripodi Calumby

Diversity-oriented Multimodal and Interactive Information Retrieval

Recuperação Multimodal e Interativa de Informação Orientada por Diversidade

CAMPINAS
2015

Rodrigo Tripodi Calumby

Diversity-oriented Multimodal and Interactive Information Retrieval

Recuperação Multimodal e Interativa de Informação Orientada por Diversidade

Tese apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Supervisor/Orientador: Prof. Dr. Ricardo da Silva Torres

Este exemplar corresponde à versão final da Tese defendida por Rodrigo Tripodi Calumby e orientada pelo Prof. Dr. Ricardo da Silva Torres.

CAMPINAS
2015

Agência(s) de fomento e nº(s) de processo(s): CAPES, P-4388/2010; CNPq, 140977/2012-0

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

C138d

Calumby, Rodrigo Tripodi, 1985-
Diversity-oriented multimodal and interactive information retrieval / Rodrigo Tripodi Calumby. – Campinas, SP : [s.n.], 2015.

Orientador: Ricardo da Silva Torres.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Recuperação da informação. 2. Interação humano-computador. 3. Aprendizado de máquina. 4. Programação genética (Computação). I. Torres, Ricardo da Silva,1977-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Recuperação multimodal e interativa de informação orientada por diversidade
Palavras-chave em inglês:
Information retrieval
Human-computer interaction
Machine learning
Genetic programming (Computer science)
Área de concentração: Ciência da Computação
Titulação: Doutor em Ciência da Computação
Banca examinadora:
Ricardo da Silva Torres [Orientador]
Rodrygo Luis Teodoro Santos
João Paulo Papa
Ariadne Maria Brito Rizzoni Carvalho
Hélio Pedrini
Data de defesa: 18-12-2015
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Rodrigo Tripodi Calumby

Diversity-oriented Multimodal and Interactive Information Retrieval

Recuperação Multimodal e Interativa de Informação Orientada por Diversidade

Banca Examinadora:

- Prof. Dr. Ricardo da Silva Torres (Presidente)
  Universidade Estadual de Campinas
- Profa. Dra. Ariadne Maria Brito Rizzoni Carvalho
  Universidade Estadual de Campinas
- Prof. Dr. Hélio Pedrini
  Universidade Estadual de Campinas
- Prof. Dr. João Paulo Papa
  Universidade Estadual Paulista
- Prof. Dr. Rodrygo Luis Teodoro Santos
  Universidade Federal de Minas Gerais

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 18 de dezembro de 2015

Dedication

This thesis is dedicated to those whose behavior and knowledge shaped my character and gave me more than I could have ever asked for.

To my grandmother Prof. Voninha, teacher and sentinel. Realizing that humankind returned to zero and paraphrasing Ruy Barbosa about the triumph of the nullity, she pursues and appreciates generosity, virtuousness, honor, and honesty.

To my grandfather Prof. Aldo (in memoriam). Teacher and model. My first advisor. Roughly twenty years ago, he brought home the very first computer I ever touched and, without noticing, planted the computer science curiosity. Here I am.

To my mother Rosa Maria, whose unconditional love and guidance inspired me to fulfill this task and kept me aware of the important things life brings us. To my sister Maria Luiza, whose caring look always made me feel home wherever we were. To my aunt Carmela, my cousin Ana Luíza, and my friend Donata, which so motherly welcomed and hosted me. To my uncle Vicente Tripodi and my friend Regina. To my father Francisco and his wife Eliana. To my baby brothers Arthur and Eva.

To my parents-in-law, João Araujo (in memoriam) and Vania, who welcomed me as son and allowed me to hold the greatest treasure I could dream of, my beloved wife Renata. To my sisters-in-law Adriana and Vanessa and their spouses.

Specially to Renata Freitas, who accepted and became Mrs. Calumby during this work. I once again offer her my life and dedication for in her heart I found peace and in her eyes I found God. He made us one and all of it possible.

Dealing with ambiguity is far easier if you learn how to enjoy it.

(Anuranjita Kumar, Can I have it all?)

It's not about adding diversity for the sake of diversity, it's about subtracting homogeneity for the sake of realism.

(Mary Robinette Kowal)

What we have to do... is to find a way to celebrate our diversity and debate our differences without fracturing our communities.

(Hillary Clinton)

We need to reach that happy stage of our development when differences and diversity are not seen as sources of division and distrust, but of strength and inspiration.

(Josefa Iloilo)

Acknowledgements

The period of this PhD work was a rich and remarkable journey through science and through soul. This thesis is the result of a work from multiple hands and minds, and for them, I now express my deepest gratitude.

To my advisor, Dr. Ricardo da Silva Torres, my extreme appreciation. His work made me a better student, researcher, and professional. Beyond it, his lessons made me a better person. He introduced me to high-end research while taught me how doing it is worth when our work makes our lives better and allows us to share and support the ones around. Professor Torres accepted the challenge of guiding me in this work even after starting a full-time job and provided me great opportunities. From my first lecturing experiences to international communications, his countless suggestions greatly refined my actions, materials, and methods. I appreciate all the respect, dedication, understanding, and vote of confidence. Indeed I do.

For his end-to-end contribution, I am thankful to Professor Marcos André Gonçalves. His scientific strictness and writing cleverness made our experiments, findings, and papers more accurate and interesting. Professor Marcos led me to exercise reasoning and justification skills over statements and inferences.

I thank the examiner board in my research defense and final defense, all paper reviewers, and journal editors, for the important contributions that significantly improved this work.

I appreciate the contribution from many colleagues, which directly or indirectly contributed to this work, for the insightful discussions and the many co-authored works. In particular, my acknowledgements to Dr. Daniel C. G. Pedronette, Dr. Otávio A. B. Penatti, Javier Muñoz, Dr. Giovani Chiachia, Dr. Jurandy Almeida, Dr. Fabio A. Faria, and Dr. Jefersson A. dos Santos. In special, I appreciate all the partnership and support from Dr. Lin T. Li.

I thank Unicamp, CAPES, and CNPq for the scholarships that supported this work. I thank FAPESP, AMD, Samsung, and Microsoft for the infrastructure of the Recod laboratory. I thank the administrative and technical staff from Unicamp and UEFS that made this work possible.

I thank all my co-workers and students from the University of Feira de Santana. Especially, my sincere regards to Prof. Ana Lucia L. M. Maia and Prof. Tiago A. Coelho for all the friendship and partnership along these years. You made my work lighter and happier.

I want to highlight all my gratitude and love to my friends and family that inspired me through this journey and understood some moments of absence.

I express my profound gratitude to God. His watch and protection made my path safe. His words made my mind peaceful and my heart calm. His lessons and guidance brought me resilience and made me stronger and wiser.

I wish I had three pages.

Resumo

Os métodos de Recuperação da Informação, especialmente considerando-se dados multimídia, evoluíram para a integração de múltiplas fontes de evidência na análise de relevância de itens em uma tarefa de busca. Neste contexto, para atenuar a distância semântica entre as propriedades de baixo nível extraídas do conteúdo dos objetos digitais e os conceitos semânticos de alto nível (objetos, categorias, etc.) e tornar estes sistemas adaptativos às diferentes necessidades dos usuários, modelos interativos que consideram o usuário mais próximo do processo de recuperação têm sido propostos, permitindo a sua interação com o sistema, principalmente por meio da realimentação de relevância implícita ou explícita. Analogamente, a promoção de diversidade surgiu como uma alternativa para lidar com consultas ambíguas ou incompletas. Adicionalmente, muitos trabalhos têm tratado a ideia de minimização do esforço requerido do usuário em fornecer julgamentos de relevância, à medida que mantém-se níveis aceitáveis de eficácia.

Esta tese aborda, propõe e analisa experimentalmente métodos de recuperação da informação interativos e multimodais orientados por diversidade. Este trabalho aborda de forma abrangente a literatura acerca da recuperação interativa da informação e discute sobre os avanços recentes, os grandes desafios de pesquisa e oportunidades promissoras de trabalho. Nós propusemos e avaliamos dois métodos de aprimoramento do balanço entre relevância e diversidade, os quais integram múltiplas informações de imagens, tais como: propriedades visuais, metadados textuais, informação geográfica e descritores de credibilidade dos usuários. Por sua vez, como integração de técnicas de recuperação interativa e de promoção de diversidade, visando maximizar a cobertura de múltiplas interpretações/aspectos de busca e acelerar a transferência de informação entre o usuário e o sistema, nós propusemos e avaliamos um método multimodal de aprendizado para ranqueamento utilizando realimentação de relevância sobre resultados diversificados.

Nossa análise experimental mostra que o uso conjunto de múltiplas fontes de informação teve impacto positivo nos algoritmos de balanceamento entre relevância e diversidade. Estes resultados sugerem que a integração de filtragem e re-ranqueamento multimodais é eficaz para o aumento da relevância dos resultados e também como mecanismo de potencialização dos métodos de diversificação. Além disso, com uma análise experimental minuciosa, nós investigamos várias questões de pesquisa relacionadas à possibilidade de aumento da diversidade dos resultados e a manutenção ou até mesmo melhoria da sua relevância em sessões interativas. Adicionalmente, nós analisamos como o esforço em diversificar afeta os resultados gerais de uma sessão de busca e como diferentes abordagens de diversificação se comportam para diferentes modalidades de dados. Analisando a eficácia geral e também em cada iteração de realimentação de relevância, nós mostramos que introduzir diversidade nos resultados pode prejudicar resultados iniciais, enquanto que aumenta significativamente a eficácia geral em uma sessão de busca, considerando-se não apenas a relevância e diversidade geral, mas também o quão cedo o usuário é exposto ao mesmo montante de itens relevantes e nível de diversidade.

Abstract

Information retrieval methods, especially considering multimedia data, have evolved towards the integration of multiple sources of evidence in the analysis of the relevance of items considering a given user search task. In this context, for attenuating the semantic gap between low-level features extracted from the content of the digital objects and high-level semantic concepts (objects, categories, etc.) and making the systems adaptive to different user needs, interactive models have brought the user closer to the retrieval loop allowing user-system interaction mainly through implicit or explicit relevance feedback. Analogously, diversity promotion has emerged as an alternative for tackling ambiguous or underspecified queries. Additionally, several works have addressed the issue of minimizing the required user effort on providing relevance assessments while keeping an acceptable overall effectiveness.

This thesis discusses, proposes, and experimentally analyzes multimodal and interactive diversity-oriented information retrieval methods. This work, comprehensively covers the interactive information retrieval literature and also discusses about recent advances, the great research challenges, and promising research opportunities. We have proposed and evaluated two relevance-diversity trade-off enhancement work-flows, which integrate multiple information from images, such as: visual features, textual metadata, geographic information, and user credibility descriptors. In turn, as an integration of interactive retrieval and diversity promotion techniques, for maximizing the coverage of multiple query interpretations/aspects and speeding up the information transfer between the user and the system, we have proposed and evaluated a multimodal learning-to-rank method trained with relevance feedback over diversified results.

Our experimental analysis shows that the joint usage of multiple information sources positively impacted the relevance-diversity balancing algorithms. Our results also suggest that the integration of multimodal-relevance-based filtering and reranking is effective on improving result relevance and also boosts diversity promotion methods. Beyond it, with a thorough experimental analysis we have investigated several research questions related to the possibility of improving result diversity and keeping or even improving relevance in interactive search sessions. Moreover, we analyze how much the diversification effort affects overall search session results and how different diversification approaches behave for the different data modalities. By analyzing the overall and per feedback iteration effectiveness, we show that introducing diversity may harm initial results whereas it significantly enhances the overall session effectiveness not only considering the relevance and diversity, but also how early the user is exposed to the same amount of relevant items and diversity.

List of Figures

List of Tables

List of Acronyms and Abbreviations

ACM Association for Computing Machinery

BoVW Bag of Visual Words

BPREF Binary Preference

CBIR Content-based Image Retrieval

CR Cluster Recall

GMAP Geometric Mean Average Precision

GP Genetic Programming

IIR Interactive Information Retrieval

IR Information Retrieval

JASIST Journal of the Association for Information Science and Technology

kNN k-Nearest Neighbors

LTL Long-term Learning

MAP Mean Average Precision

MMR Maximal Marginal Relevance

MRA Median Rank Aggregation

MSD Max-Sum Dispersion

NDCG Normalized Discounted Cumulative Gain

RF Relevance Feedback

RRF Reciprocal Rank Fusion

SIGIR Conference on Research and Development in Information Retrieval

STL Short-term Learning

SVM Support Vectors Machine

TF-IDF Term frequency - Inverse Document Frequency

Contents

Chapter 1

Introduction

In this chapter, Section 1.1 introduces the research field and some challenges related to this thesis. Section 1.2 presents the objectives of this thesis and the research questions we investigated. In turn, Section 1.3 summarizes the main contributions of this work. Finally, Section 1.4 presents the structure of the remainder of the thesis.

1.1 Context and Challenges

The information retrieval community has continuously worked on the development of better ranking models, which produced high quality systems able to provide the user with plenty of relevant items. However, some side effects have to be considered in order to properly address the user information need. For instance, retrieving several highly relevant items may not completely satisfy the user need, if most of them correspond to near duplicates [45] or answer to a single search facet.

In this context, multimedia retrieval engines have to be robust enough for handling ambiguous or underspecified queries. Therefore, ambiguous queries may be decomposed into some, possibly several, search interpretations/aspects. For tackling these issues some studies have proposed the use of explicit [2, 170, 205] or implicit [45, 208] diversity promotion techniques. Such studies have shown that introducing diversity is helpful for enhancing user satisfaction and optimizing the search experience by adaptively retrieving more items for the most likely intent [114] or, as a post-retrieval process, at least some relevant items for the maximum number of search interpretations [172]. In turn, some complex queries inherently demand diverse results to be properly answered and may be better fulfilled with diversity promotion methods.

For instance, let us focus on a search scenario in which there is no specific single "correct answer," i.e., several different items may be considered as satisfactory, with each one carrying its particularities. In this scenario, the retrieval system may not be able to select the best item from the group of possible answers and then should provide the user with a set of possibilities. Hence, instead of biasing the result towards a, although correct, unique concept, it allows the user to browse and pick the most suitable items. In a different scenario, a given query may only be properly answered not by a single target item but by a complementary group. More specifically, these items may share common

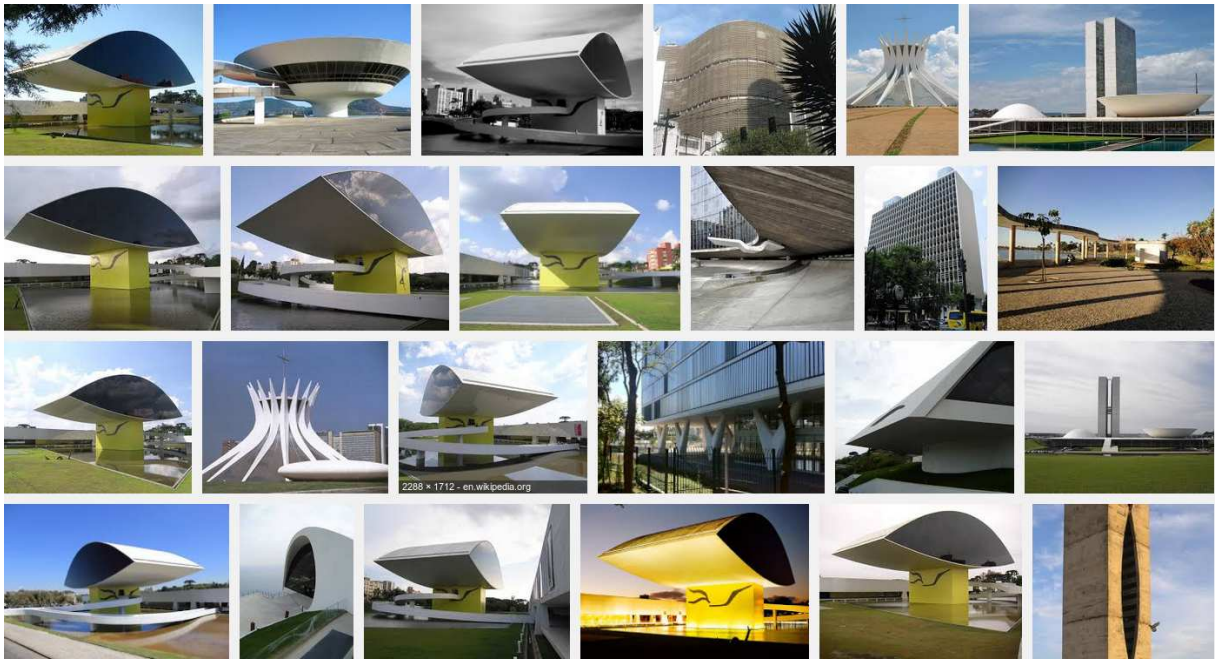

Figure 1.1: Result for "Oscar Niemeyer Buildings" on Google Images.

aspects whereas each one brings extra information. For example, in an image retrieval context, let us assume the user is looking for "a modern architecture building." This is a broad query whose answer items may differ in several aspects such as their location in the world, construction materials, usage objectives, colors, or even the perspective. Here, the system should provide the user with a set of relevant items so she can choose the most interesting to her. Differently, another user may wish to collect images of buildings designed by a specific architect, which means gathering a set of items representing the architect career portfolio and not several pictures from a famous building.

As an illustrative example, Figure 1.1 presents real results for the query "Oscar Niemeyer Buildings" on Google Images,[1] executed in November 4, 2015. The first page of results, presented in Figure 1.1 brings a total of 24 pictures from only 11 different buildings out of the hundreds of projects designed by Niemeyer, which means that the user received roughly only 46% of the total number of different buildings that could be presented in the first page. Moreover, 11 out of the 24 pictures presented, circa 46%, correspond to the same building, the Oscar Niemeyer Museum in Curitiba, Paraná.

Considering these challenges, in summary, the proposal of this PhD work was to investigate the potential benefits, drawbacks, and behavior of diversity promotion when jointly applied with multimodal and interactive strategies for learning to rank. Our hypothesis, research questions, and proposals are described in the following.

[1] http://images.google.com (with the "labeled for reuse" search option enabled) (As of October 6, 2015).

1.2 Hypothesis, Research Questions, and Proposals

Integrating multiple sources of evidence has been reported as suitable for reducing the semantic gap and improving retrieval effectiveness. In turn, interactive learning has been shown as an effective method for capturing user preferences, system adaptiveness, and for the enhancement of search success. On the other hand, diversification methods can avoid the construction of highly relevant results related only to a single search interpretation, caused by the imbalance in the number of relevant items for each query aspect or as a consequence of ambiguous or poorly defined queries [38]. However, a natural diversification drawback and research challenge we have to consider is that the promotion of diversity may erroneously place non-relevant items among the relevant ones. Finding the proper balance between relevance and diversity is still a hard task, as optimizing one may be detrimental to the other [195].

Considering these factors, the main hypotheses that guided our work are presented in the following:

- Combining multimodal information is effective on enhancing retrieval effectiveness. It should hold when diversity is considered. The integration of multiple sources of information has been shown beneficial for similarity estimation. Consequently, it is also assumed effective for diversity promotion, since implicit diversity (see Section 2.4.6) may be directly associated with similarity aspects;

- Learning-to-rank methods usually outperform traditional multimodal retrieval methods and it should also happen when diversity is integrated. The usage of diverse training data provides the system with additional information, which is assumed to be better exploited with adaptive learning models than traditional static methods;

- Diversifying retrieval results covers different search aspects and should improve user-system information exchange and consequently speed up interactive learning. The user relevance feedback over diverse results is assumed to boost learning methods since the system may obtain information that more widely covers the search/feature space and consequently enhance learning convergence; and

- Assuming learning with diversity positively impacts retrieval effectiveness, the retrieval methods using different modalities may equally benefit from the diverse information. Since diverse information boosts interactive learning, this is supposed to hold whatever retrieval evidence (textual, visual, etc.) is used.

These statements led us to some research questions that oriented our work:

- Does multimodal data integration contribute for optimizing the relevance-diversity trade-off?

- Is it possible to improve diversity with small or even no impact on relevance?

- Is it possible to boost interactive learning with diversity?

These questions drove us through a preliminary study of the many possible scenarios that could be exploited. We investigated the possibility of improving diversity with small or even no detrimental impact on relevance (Chapter 3). Moreover, we introduced an approach for enhancing the user experience by interactively learning with user feedback over diversified results produced by a multimodal image retrieval engine (Chapter 4). The experiments demonstrated that learning with diverse items helps improving overall diversity during a search session while simultaneously retrieving more relevant items in fewer feedback iterations.

Considering the successful application of diversity for interactive learning-to-rank we also studied its impact on different search scenarios. In particular, we would like to understand the relationship between diversity and multimodal retrieval approaches. Therefore, we investigated other research questions, such as:

- Regarding the potential drawback of erroneous diversification, how the "strength/intensity" of the diversification affects the overall results?
- Accounting for the scenario in which different sources of information are available for the ranking engines to process, how the diversification approaches behave when used along with different retrieval modalities and vice versa?
- Since alternative methods may differently benefit or harm with diverse promotion, what is the impact of diversification on our learning approach and on baseline fusion techniques in the context of multimodal relevance feedback?

From these questions, we propose a profound analysis with different research directions and significantly extended experimental protocol, as well as a thorough analysis of the results (Chapter 5).

1.3 Contributions

As a result of the validation of the hypotheses and research questions previously presented and our experimental investigation, the main contributions of this work are summarized as:

- Considering the relevance-diversity optimization problem, we propose two multimodal relevance and diversity promotion architectures for image retrieval;
- For interactive learning-to-rank enhancement, we propose an integrated framework for multimodal interactive image retrieval with diversity-driven relevance feedback;
- To understand the relationship between diversity and multimodality, we present a thorough experimental analysis of multimodal retrieval and the impact of diversity promotion methods; and
- For assessing the diversity benefits in interactive retrieval sessions, we conduct a broad and deep analysis of the impact of diversity promotion methods in interactive learning-to-rank scenarios.

As far as we know, no previous work has ever proposed introducing diversity in a multimedia interactive learning scenario with explicit user feedback. Moreover, although diversity promotion has gained attention in the last years, there is still a lack of development of the field and limited understanding of how to exploit its benefits and how to attenuate the known drawbacks of its application.

Therefore, as an additional contribution, we present an up-to-date and comprehensive review on IIR methods, materials, challenges. Moreover, we also present an extensive list of future work alternatives and promising research directions.

1.4 Outline of the Thesis

Considering the IIR research field and its challenges, Chapter 2 presents a broad and comprehensive review on the concepts and foundations of interactive retrieval systems. Moreover, we also present state-of-the-art literature and practical materials for IIR.

For relevance-diversity trade-off optimization, we have proposed and evaluated two multimodal strategies. This study was conducted in the context the MediaEval Retrieving Diverse Social Images Task and is described in Chapter 3.[2]

Our investigation on interactive learning-to-rank oriented by diversity is described in Chapter 4.[3] Extending and deepening this study, Chapter 5 dives into a thorough analysis on the integration of multimodal data, interactive learning, and diversity promotion.

Finally, Chapter 6 presents our conclusions, the publications directly and indirectly associated with this thesis, and the many possibilities of future work, which may benefit from the knowledge, proposals, and findings described here.

[2]The results on Chapter 3 were partially published in [28] and [29].
[3]The results on Chapter 4 were partially published in [27].

Chapter 2

Concepts and Foundations

In the last decades, we have witnessed the production and storage of huge amounts of digital objects, boosted by a constantly growing data generation rate. Human beings and electronic devices have never generated so much data in such a short time [121]. These factors were promoted by important advances related to data capturing and sharing. Moreover, with the popularization of the Internet and mobile devices, a great portion of previously consumer-only people became prolific data producers.

For image retrieval applications, given the technological advances of the last decades, the resources for capturing, processing, sharing, and retrieving have tremendously evolved. Currently, with the convergence of multimedia devices and the broad access to the Internet, the production and sharing of images is an easy and popular activity. Therefore, with so much data around, the information technology industry is challenged to deliver more effective and efficient indexing and searching engines.[1]

When dealing with large repositories, finding data, which are relevant to a given user query, context, or information need, becomes a hard task. For instance, considering unstructured or multimedia data, traditional search methods rely only on textual metadata as a source for relevance estimation and ranking, implying important issues related to annotation costs and accuracy. Using textual annotations is also subject to language problems related to synonym and polysemy. In a different paradigm, with the advances on data processing capabilities, content-based methods[2] for large-scale scenarios became an important and complementary alternative. However, low-level features, widely used for multimedia data applications, such as image and video retrieval, quite often are not able to properly represent data concepts and user preferences, causing the well-known semantic-gap problem [207].

The obstacles naturally present in information retrieval tasks range from the cost of large-scale data annotation to the subjectivity of user search intents. Moreover, researchers have faced many theoretical and practical difficulties for conducting experimental studies and performing data analysis. In spite of the great advances from the

[1]The methods proposed in this thesis and their experimental analyses are focused on effectiveness aspects. Nevertheless, we also provided soft discussions on efficiency throughout the text.

[2]These methods extract information from the object itself instead of relying on metadata/annotations. For instance, the whole content of an article may be processed instead of considering only, e.g., its title, abstract, or keywords.

last decades [110, 194], the information retrieval community, specially on multimedia retrieval, still suffers from the absence of well-established standards, e.g., when considering user-system interaction models, evaluation protocols, and benchmarks.

Content-based image retrieval (CBIR) systems exploit low-level features within the images for measuring visual similarity. While CBIR is very effective for some tasks [42], the adequate encoding of high-level visual concepts through low-level visual features is a hard issue. Quite frequently, for rich queries on heterogeneous collections, the low-level features cannot accurately encode the visual concepts of the images. In addition, several works have reported the complementarity of textual and visual information [12, 25, 30]. These methods have attracted great attention from the research community as the joint usage of various information sources could be useful for attenuating the semantic gap.

Another important issue is that different users quite often have disparate interpretations of a same image or the same user may have different perceptions at different times, making the retrieval task much more difficult. Additionally, users are not always able to properly express their information needs, meaning that retrieval systems have to process poorly defined queries. In order to make systems adaptable to different users, Relevance Feedback (RF) [245] has been applied to aid per-user system optimization. In this context, users can help the system to refine results by providing feedback about the relevance of the results. The user interacts with the system by implicitly or explicitly providing relevance assessments for the retrieved items.

The system exploits the feedback information to expand queries and enhance internal learning models. These models are employed in order to refine, customize, and present novel retrieval results, which are supposed to better correspond to the user needs. Consequently, introducing user perception to retrieval methods became an important asset for effectiveness enhancement and search personalization. Hence, by interactively exchanging information with the system, the user allows her preferences to be learned and optimized, which may significantly improve the search experience. Several studies have shown the ability of relevance feedback to improve both retrieval effectiveness and user satisfaction [162, 167].

In the effort for jointly exploring several information related sciences (information retrieval, machine learning, human-computer interaction, computer vision, data mining, etc.), interactive information retrieval (IIR) became a very active research field. Moreover, for boosting the user-system knowledge transfer and personalization, recent works have gone beyond simple relevance feedback towards integrating more diverse information and techniques into the interactive search process.

Interactive learning has been explored in the information retrieval field for decades with the purpose of tackling several inherent issues. The possibility of including the user in the retrieval loop has allowed significant effectiveness enhancements over time. By taking advantage of all the data available and the collected user preferences, learning-to-rank models [129] leveraged online adaptiveness and consequently improved user search experience.

This chapter reviews several interactive retrieval related aspects focusing mainly on recent advances, important challenges, and promising research directions. We have selected and described several works from important conferences and journals. The main

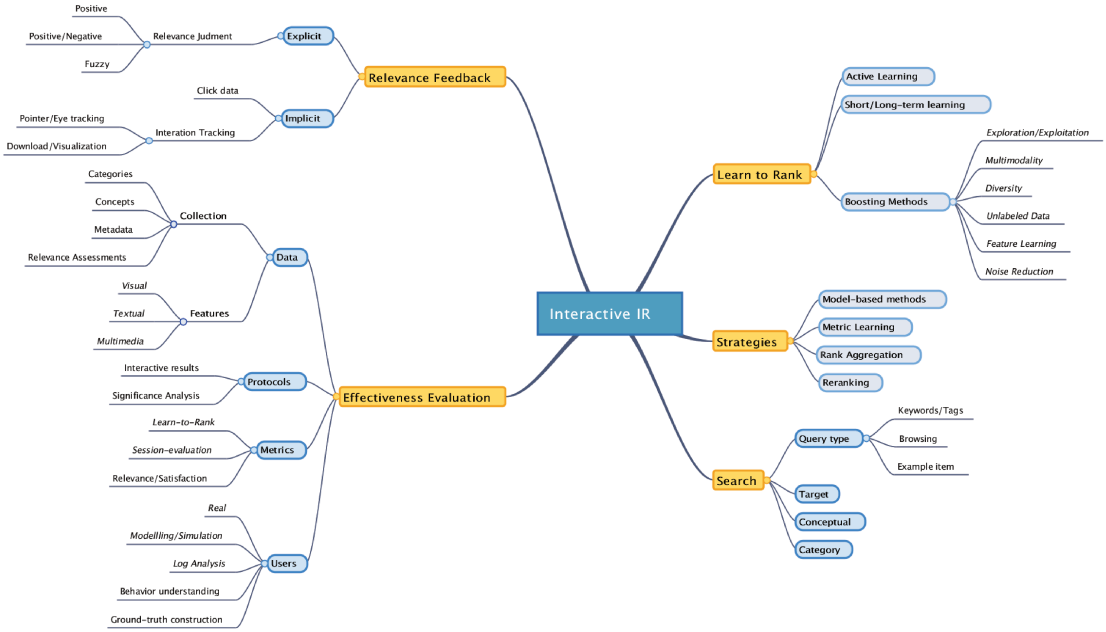

Figure 2.1: Conceptual map of the Interactive Information Retrieval field.

publication venues and periods consulted in this work were: (i) Conferences: CBMI (2011-2014), CIKM (2011-2014), CLEF (2011-2013), ECIR (2011-2014), ECML-PKDD (2011-2014), ICIP (2011-2014), ICME (2011-2014), ICMR (2011-2014), SIGIR (2011-2014), and WSDM (2011-2015); (ii) Journals: IEEE-MM (2011-2015), IEEE-TCOMP (2011-2015), IEEE-TIP (2011-2015), IJMIR (2012-2014), JASIST (2011-2015), JVCIR (2011-2015), MTAP (2011-2015), PR (2011-2015), and PRL (2011-2015). Important works from other venues were also considered. Our focus is on recent works that exploit mostly machine learning techniques and multiple modes of information (textual, visual, etc).

As a broad and comprehensive representation of the IIR field and consequently of the structure of this chapter, in Figure 2.1, we present a conceptual map covering several foundation areas and aspects that are integrated for the construction of modern interactive retrieval systems. As an overview of the IIR literature covered in this work, Table 2.1 presents a categorization and representative works on the concepts from Figure 2.1.

The remainder of this chapter is organized as follows. Section 2.1 summarizes the findings of previous overview works on the interactive retrieval field. Section 2.2 overviews traditional concepts on IIR and recent works. Next, Sections 2.3 and 2.4 describe common learning-to-rank strategies for IIR and recent boosting alternatives, respectively. With regard to experimental evaluation and user aspects, Sections 2.5 and 2.6 present common and new experimental and modeling theoretical and practical tools. Section 2.7 illustrates several interactive multimedia retrieval applications. Finally, Section 2.8 presents our final considerations.

A bibliometric analysis of the main recent publications discussed throughout this text is presented in Appendix A. In turn, Appendices B and C describe the main open challenges and promising research directions in IIR.

Table 2.1: IIR Concepts and Representative Works.

IIR Concepts	Representative works
Interactivity	
Relevance Feedback	Distance-based learning [7, 8], Random walks [118, 166], Graph Cuts and Manifold learning [242], Evolutionary methods [8, 30, 59, 95], Query-point movement [67], Query expansion [225], Query Reformulation [104], Implicit vs. Explicit RF [244].
Active Learning	Most positive/informative samples [41], Uncertainty/Density [44], Positive/Negative samples unbalance [220, 225].
Short/Long-term learning	Short-term [9, 151], Long-term [163], Short and long-term fusion: [163, 225, 228].
Learning Strategies	
Classification-based methods	SVM [44, 63, 95, 220, 242], Evolutionary algorithms [9, 27, 30, 95], Logistic Regression [23], Optimum-path forest [41].
Metric learning	Kernel Combination [232], Similarity function optimization [30, 59], Features and components weighs adjustment [176].
Rank Aggregation	Ranked-lists fusion [161].
Reranking	Multi-instance learning [46], Reinforcement learning [125].
Learning Clues	
Exploration X Exploitation	Interleaving [9, 79], Redundancy minimization [184], Exploration-exploitation nesting [123].
Diversity	Dynamic ranked retrieval [23], learn from diversity [27].
Unlabeled Data	Heuristic selection [230], Subspace learning [243], Contextual information [151].
Noise Reduction	Feedback samples similarity [87].
Feature Learning	Dynamic visual dictionaries [63, 64, 132], Adaptive feature space [219].
Multimodality	Multimodal feature space [13], Multi-form image representation [67], Multimodal ranking functions [27].
Experimental Evaluation	
Protocols	Rank-shit [123], residual collection [161], freezing [27, 111].
Datasets	PICv1 [237].
Measures	Learning-to-rank [235], Session-based [98, 104].
User Aspects	User modelling [15, 19], Ground-truth generation [71], judgment effort analysis [71, 209].

2.1 Previous Works

Thomee & Lew (2012) [194] presented an overview on interactive image retrieval (IIR) considering all papers in ACM, IEEE, and Springer digital libraries on the subject of interactive content-based image retrieval over the period of 2002 to 2011 (over 170 papers). The authors provided a detailed review by clustering interactive search topics according to the user's point of view and the system's point of view. On the user's perspective, the authors described trends and advances related to query specification, types of retrieved results, user interactions, and retrieval interfaces. On the other hand, considering the system-centric analysis, the authors described advances and trends related to image representation, indexing and filtering, active learning, common similarity measures, and long-term learning. Furthermore, the authors discussed several issues and advances related to the evaluation and benchmarking of interactive systems considering image databases and effectiveness measures. The authors concluded by presenting promising research directions.

As described in [194], from the user's point of view, the general interactive search process starts with the query specification. The system provides an initial result set and the user interacts by providing feedback. The query specification process may occur using descriptive texts [113], example images [1], random selection of images from the database [193], selected segmented regions [36], and outlines [5]. An interesting approach starts the search using keywords (possibly selected from a thesaurus) and allows the user to provide visual region selection on the result [198]. The results are usually presented as a ranked list of items that may include the best matching images and/or the most informative ones [88]. The interaction with the user continues with feedback that may be provided using different possibilities of relevance levels: positive only [100], positive/negative [239], positive/neutral/negative [233], or multiple/fuzzy relevance levels [91]. The user feedback may also be collected using region selection on images [199] or implicitly, according to user's actions [35]. The input, results, and feedback may include items from different modalities [134] (text, audio, images, etc.). In turn, the development of new interactive interfaces have focused on better collection browsing [192] and results presentation [217], as well as handling multiple query and feedback modalities [201] (grouping, region selection, image marks, etc.). Finally, the user-centric trends and challenges are related to: region-based retrieval; clustered/linked/3D results interfaces; and multi-modal (input/output/feedback).

On the other hand, regarding the system's point of view, the first aspect we have to consider is the image representation. In the last years, we have witnessed the shift from low-level to mid-level, and high-level image representations, including the bag-of-visual-words approach [103]. In interactive retrieval, this approach can be explored for target visual words prediction based on user feedback. Consequently, the system is able to rank images using not only low-level features, but also higher-level visual words. For efficiency improvement, recent work has explored indexing and filtering alternatives. For instance, clustering techniques have been used to reduce the number of candidate images, as well as hierarchical and hashing indexing structures.

Regarding effectiveness enhancement, a quite common approach is the use of active learning methods. Active learning is used to reduce the interaction effort and maximize accuracy, by choosing the most informative images, while promoting the diversity among the samples to be labeled. Moreover, the information obtained with the feedback can be used to create better models for the feature space. For this purpose, recent works have explored several directions, such as: Feature selection and weighting – principal component analysis (PCA) [60], discriminant component analysis [81] or linear discriminant analysis [86]; Manifold learning [22]; Synthetic and pseudo-imagery [191]; Learning Methods – Artificial Neural Networks [226] and SVM [240]; Kernels [229]; Learner combination [189]; and Probabilistic classifiers [4].

Similarly, long-term learning approaches (see Section 2.2.3) have been studied with the objective of efficiency and effectiveness improvement. In this line, inspired by recommender systems, collaborative filtering approaches have been used to accumulate information about different users. This information may be obtained from log analysis and used for reducing the interaction effort, improving retrieval accuracy, and reducing the processing time. Considering the aspects related to similarity measures and collection ranking, recent work has considered not only the relevance according to a query, but also how close the image is to the nearest relevant and the nearest irrelevant neighbor. At the same time, great effort has been made for better combination of multiple similarity measures.

According to Thomee & Lew (2012) [194], trends and advances related to the system's point of view focus on: tackling the small training set problem; handling many clusters of positive images or closeness of relevant and irrelevant clusters; concept-based retrieval with high-level features using bag-of-words, manifold learning, long-term learning, and multiple information sources.

Li & Allinson (2013) [124] presented an overview of relevance feedback-based methods for Content-based Image Retrieval (CBIR). Different from [194], in [124], the RF methods were grouped according to two learning models: short-term learning and long-term leaning. The authors also provided some insights on future work and research directions. The authors report that relevance feedback is a technique that leads to improved retrieval performance by the update of query and similarity measures based on user's preference. With the use of relevance feedback, the traditional short-term learning and also more recent long-term learning methods allow improving the retrieval performance in terms of effectiveness and efficiency. The authors also highlight that most long-term learning techniques are jointly applied with short-term methods and improved retrieval performance has been reported in terms not only of effectiveness, but also of efficiency.

As a historical analysis, the work from Kelly & Sugimoto (2013) [110] overviews 40 years of Interactive Information Retrieval (IIR) evaluation works (1967 to 2006). From 2791 journal and conference papers, 127 were selected for systematical analysis. The works were coded using features such as author, publication date, sources, and references. Moreover, the properties of the research method used were extracted, such as the number of subjects, tasks, corpora, and measures. In a bibliometric analysis, the results reveal the growth of IIR studies over time, the most frequently occurring and cited authors and sources, and the most common types of datasets and measures used.

Similarly to [194], the authors of [110] defined different scopes for the IIR studies. Some works were defined as system-focused, which do not use real test subjects, but there may be a human involved on topic creation and result evaluation. Other studies were characterized as primarily focused on understanding the information-seeking behavior just like it naturally happens in different contexts. Alternatively, as previously described by Kelly (2009) [108], the works that fit both descriptions were defined as the classic core of IIR. Such works include experiments conducted for evaluating the systems and also the retrieval interfaces. Although the IIR research evolves based on many different studies, the evaluation efforts are considered as a core component in which the system-oriented and user-oriented approaches are jointly explored.

The bibliometric analysis in [110] also revealed IIR as a relatively young field with most of the research works published at the late part of the review period. They have also noticed that it is was also a concentrated research field with half of the publications only in three venues: JASIST, IP&M, and SIGIR Proceedings. This fact has changed in last few years with IIR works published in several conferences and journals, as we show in Appendix A.

Complementary to [110, 124, 194], this chapter targets reviewing IIR concepts and foundations and broadly covers recent proposals and current open issues and challenges. As mentioned before, we focus on recent work, mostly related to machine learning strategies (e.g., learning-to-rank), considering multiple modes of information. In this sense, our work updates and complements previous efforts in summarizing and understanding such a multidisciplinary research field.

2.2 Interactive Retrieval

According to Kelly & Sugimoto (2013) [110], "*interactive information retrieval (IIR), blends research from information retrieval (IR), information behavior, and human computer interaction (HCI) to form a unique research specialty that is focused on enabling people to explore, resolve, and manage their information problems via interactions with information systems.*" In the image retrieval context, Thomee & Lew (2012) [194] state that the interactive search methods are developed for finding relevant imagery by allowing an interactive dialog between the user and the search system. The interactive methods are also useful on scenarios when the user cannot express the concepts she has in mind by a known word.

In this section, we review some concepts related to interactive retrieval systems such as learning-to-rank applications. Therefore we consider Relevance Feedback and its Implicit and Explicit variations (Section 2.2.1), Active Learning strategies (Section 2.2.2), and Short-term and Long-term Learning (Section 2.2.3).

2.2.1 Relevance Feedback

Relevance feedback is a common interactive retrieval technique that allows the user to provide the system with relevance grades for the items retrieved in response to a given query. It can be applied for instance in order to reduce the semantic gap between user

information need and low-level extracted features. Basically, the user receives the group of items retrieved and judges their relevance in relation to her information need. Usually, the user can mark each retrieved item as positive (relevant) or negative (non relevant). Some methods also allow the neutral grading or even multiple grading levels for positive and negative samples.

In summary, one can classify the RF techniques into three groups: explicit feedback, implicit feedback, or pseudo feedback. The first two regard if the relevance information is explicitly provided by the user or automatically captured by the systems by monitoring user interactions. The pseudo feedback is an automatic feedback method that does not require user interaction. For instance, a system can collect items considered as relevant with high classification confidence, and automatically use them as positive samples for improving the search results [46].

Recently proposed relevance feedback approaches have relied on several methods such as Random walks [118, 166], Genetic Algorithms [8, 95], Graph Cuts [242], Manifold learning [242], Distance-based methods (e.g., kNN) [7, 8], Genetic Programming (GP) [30, 59], Query-point Movement [67], Query Expansion [225], and Query Reformulation [104].

In [118, 166], positive and negative feedback samples were used as staring point for random walks. The ranking scores of the unlabeled items were computed as the probability that a random walker in the graph starting at that image reaches a relevant sample before finding a non-relevant one.

In [8], the authors combined genetic algorithms and distance-based learning for relevance feedback in CBIR. The feature vectors of positive samples were genetically evolved towards positive regions of the search space. For mapping the evolved genotypes to real images, a distance-based method was applied considering also the negative samples obtained from user feedback. Similarly, in [95], the authors boosted an SVM-based RF approach by optimizing feedback samples' features using genetic algorithms.

In [242], a method was proposed to combine manifold structure information and visual features using a graph-cut method based on an energy minimization approach.

As discussed in [7], distance-based methods and similar approaches (e.g., margin-based) suffer from problems such as unbalanced number of positive/negative samples, small sample sizes, variations of the feature space density and the lack of representativeness of the labeled samples. To overcome such problems, the work in [7] successfully incorporated a reliability factor for estimating relevance, which in practice combines the distance to the nearest positive and the nearest negative neighbors for relevance probability estimation.

In this thesis, RF is used for interactively capturing user preferences over retrieval results which was subject to diversity promotion. The user is supposed to indicate relevant items whereas the items not marked are implicitly considered as non-relevant. With this diversity-oriented feedback, the learning system is supposed to discover better multimodal ranking functions than when diversity is not promoted.

Other RF-based learning-to-rank proposals and strategies are described throughout this section. For the interested reader, we refer to Section 2.3 and Section 2.4 for more details.

Implicit vs. Explicit RF

While very useful for system-user adaptiveness, explicit relevance feedback is not an easy task and users may not be interested in providing relevance grades through many iterations. As an alternative, user interactions may be captured and reasoned as implicit feedback signals. Common user interactions are click on a link, document download, image visualization, mouse hovering, and page inspection time. Alternative signals can be captured as multimodal feedback including eye tracking, voice commands, screen touching, and gestures.

Although explicit and implicit RF present different practical challenges and information gain potentials, some works suggest that their combined usage may be beneficial to the overall system effectiveness and user satisfaction. For instance, Zhang et al. (2013) [244] proposed a hybrid RF method that combines explicit graded relevance feedback from the user with implicit information obtained from user browsing behavior. The images' grading values and implicit preference values were used to iteratively train a (SVM) preference-based classifier for determining the search results after each feedback iteration.

2.2.2 Active learning

One significant goal of interactive learning is maximizing the information transfer between the user and the retrieval system. The objective of active learning strategies is to select the items from the collection which when labeled by the user will help to optimize the results in the next iteration. Additionally, by selecting the proper unlabeled samples for user judgment, the system aims at reducing the number of samples that are necessary to train internal models, moving the search towards relevant items faster.

In this context, instead of providing the user with the most positive (relevant) items, the system may proceed through some iterations retrieving the most informative (close to the classification boundary) items. After a few iterations and the labeling of a "proper" amount of informative items labeled, the system may use the cumulated information to generate the final result list. Some works have also combined these strategies by including the most positive and most informative items in every iteration with different participation rates. The amount of positive and informative items can be dynamically adjusted according to result convergence or user satisfaction. Sharing some of these goals, the exploitation-exploration trade-off methods and the diversity promotion approaches are discussed in Section 2.4.1 and Section 2.4.6, respectively.

Differently from traditional active learning techniques, which explore user feedback for the most ambiguous (relevant and irrelevant) samples, in [41], an active learning model is proposed for feedback over the most informative samples selected only from the set of relevant images. The method in [41], based on the optimum-path forest classifier [148], requests feedback for the items classified as relevant that are also close to irrelevant samples. For this, the relevant items are ranked according to the absolute cost difference to positive and the negative prototypes with optimum cost. The prototypes are part of the Optimum-path forest technique and are the training samples that link relevant and irrelevant paths on a minimum spanning tree constructed with the training samples. The

authors state that this strategy reduces the number of false positives. The experimental evaluation has shown more effective performance when compared to a traditional SVM-based active learning method [197], with significantly lower processing time.

In the context of remote sensing image retrieval, the work described in [44], which extends [43], proposed an active learning method based on uncertainty, diversity, and density. The uncertainty and diversity criteria aim at maximizing the classifier accuracy. In turn, the density criterion aims at finding representative samples of the image distribution on the feature space. For exploring uncertainty, the samples for user feedback are initially selected with a traditional margin sampling SVM approach. These most informative samples are clustered for diversity purposes using a kernel-based k-means clustering technique. Finally, from each cluster, a representative sample is selected according to a density criterion based on the average distance from each image to all other images in the cluster. This method outperformed a similar SVM-based active learning approach with marginal sampling and a diversity criteria based on the distance between the most informative samples [55]. These results highlight the importance of the representativeness of image distribution on the feature space which, in this case, was targeted using samples from high density regions.

The work in [220] presented a comprehensive overview of SVM-based relevance feedback and active learning methods and highlighted related open issues. Relevance-based ranking using SVM classifiers, specially with a few training samples, often outperformed other learning alternatives. Nevertheless, some limitations are still present. Such difficulties, attenuated over time, are related to the SVM methods' limitations on equally handling positive and negative samples and on differentiating the relative relevance among positive samples. Moreover, these learning methods suffer from the fact that positive samples may be clustered in the feature space while the negative samples can be widely spread. Additionally, good effectiveness was frequently achieved with proper parameter optimization and can be quite affected by unbalanced number of samples from the different classes. Hence, for attenuating such issues, the authors in [220] proposed the ensemble of sub-features vectors specialized classifiers. Moreover, for enhancing previous similar ensemble proposals, a weight vector for component classifiers was dynamically computed from positive and negative samples which allowed superior effectiveness.

Specifically for the realm of active learning for learning to rank, in [179], a *lazy* association rule-based active method is proposed, which selects a small training set from scratch (which is essentially the method originally proposed by the same authors in [178]). This seed set provides the basis for the application of a query-by-committee (QBC) second-stage method to improve and expand the selection, yielding state-of-the-art results on the LETOR 3.0 web datasets (see Section 2.6.2 for a better description of the datasets). The first phase of the proposed technique depends on a loosely defined concept of "diversity" (e.g., "exploration"): intuitively, the association rule method tries to "cover" the feature space with the minimum number of representative instances, whilst the QBC stage depends on the variation of the committee models and algorithms to select "interesting" (e.g., "exploitation") instances from those remaining in the unlabeled set. This is the only method that tries to apply both active learning "objectives," albeit in a two-stage manner. Although the method yields good results, it is extremely inefficient since, by

being lazy, it generates a model for each single unlabeled instance to be evaluated, and thus does not scale to be used in datasets larger than a few thousands of documents.

As previously described, active learning methods usually work in a two-step strategy. Therefore, it includes a training/learning phase from feedback over actively selected result items and a second phase in which all the knowledge cumulated through iterations is used for generating the final result which will be presented to the user as the answer for her query. Analogous to active learning methods, the diversity-oriented proposal of this thesis intends to present the user retrieval results that may speed up the preference learning. However, in our proposal, there is no preliminary active training step and the system is supposed to present the user the most relevant and diverse result at every iteration of the retrieval session, which means targeting both objectives simultaneously.

2.2.3 Short-term and long-term learning

The traditional interactive learning methods described in the previous sections usually provide system optimization and user adaptiveness considering only the feedback information obtained for a given query session, named short-term learning (STL). However, in such methods all the optimization effort and constructed knowledge is immediately lost at the end of the session since no information is stored for speeding up the learning on further sessions. Hence, for taking advantage of historical interactive sessions, several works have been proposed on long-term learning (LTL) of semantic relationships among the images of the collection. Different from STL methods, which rely only on intra-query learning, LTL takes advantage of relevant patterns discovered at previous iterations. Consequently, this accumulated knowledge can be exploited for reducing the labeling effort and improving retrieval results.

The STL, a.k.a intra-query learning, methods explore the information obtained from a single retrieval session. As described in [124], these methods can be categorized regarding how the labeled samples are treated, such as:

1. One-class (for positive samples only): these approaches focus the learning procedure on most positive samples, e.g., SVM with sphere hyperplanes, in which the inner one embraces most of the positive samples whereas the outer one pushes negative samples away. Other methods lately applied were PCA and Gaussian Mixture Models (GMM);

2. Two-class (one class for positive samples and another for negative samples): these approaches focus the learning procedure on informative samples. The most common approaches are active learning SVM, co-training techniques, random subspace methods, asymmetric bagging, and manifold learning; and

3. Multi-class (several classes for positive samples or negative samples): these methods are modeled as non-binary classification problems for handling multiple positive/negative classes.

In turn, the LTL methods aggregate user log information along feedback sessions. These methods can be categorized regarding how the knowledge is used, for instance:

1. Latent semantic indexing-based techniques: among such methods the most commonly used is the Singular Value Decomposition. Chen et al. (2005) [34], for instance, explored semantic regions segmented from images and user feedback for constructing the long-term knowledge base;

2. Correlation-based approaches: these methods rely on the creation of sets of images that are semantically correlated. Therefore, the LTL can be performed by putting the relevant items for a query into each other's peer index whereas the removal is performed for irrelevant samples. The correlations between images in the database and the current feedback can be estimated by collaborative filtering. Urban & Jose (2006) [200], for example, proposed an image-context graph for representing the correlation between images, terms, and low-level features;

3. Clustering-based algorithms: these methods can be used to refine retrieval results using the information from conceptual groups of semantically related items accumulated from previous feedback sessions. For instance, Han et al. (2005) [72] proposed semantic-correlated clusters constructed based on co-positive-feedback frequency and the co-feedback frequency between the images;

4. Feature representation-based methods: these methods try to improve retrieval effectiveness by properly adjusting relative feature weighs using accumulated feedback information [39]; and

5. Similarity measure modification-based approaches: once a feedback session is finished, the internal relevance scoring functions are adapted based on the provided feedback. Therefore, this adjusted score can be used in future sessions [82].

LTL methods usually rely on storing pairwise relevance correlation, usually aggregated on an affinity matrix between images or between images and semantic concepts [163]. The semantic relationships between images can be extracted by analyzing user interactions over time on multiple retrieval session logs. Using STL and LTL knowledge allows not only computing and adjusting relevance to queries according to, e.g., visual similarity, but also considering semantic relationship scores. A list and brief description of several previous LTL methods can be found in [228].

Some common difficulties inherent to RF-based systems are the availability of just a few training samples, the imbalance between the amount of positive and negative samples, and also the labeling effort and high computational costs. For attenuating these issues, Wu et al. (2013) [225] proposed not only combining short and long-term learning, but also integrated semi-supervised learning and active learning sessions in a CBIR system. In that work, the long-term knowledge and random sampling was exploited for extending and balancing the positive and negative training data, respectively. The resulting samples were used in a semi-supervised process for optimizing visual similarity and consequently

the retrieval effectiveness. For efficiency purposes, the visual similarity between unlabeled images to the positively and negatively labeled sets from previous iterations are incrementally computed and the cost is reduced to the similarity computation in relation to the current feedback samples. For the final ranking, the semantic and visual similarities are non-linearly combined. This combination of several effectiveness and efficiency techniques allowed outperforming several methods that rely on semi-supervised, active learning, and/or hybrid short/long-term learning methods.

More recently, for content-based image retrieval with relevance feedback, Xiao et al. (2014) [228] proposed integrating short and long-term information using a simple weighted linear combination of a visual-based short-term similarity score and a high-level long-term-based semantic score. The visual score is computed and updated using the amount of relevant samples obtained from feedback. The long-term procedure relies on storing and updating the semantic correlation of images for a set of queries and the semantic descriptions of the queries were constructed according to the semantic features from the positive feedback samples.

Alternatively, Rashedi et al. (2015) [163] evaluated different fusion methods including fusion of retrieved images, rank fusion, and similarities fusion. Additionally, a statistical semantic clustering method was proposed for long-term learning and reasoning. The proposed long-term method relies on detecting the proper semantic category of a query using positive and negative feedback samples present in the already discovered semantic categories available in a learning knowledge base. If no existent semantic category adequately fits the new query then a new category is dynamically created using the feedback information. During the learning process, similar categories may also be merged for unifying semantically close samples.

Although some works have shown the benefits of aggregating long-term information for retrieval enhancement, in this thesis we rely only on intra-query learning given diversity aspects are still not well understood and must be carefully evaluated before introducing more complex methods. Therefore, the learning procedure with feedback over diverse data is conducted only along each search session.

2.3 Interactive Learning Strategies

Applying machine learning techniques is a common procedure for knowledge construction according to implicit or explicit user interactions. In this section, we describe several interactive learning proposals that explore effectiveness improvement techniques such as Model-based methods (Section 2.3.1), Metric learning (Section 2.3.2), Rank aggregation (Section 2.3.3), and Reranking methods (Section 2.3.4).

The frameworks proposed in this thesis were constructed as a combination of methods from all the categories of strategies mentioned. The methods used for specific modules are described in Chapters 3 and 4.

2.3.1 Model-based methods

Beyond feature weight adjustment and (multi) query-point movement [13], several interactive learning-to-rank approaches model the RF task as a classification problem for separating relevant from non-relevant samples according to user preferences. Among the model-based learning methods, the most commonly used is the SVM technique as in [63, 95, 242]. In these methods, the labeled samples are used to construct separation hyperplanes using positively and negatively labeled samples as training instances.

For greedy methods, the items classified the farthest from the separating hyperplane in the positive side are selected as the next samples for answering the user query and posterior labeling. Differently, in active learning approaches like the proposed in [44, 220], the samples that are the closest to the separating hyperplane are selected as the most informative items that when labeled may provide the best contribution for the model improvement and hyperplane adjustment.

As described in Section 2.2.1, besides the SVM technique, several other machine learning methods have been explored for capturing user preferences such as Genetic Algorithms [9, 95] and Programming [27, 30], Logistic Regression [23], Optimum-path forest [41], etc. Since classification-based methods are the most common approaches, consistently covered in the literature, and applied in several works described in the next sections, we do not include further details here and also direct the interested reader to the work in [130].

2.3.2 Metric Learning

Analogous to feature components weight learning, when retrieval systems consider multiple features, with early or late fusion approaches [181], users' preferences may be explored for adjusting inter-feature importance and have been successfully applied for steering the search engine towards the features that more properly represent the high-level user needs. This learning alternative is also usually applied in multimodal systems as described in Section 2.4.5. For instance, the authors in [232] proposed a RF method using cost functions for distance metric learning for the linear combination of multiple kernels. The local analysis conducted with user's feedback for the adjustment of base kernels weights outperformed baseline methods with global optimization (SVM-RC [175] and LMNN [222]).

For the automatic and adaptive combination of similarity functions from different visual features, the work in [59] proposed a genetic programming framework for CBIR with RF. This method considers user feedback for creating better similarity combination functions that more adequately express the user need. Therefore, the ranking functions are evolved using positive and negative feedback images as training samples. Similarly, in [30], the authors proposed a multimodal image retrieval framework that uses GP for the combination of similarity measures from visual (e.g., color and texture) and textual (e.g., BM25 and Cosine) features. This method creates optimized multimodal functions that automatically adjust the importance of the different modalities and the different features from each modality according to user preferences expressed through RF.

Alternatively, with a hybrid approach, Shamsi et al. (2014) [176] proposed not only adjusting the different feature weights, but also the weights of each component of the

features. The weights of the feature components were adjusted according to the mean and standard deviation values of the features of relevant samples from feedback, while the weight for each feature was adjusted according to the rank positions of the relevant samples on feature specific ranked lists.

2.3.3 Rank Aggregation

Interactive learning methods based on ranked lists fusion work by requesting and exploiting user relevance feedback for the items present in a single list, which is actually created from the fusion of different, possibly several, intermediate lists. These intermediate lists are constructed, e.g., using different retrieval models or features. While some works have applied rank fusion strategies for traditional IR tasks, there are limited research when it comes to IIR approaches.

For exploring relevance feedback over fusion-based improved ranked lists, the work in [161] proposed a meta fusion method that combines different fusion scores in order to create the final ranked list considering not only the relevant items from user feedback, but also the inherent effectiveness of the intermediate lists. The first score is computed based on a query expansion ranking model using the positive feedback examples, whereas the second considers the relative effectiveness of the intermediate lists for weighing the document scores. The proposed meta fusion method simply applies a weighted linear combination over the two ranking scores. The experimental results have shown significantly superior effectiveness when compared to standard single ranked list settings.

Leveraging the relative effectiveness of intermediate lists is an interesting optimization technique as it allows the automatic definition of the importance of the ranks constructed using different ranking functions, features, or even query modeling approaches.

2.3.4 Reranking

Considering that the retrieval results are usually not optimal and the existence of noisy items even when highly effective ranking methods are applied, using reranking methods allows integrating multiple sources of information in order to refine initial results. Let us examine for instance the multimedia retrieval tasks. As highlighted in [133], text-based approaches have achieved limited success by not including all the information encoded in different modalities such as visual content or audio features. For enhancing text-based multimedia search, many works have proposed visual reranking strategies for improving initial results lists constructed only using textual metadata. In fact, reranking strategies can be applied for improving results in cross-modality tasks or even when multiple features from a single modality are combined.

For improving ranking on text-based web image search, the work in [46] proposed a bag-based reranking model using textual and visual features. Accordingly, the images initially retrieved using user-provided tags were reranked using a bag-based multi-instance SVM model. The multi-instance methods [6] assume that a positive bag contains at least one relevant instance while there are only irrelevant instances in negative bags. Therefore, the learning procedures consider only the bag labels instead of instance specific labels.

In [46], for creating the training bags, the initially retrieved images were clustered using textual and visual features. The clusters were ranked according to the average ranking scores of the images in each cluster. The highest ranked clusters were used as pseudo-positive bags for training a multi instance SVM. Differently from traditional bag-based methods, in [46], relevant and irrelevant bags are assumed to contain a given proportion of relevant instances, i.e., a given bag is considered irrelevant if it does not contain enough positive samples. Alternatively, the authors also evaluated the effectiveness of manually labeling the bags by user simulation. The pseudo-feedback method outperformed several baselines including [83, 246]. Additionally, the user labeling simulation allowed further effectiveness improvements over the pseudo-feedback method.

Also exploiting image social tags, the work in [125] proposed a multimodal relevance feedback method for image reranking boosted by an image-tag relationship graph model. The image-tag graph was optimized by a mutual reinforcement approach, i.e., the scores of images connected to high-ranked tags and the scores of the tags connected to high-ranked images were increased. The relevance feedback information (positive/negative images/tags) is used to adjust the scores of the labeled samples in the graph which are iteratively propagated through the graph with the reinforcement process. This method achieved superior effectiveness in relation to the several baselines including traditional query-point movement, SVM-based RF [241], VisualRank [101], and clustering-based reranking [83].

For the interested reader, an extensive overview of reranking methods, as well as the description of several previous interactive reranking proposals can be found in [133].

2.4 Learning Boosting Clues

In this section, we review several information sources used for boosting the interactive learning methods, which go further than only capturing implicit or explicit relevance feedback. We consider important recent contributions on the Exploration-Exploitation dilemma (Section 2.4.1), Diversity Promotion (Section 2.4.6), Semi-supervised learning (Section 2.4.2), Noisy Feedback Reduction (Section 2.4.3), and other alternatives such as Feature learning (Section 2.4.4), and multimodal feature combination (Section 2.4.5).

2.4.1 Exploration and Exploitation

Hofmann et al. (2011) [79] regard exploitation as a step that uses what has already been learned to produce relevant results, while exploration is the search for new solutions to obtain feedback for effective learning. According to Suditu & Fleuret (2012) [184], in the exploration phase, the user informs the system in a broad way which categories are of interest. On the other hand, during the exploitation phase, the user provides more detailed requirements on the visual properties of the search interests and the system can more effectively handle the subset discovered during exploration. More recently, Arevalillo-Herráez et al. (2015) [9] stated that exploitation approaches focus on the search inside the frontiers of previous relevant retrievals, attempting to exploit already known regions

of interest of the feature space. Differently, exploration methods focus on finding other relevant areas.

On-line learning to rank is considered a promising approach specially for applications with little training data available or when collecting a large amount of training data is a costly task. For instance, it is useful for learning user preferences on newly deployed systems. Nevertheless, the information gathered through this kind of system is in general biased towards the limited amount of items that are examined by the users, frequently not reflecting the actual information distribution of the existing data. Moreover, these issues avoid the exploration of different but equally relevant solutions that circumstantially do not exactly fit the current extracted knowledge. For dealing with such issues, besides using the already learned ranking models, the systems can expand retrieval capabilities by explicitly exploring new different solutions, for instance different regions on the feature space. These new solutions may be interleaved with the optimized ones for combining exploration-and-exploitation-based learning procedures. However, when reasonably good solutions are found, the improvement obtained with exploratory methods becomes limited. Therefore, a proper exploration-exploitation balance is fundamental for avoiding harming the system's effectiveness by mistakenly introducing exploratory but non-relevant solutions [79].

In this context, the work in [79] presents an on-line learning-to-rank method based on implicit feedback that optimizes the balance between exploration and exploitation strategies for retrieval effectiveness improvement. This learning method works by optimizing a linear feature combination function using two result lists for a given query, one exploitative and one exploratory. These two lists are interleaved (with the first one randomly picked). The effectiveness of each list is assessed according to implicit feedback (click data). The exploratory weight vector is created by randomly moving the exploitative vector. If the exploratory list outperforms the exploitative one, the exploitative weight vector is updated according to a given constant step towards the exploratory vector. Instead of simply interleaving the two retrieved lists, the method probabilistically selects the list from which a retrieved item will be picked for each position of the final list. The effectiveness of the method is directly affected by the proper adjustment of the exploratory probability. Their experimental analysis has shown that achieving the proper balance between exploration and exploitation can significantly improve the retrieval performance of on-line systems. Additionally, experimental results have led to the conclusion that "measuring final performance is not enough when evaluating on-line learning-to-rank algorithms," the different instantiations of the click model [68] "also result in qualitative differences in cumulative performance," and the "performance on some datasets is more strongly affected by noisy feedback." The authors highlight the necessity of conducting new experiments using better click models and even exploring click log data or real-life settings. The authors also suggest that future improvements may be achieved by combining active learning methods with exploration strategies.

For dynamically optimizing the exploration-exploitation trade-off, the work in [184] proposed an extension of [56, 57], which is a query-free approach that starts the search by heuristically sampling the dataset and proceeds by refining results based on user relevance feedback. For estimating the conditional probability of relevance of the images in

relation to feedback events, the authors in [56, 57] used a Bayesian framework. These probabilities are used to select the image to be showed next and are computed according to the proximity to the feedback images. Additionally the images to be presented to the user are selected not directly based on the relevance probability but with a sampling procedure that tries to optimize information gain from feedback by minimizing the redundancy on the result. The redundancy is minimized by iteratively selecting the image with the highest relevance probability that does not belong to the neighborhoods of the already selected ones. This redundancy minimization process, as an exploration-based method, tends to evolve quickly to the relevant regions of the feature space but continues trying to cover all dataset over the iterations even when an image from a relevant region is found. For eliminating such limitations, the work in [184] proposed a dynamic control of the images selected for displaying based on the estimation of consistency among the system internal state and the user search objective. The exploration-exploitation trade-off is optimized by adjusting the images' neighborhood at each iteration using a heuristic consistency score between probability of relevance of the feedback image and the other images shown. If the feedback image's probability is relatively high, it means that the distribution of probabilities is already close to the user interpretation. The neighborhood adjustment score is computed according to the accumulated consistency score over the iterations. Experimental evaluation has shown the statistical superiority of the adaptive method over the baseline for three of the four similarity measures tested.

The authors in [9] present a hybrid approach joining exploration and exploitation using several combinations of a multi-objective genetic algorithm along with the nearest neighbor method. The genetic algorithm naturally explores the feature space by iteratively moving query points according to positive feedback. On the other hand, the nearest neighbor method intrinsically exploits the already found areas of interest of the feature space. For the hybridization process, the results of both methods are probabilistically aggregated based on a dynamic weight selection that reduces the importance of exploration along the feedback iterations. Experimental evaluation has shown that such a combination improves the session effectiveness specially on late iterations.

In a slightly different formulation, for the high-precision and high-recall tasks, combining exploration-exploitation optimization and diversity promotion, the work in [123] proposed a retrieval method for maximizing precision and recall by using a double-loop system that combines an interactive classifier optimization according to relevance feedback and the iterative feature space exploration based on query expansion. This process is recommended for users interested in the completeness of the results and that are willing to make an effort on interactively providing relevance feedback for many items. This process works by exploiting the relevant feature space regions for optimizing the classifier based only on the current pool of retrieved documents. During this process, the user query is constantly updated with the feedback provided. When the classifier is sufficiently stable, an exploration phase is initiated with a new updated query issued to the retrieval engine and the optimized classifier used for selecting the new documents to be shown to user. At this point, with the new explored information, the classifier optimization interactions can continue. The classifier optimization phase can also conduct an active learning process. In summary, this method can be considered a global search system with local search-

based optimization. This framework has been instantiated for five different variations: traditional relevance feedback (Rocchio's method), passive (SVM-based ranked search), unanchored passive (new queries constructed from scratch), active (SVM-based active learning), and diverse active (relevant low-ranked documents are selected to expand the search space). The experimental evaluation has shown that all the proposed instantiations of the framework outperformed the traditional iterative relevance feedback method. Among the framework variations, the active and diverse active instances were the best performing ones, highlighting the potential of exploring the feature space. It is important to mention that the experimental results have shown that the proposed method suffers from the cold-start problem of supervised learning and its success is directly affected by the user effort on labeling documents. The best performing instances of the framework only outperformed the baseline after 50 to 90 judged documents. Moreover, the benefits of the diversity-based method emerged only when around 150 judgments were collected.

Learn-to-rank methods may demand sufficient training information before producing result superior to traditional methods. In fact, our diversity-oriented approach was also partially harmed by the reduced amount of information in the initial retrieval and had also to deal with the natural drawbacks of the diversification. An experimental analysis of such problem and retrieval alternatives are discussed in Chapter 4. For clarity purposes, diversification literature and interactive diversity-oriented works are discussed in Section 2.4.6.

2.4.2 Unlabeled Data

One of the main problems that data classifiers have to face is the limited amount of labeled training samples. Moreover, the feedback information obtained from top-ranked documents is usually biased for the lack of representativeness of the actual relevant items or feature distribution in the dataset, and also the limited information gain when only near-duplicate items are judged. Additionally, constructing labeled training sets is always an expensive task and sometimes error prone. Even when considering object annotation or tagging the systems are subject to inconsistency, for instance because of the use of different dictionaries or as a consequence of different user interpretations of the same object. In IIR, as the amount of unlabeled data is significantly superior to the labeled set and users are not supposed to provide many labels, using unlabeled information is considered as an important boosting factor for learning strategies. Furthermore, at the beginning of a search session, the query pattern information provided by the user is usually extremely limited, which may be improved by integrating unlabeled data to the initial training pattern.

In this field, Xing et al. (2011) [230] discussed about the biased feedback problem that arises when the feedback is not representative of the existing relevant items in the collection. They have experimentally evaluated the bias and reported its greater influence on relevance feedback in the cases of low similarity between query documents and the documents in the collections and also when the documents on the feedback set are too similar. For tackling these issues, the authors proposed extending the feedback set by heuristically selecting unlabeled documents. The best results were achieved when the

unlabeled documents were selected according to a combined score of similarity of positively labeled documents, negative labeled documents, and the portion of new words in relation to the positively labeled documents. The information gain obtained from the novel unlabeled documents was important for improving the amount of relevant items retrieved after feedback and this heuristic outperformed density- based and centroid-based methods.

The authors in [243] argued that traditional SVM-based approaches treat positive and negative feedback samples equally, which is considered not appropriate since these two sample groups have distinct properties. For instance, the positive samples tend to share similar concepts with the query whereas the negative samples may represent several non-related concepts. Another discussed issue related not only to SVM-based RF method, but also generally present on image-based RF schemes is the small size of the group of samples available. In order to reduce such problems, the work in [243] proposed a method based on subspace learning for approximating the relevant samples while separating irrelevant ones using a maximal margin analysis. This method uses a graph-embedding approach for the reduction of the feature space dimensionality. Therefore, positive feedback, negative feedback, and also unlabeled samples are projected into the new learned subspace. The unlabeled information is explored by introducing a Laplacian regularizer and a trade-off for the contribution of labeled and unlabeled samples for the SVM. The experimental analysis reported the superior effectiveness of the methods in relation to other dimensionality reductions method and traditional SVM approach.

Pedronette et al. (2014) [151], proposed exploiting contextual information (feature space neighborhood) for semi-supervised learning for image retrieval with relevance feedback. The proposed method uses the pairwise recommendation reranking algorithm [153] for exploiting unlabeled data in conjunction to pairwise supervised recommendations using feedback samples. In the proposed method, the contextual information is used for adjusting the distances between images that simultaneously occur on the neighborhood of a sample in order to approximate relevant images considering positive feedback while also increasing the distance for irrelevant samples. The experimental analysis has shown the effectiveness of the methods for different content-based image retrieval tasks using shape, color, and texture visual features. Additionally, the proposed method was also evaluated in a multimodal setting combing visual and textual information. The proposed method outperformed a similarity combination function optimization baseline.

2.4.3 Noisy Feedback Reduction

Although user feedback has been shown to effectively improve retrieval effectiveness, search systems have to deal with the problem of noisy feedback that arises when the relevance assessments are not conducted accurately or even erroneously. It is not rare that a user provides confusing or incorrect feedback samples, which directly impact the convergence of learning models.

Considering real user conditions and the possibility of mislabeled feedback samples, the work in [87] proposed a two-step feedback noisy-smoothing method for avoiding harming the learning models with erroneous training data. The authors argued that positively

labeled irrelevant images may decrease the precision of relevance feedback given images similar to those negative examples are likely to be ranked higher after feedback. Additionally, negatively labeled relevant images may harm the recall of relevant items because similar images will be ranked lower. For tackling such issues, the first step of the method uses the similarity of the positive samples in relation to the other positive samples and also to the negative samples to estimate a confidence degree of relevance in order to filter out non-relevant samples mistakenly marked as relevant by the user.

Similar to [243], the authors of [87] also argue about the traditional SVM limitation on treating positive and negative samples equally and also make no distinction according to the relevance probabilities of the samples. In order to properly handle positive and negative samples and exploring different relevance probabilities, the authors proposed a second step to further optimize the learning step with the remaining images. In this second phase, each training sample is labeled with a relevance probability based on their proximity to the other relevant and irrelevant samples. These new relevance probabilities are used to train a fuzzy SVM that properly explores the different relevance confidence degrees for finding the decision boundaries. The experimental evaluation on a medical image collection demonstrated the superior effectiveness of the two-step noise reduction method considering several baselines including the traditional SVM and a relevance score combination method.

As described in Section 2.2.3, Rashedi et al. (2015) [163] also achieved noisy feedback reduction by jointly fusing short-term and long-term learning models.

2.4.4 Feature Learning

An effective approach for CBIR is the construction of dictionaries of visual features [103]. In this context, the dictionary is usually built in a batch learning procedure. When a large training set is to be considered, it requires a costly off-line procedure, which produces global dictionaries based on features extracted from training samples. However, in interactive retrieval scenarios, the training information is produced in an online incremental fashion.

Some recent works have proposed the interactive dictionaries construction according to user preferences [132]. In a further step, Gosselin et al. (2011) [64] also introduced an active learning step for incremental kernel learning and dynamic dictionary construction using the features extracted from relevant feedback samples. This dynamic model outperformed the traditional batch constructed visual dictionary on an image retrieval task. Extending the online dictionary learning idea, Gosselin (2012) [63] proposed a multiple kernel learning method with linear combination of base kernels for specific visual features. These dynamic online methods are specially interesting for image retrieval from dynamic databases in which new items are frequently introduced or even removed, which requires adaptivity skills for retrieval strategies and feature representation methods.

Similar to [64], Wang et al. (2014) [219] also explored an active learning RF method for interactive feature reconstruction. Instead of dynamically constructing new visual dictionaries, the proposed method considers the features of positive feedback samples as input for a covariance matrix based kernel empirical orthogonal complement component

analysis (OCCA [188], which is analogous to the principal component analysis). In this method, the features of positive samples are mapped to a high-dimensional space and their covariance matrix is calculated. Afterwards, the kernel empirical orthogonal complement components of the covariance matrix are computed and the image features are mapped to a new subspace for re-training an SVM-based classifier.

2.4.5 Multimodality

Due to the limitations of single modality approaches, combining multiple feature types has attracted great attention of the research community. Integrating multiple sources of relevance evidence has been proven to enhance retrieval effectiveness by wisely exploiting the complementary aspect or reinforcement criteria of different modalities. In the multimedia retrieval context, multiple modalities are naturally available, for instance considering the visual, audio, and text information within a video. Beyond it, interactively adjusting feature combinations was also considered an effective solution for attenuating the semantic gap.

For instance, Axenopoulos et al. (2012) [13] enhanced a multimodal object retrieval system by incorporating relevance feedback. For fast and effective retrieval, the information from all objects' modalities were mapped to a low-dimensional multimodal feature space. Therefore, multimodal items composed of 3D objects, 2D images, and audio data were described according to the individual modalities and indexed using the unified multimodal feature. Additionally, mapping query items that include at least one type of modality to the multimodal feature space allows the retrieval of the multimodal objects.

In [67], Guldogan et al. (2014) proposed using implicit relevance feedback for personalizing an adaptive image retrieval method based on different modalities, which were named by the authors as multi-form image representation. Therefore, the weighs of the different forms, and consequently their contribution to the result in each retrieval iteration, were dynamically adjusted according to the user behavior while also using a query point movement strategy.

For RF-based multimodal similarity function optimization and per-session system adaptiveness, Calumby et al. (2014) [27] proposed a multimodal image retrieval framework that combined visual and textual similarities into multimodal ranking functions according to users' feedback. For automatic optimization and nonlinear combination of several visual and textual similarities, a genetic programming framework was proposed. Therefore the user preferences were mapped to dynamically discovered ranking functions, which automatically represented the selection and importance of each modality according to the user feedback.

Aligned with the successful integration of multiple modalities for image retrieval in the last years, the experimental evaluation described in this thesis was conducted using several kinds of modalities, e.g., image visual features, metadata, textual descriptions, geographic information, etc. We have evaluated our proposals in different retrieval scenarios based on single modalities and also compared their effectiveness in relation to multimodal sessions.

2.4.6 Diversity

Promoting diversity in retrieval results has emerged as an effective way of maximizing the satisfaction rate in several different scenarios [114, 205]. For instance, it has been applied for tackling ambiguous or underspecified queries for which there is no specific answer item or search aspect [208]. By covering as many query interpretations as possible, a retrieval system may not provide several relevant answer items for a given aspect but at least some relevant samples for each possible user interpretation.

The diversity problem, is an instance of the maximum coverage problem, which has NP-hard computational complexity. Therefore, it is necessary to apply polynomial-time approximations in order make it possible, e.g., using greedy algorithms. For a detailed complexity analysis and demonstrations we refer the reader to Santos et al. (2015) [173].

In general, the diversity promotion methods are divided into two categories: explicit diversity or implicit diversity. These categories are described next. Moreover, we describe several methods which integrate diversity and multimodal or interactive approaches. We also present some applications that directly benefit from diversity promotion.

Explicit Diversity Promotion

State-of-the-art diversity-promotion methods explore different query intents that are explicitly stated or detected at runtime. These techniques are usually employed over traditional information retrieval engines including probabilistic and language modeling methods. In this context, the most effective methods include the IA-Select [2] and xQuAD [170].

IA-Select is a probabilistic method that assumes an existing taxonomy of information, which is used to model user interpretations of a query and categorize the documents in the collection. Both queries and documents may belong to multiple categories. Moreover, the method also assumes that usage statistics have been collected for representing the probability distribution of intents. Therefore, the IA-Select algorithm focuses on the minimization of a diversification objective function and provides a greedy procedure for its approximation. It works by iteratively selecting the document that maximizes a utility function, which considers the likelihood of a document to satisfy the user intent given the query and also considers the conditional probability that a query belongs to a given category assuming that all previously selected documents fail to satisfy the user.

Differently, instead of relying on a predefined taxonomy, the xQuAD framework represents the multiple search aspects as sub-queries, which reflect, for instance, the interpretation multiplicity of an ambiguous query. Therefore, the framework iteratively reranks an initial search result using a probability mixture model that balances the likelihood of a given document to be observed given the initial query (relevance) and the likelihood of observing this document but not the previously selected documents (diversity). For estimating the diversity contribution of a document, the possibly several sub-queries are explicitly considered. Therefore, the diversity score is marginalized by considering the relative importance of each sub-query in relation to the others, which is combined with the likelihood of the document answering the given sub-query. The authors also highlight that several alternatives may be used for estimating the probabilistic components of the framework, i.e., document relevance may be estimated with a probabilistic retrieval

method (e.g., language modeling or BM25) and sub-query importance may estimated using its frequency on query logs or with the number of hits for that sub-query according to a given search engine.

Vargas et al. (2012) [205] reviewed the formal foundations of the two previously described state-of-the-art methods, IA-Select and xQuAD, which outperformed previous implicit diversity promotion alternatives. The authors propose a new definition for these algorithms using a formal relevance-based model that revealed xQuAD as an IA-Select generalization. The relevance-based version of xQuAD (RxQuAD) was evaluated for search and recommendation tasks, and achieved equivalent or superior improvements according to several intent-oriented measures in three different datasets. Since explicit diversification methods are not the main focus of our study, we leave deeper descriptions to the original works and direct the interested reader to [173, 205].

Implicit Diversity Promotion

Considering multimedia data and content-based methods, it is very common to rely on a different paradigm, named implicit diversity. These methods usually exploit the similarity between objects in order to assess their conceptual/aspect closeness and consequently measure their ranking/diversity scores.

The work in [45] proposed a method for the simultaneous optimization of relevance and diversity for image retrieval. Different from other studies, which use two-step diversification methods, the authors in [45] propose an implicit diversification technique inspired on dynamic programming, avoiding the reranking step. Experiments have shown that the method achieves diversity superior to the state-of-the-art with slightly inferior precision values.

Similarly to [45], the work in [204] proposes an implicit diversification method for text-based image retrieval on the Web based on image metadata (e.g., title, description, and tags). In that work, diversity is considered in a pseudo-relevance feedback scheme that exploits query expansion according to different query senses. Experiments have shown that term expansion is very useful and that a retrieval method using only tags was able to produce the best balance between the primary and secondary topics of the queries.

For implicit diversification, several works have relied on the Maximal Marginal Relevance (MMR) method [31], which is actually the ground idea of the intent-oriented methods previously described. In MMR, a relevance-based list of candidate items is diversified to maximize a score that combines relevance and diversity. The resulting diverse list contains the objects that maximize the score according to Equation 2.1, in which: $R = IR(C, Q, t)$ is the ranked list retrieved by the IR system given C (a document collection), Q a query or user profile, and t (a relevance threshold for selecting the documents retrieved or the number of documents); S is the subset of documents in R already selected; $R \backslash S$ is the set of documents not selected yet; Δ_1 is a similarity metric between a document and the query; Δ_2 maybe equal to Δ_1 or another similarity metric; and λ is the relevance-diversity weighing factor.

$$MMR = \underset{d_i \in R \backslash S}{\arg\max} \left[\lambda(\Delta_1(d_i, Q)) - (1 - \lambda)(\max_{d_j \in S} \Delta_2(d_i, d_j)) \right] \tag{2.1}$$

Algorithmically, as a greedy solution for maximizing the MMR score, the most relevant item in the candidate set is initially inserted into a new list. Iteratively, the next item to be added to the diverse list is the most relevant one in the candidate set that is also the most diverse in relation to the images already in the reranked list. This process may continue until all candidate items are reranked or a given reranking depth is achieved.

Another common approach for implicit diversification relies on the use of clustering techniques for producing diverse results [77]. Similar to the reranking algorithms the clustering technique may be applied to previously selected items, for instance using a traditional similarity-based ranking engine. For results construction, representative items can be selected from the different clusters.

Diversity Promotion, Multimodality, and Interactive Learning

The work in [162] proposed a set of interactive approaches for text retrieval with diversity promotion using implicit feedback. Here an MMR-like diversification is applied using the relevance model learned from feedback. Other studies like [20, 171] also present efforts for effectiveness maximization based on the detection of the diversification need for specific queries. In [79], the authors introduce a learning-to-rank optimization method using implicit feedback from click data based on a probabilistic combination of exploration and exploitation-based retrieval. Likewise, in [106] an exploration-exploitation trade-off optimization method is presented for relevance feedback sessions focused on Web search using small screen devices. This method was considered to be useful for the learning process specially for hard queries. Experiments on per-query trade-off optimization have shown that the method achieved better results than traditional utility maximization relevance feedback approaches.

For effectiveness improvement several works have proposed selecting and fusing different features and modalities [33, 102, 114, 218]. For instance, the work in [33] proposed a video indexing and retrieval method by fusing visual and audio signals. Similarly, in [48], audio, visual, and textual evidences are explored for movie summarization. In [114], the authors have proposed not only fusing multiple modalities (visual and textual) from videos, but also refining retrieval results according to different types of queries: mono-intent or multi-intent queries. For mono-intent queries the results were reranked for presenting in the first positions the videos most likely to satisfy the detected prevailing intent. Differently, for multi-intent queries the initial result list was reranked for including diverse videos corresponding to the different intents.

Diverse information has also been shown useful on classifier training as an alternative for maximizing the data distribution coverage and consequently the classifier robustness and convergence rate [123]. To some extent, the diversity promotion methods also relate to exploration-exploitation approaches (Section 2.4.1). By properly optimizing the relevance-diversity trade-off, the system provides the user with very relevant (exploitation) and diverse (exploration) items, specially when iterative methods are used, e.g., the MMR. Therefore, instead using the learned models for presenting the user only items which are very close to the query pattern in the feature space (exploitation), the diversi-

fication procedure may allow expanding the search boundaries or at least improving the representativity distribution of the items from different areas in the search space.

Nevertheless, while it is a very active research field only a few works have investigated the relationship between diversity and user preferences. For instance, Brandt et al. (2011) [23] presented a dynamic ranked retrieval strategy that uses a skip/expand dynamic result tree and a utility gain optimization strategy for maximizing recall and diversity effectiveness. Differently from greedy static raking methods, which iteratively append the document that provides the best utility gain, the first algorithm dynamically selects the items of each level considering their marginal utility and the user navigation feedback. The second algorithm selects the new document not only by trying to maximize the utility gain of the newly expanded level, but also by maximizing its subtrees' utility using a look-ahead estimate (based on static ranking). The experimental evaluation has shown significant effectiveness improvement of the dynamic methods in relation to the static ranking.

Raman et al. (2012) [162] proposed a set of interactive approaches for text retrieval with diversity promotion using implicit feedback. A diversification method based on the Maximal Marginal Relevance (MMR) [31] (described in Chapter 3) is applied using the relevance model learned from feedback. Similarly, in [27] we introduced a new genetic programming framework for improving relevance feedback session effectiveness on multimodal image retrieval scenarios with diversity (see Chapter 4). For improving the learning models, the relevance feedback was taken over diversified results. Genetic programming was applied for the discovery of adapted nonlinear similarity combination functions. The functions were optimized after each feedback iteration and then used for ranking the residual collection. We have shown that learning with diversity can improve session effectiveness not only in terms of diversity, but also in terms of the amount of relevant images retrieved. Experimental analysis has shown that the user feedback over the diversified results allowed retrieving more relevant items and also in earlier iterations.

Diversity Applications

The relevance-diversity trade-off is an important problem associated with several search scenarios. Promoting diversity in retrieval results has been shown to positively impact the user search experience specially for ambiguous, underspecified, and visual summarization queries. The *Retrieving Diverse Social Images Task* [92, 94] combines such problems into a challenge on visual summarization for social photo retrieval in a tourism related context. In this context, Figure 2.2, presents a toy example of redundant and diverse results with the images available in this collection for the "Arc de Triomphe" location.

In the ImageCLEF Photo Retrieval Task 2008 [190], the authors in [58] proposed a multimodal approach with relevance feedback for improving boolean queries constructed based on narrative texts describing what kind of images were relevant, or not, for each topic. For subtopic coverage improvement, the authors used textual and visual clustering and MMR-like diversification approaches on the results of optimized boolean queries. Similar to other studies, the authors have reported precision losses when diversification was applied.

(a) Visually redundant result.

(b) Visually diverse result.

Figure 2.2: Toy example of redundant and diverse result sets for the "Arc de Triomphe" location from the collection used in the Retrieving Diverse Social Images Task.

The work in [158] presented an application of diversity concepts for the generation of visual summaries for geographic regions by using a geographic gazetteer. By incorporating relevant entities from the gazetteer, it was possible to improve the visual diversity and user satisfaction when applied to two commercial image search engines. For multimedia social event summarization in microblogs, the work in [21] explored cross-media relationships (textual and visual) with a diversity criterion for identifying representative exemplars.

In a different context, Halvey et al. (2009) [70] evaluated the impact of diversification methods for video retrieval on user satisfaction, including random, clustering-based, and MMR-like [31] methods. Interactive retrieval experiments with real users have demonstrated a better satisfaction with diversified results, which were considered more appropriate and complete.

Diversification Effectiveness Assessment

There are several measures for evaluating the effectiveness of diversity promotion methods [173]. Considering the applications previously described, a widely adopted measure is the Cluster Recall (CR) [238]. The CR measure is defined as the number of subtopics covered in the results and may be computed for different cutoff points (CR@N). It is presented in Equation 2.2, where d_i is the i^{th} document, $subtopics(d_i)$ is the number of sub-topics to which d_i is considered relevant, and n_s is the total number of sub-topics for a given topic.

Additionally, for assessing integrated effectiveness for relevance and diversity, a common approach relies on the F1 measure. It integrates two separate measures by computing their harmonic mean and may also be taken for different cutoff points. For instance, Equation 2.3 presents an instance of F1 for combining Precison@N (P@N) and CR@N. The P@N is the traditional precision measure, which corresponds to the percentage of relevant images in the top-N positions.

$$CR@N = \frac{|\cup_{i=1}^{N} subtopics(d_i)|}{n_s} \quad (2.2) \qquad F1@N = \frac{2 \times P@N \times CR@N}{P@N + CR@N} \quad (2.3)$$

2.5 User Aspects

Although the user relevance assessment behavior is one important issue in IR, most existing work relies on ideal user modeling, which is also evident when considering interactive experiments since interaction modeling requires more complex user behavior representation [19]. Differently from the document relevance theories, e.g., Probability Ranking Principle (PRP) [164], user-system interaction theoretic models for describing, predicting, and explaining search behavior are still an open issue [15]. Nevertheless, an interesting extension of the PRP theory for interactive retrieval was presented by Fuhr (2008) [61]. Moreover, some works [79, 87, 98, 221] have already studied the impact of different user behavior on systems' performance.

It is also important to consider the assessment cost when real user experiments are conducted. Although some works have conducted live experiments with real users, it is still an expensive process, audience bias-prone, and also hardly reproducible. Therefore, proper user modeling and simulation play an important role on the IIR field and some works have highlighted that assuming some constraints, well-founded and strictly-defined user patterns can be successfully applied on systems evaluation and optimization with correlation to real user experiments [15].

User assessment behavior has to be carefully considered on experimental design for labeling the relevance of the data, not only as part of the online retrieval process, but also when used for the creation of relevance ground-truth for test collections [71]. As described in [15, 19], interactive search sessions require the user to make effort on several tasks such as query (re)formulation, result scanning, clicks and/or relevance assessments, document/image inspecting, stopping decisions, etc. All these actions contribute to the search cost and consequently impact the user experience.

In an interactive search context, Baskaya et al. (2013) [19] simulated different user behaviors in relation to search goals and constraints, query formulation strategies, snippet scanning, stopping strategies, and user response in relation to documents' relevance. Additionally, ideal and fallible human behavior were simulated (considering scanning and correct assessment probability) and contrasted considering session effectiveness. By probabilistically modeling user interaction patterns in a keyword-based interactive information retrieval task, it was noticed that the human behavior on multi-query sessions may lead to improved effectiveness when compared to a similar single query session. Moreover, experiments have shown the nonexistence of a general search behavior that leads to optimum or superior effectiveness, which is actually deeply related to the information need (topic) and target collection as similarly stated in [15].

In [15] (extended from [16]), Azzopardi proposed a more realistic theoretical modeling for search behavior understanding and prediction based on the search economic theory [14]. As proposed in [14] and evaluated in [16], the cost function for interactions considers the number of queries in a session and the amount of documents assessed per query along with their respective costs. In [15], the cost function is updated to incorporate the number of result pages viewed, the number of snippets inspected per query, the probability of document assessment, and their respective costs. An empirical analysis compared real-user behavior extracted from search logs with the proposed theory consid-

ering the relationship between interaction patterns, cost, and performance. The results have shown an alignment between the predicted and the observed behavior from real users. Nevertheless, although more realistic than previous proposals, this new model still demands further improvements specially for considering some kinds of approximations and still limited user constraints and bias.

While several works have been conducted on modeling and considering user aspects on retrieval simulation and assessment, there is still a lack of studies on the judgment process and labeling effort of individuals on image retrieval tasks.

Similar to their previous work in [209], on the relevance assessment effort evaluation for text retrieval, Halvey & Villa (2014) [71] conducted user experiments to investigate judgment effort and accuracy impact for image retrieval considering the topic difficulty, visual-semantic topic characteristics, and image size. In summary, the experiments have shown that the size of the images had no impact on the judgment effort, but larger images took more time for relevance assessment. Moreover, the judgment accuracy decreased, while the time to provide a judgment and the user perceived effort increased when topic difficulty increased or when topics moved from visual to semantic. Finally, judgment time and the user perceived effort also increased with the difficulty increase.

These findings suggest for instance that retrieval systems could be dynamically adjusted in relation to the number and the size of the images to be presented, considering the underlining difficulty of semantic characteristics of the user query. In a different direction, the outcomes from [71, 209] could have positive impact on user behavior modeling, such as in [15, 16], by simulating and assessing different user patterns considering different topic difficulties and semantics, which should also be incorporated into effectiveness evaluation measures.

2.6 Effectiveness Evaluation and Benchmarks

Different from traditional IR, IIR evaluation also includes user-oriented methods for the assessment of search systems and their components and tries to understand user actions from cognitive and behavioral perspectives [69].

Kelly et al. (2009) [109] discussed about the major challenges for interactive systems evaluation, such as: (i) there are poor or inadequate user and task models; (ii) real search task involves dynamic corpora with different document type and constant quality variation; (iii) real search tasks are complex and include evolving objectives not captured by traditional measures; and (iv) an interactive search task may be conducted with different query sessions. All these challenges bring important experimental difficulties and demand specific and combined studies.

In the historical overview (1967 to 2006) of [110], the authors concluded that large portions of IR and IIR research are evaluations in the form of experimentation or quasi-experimentation. As observed in history, recent works, and meetings, for the IR technology understanding and evolution, researches have not only developed new techniques, but also properly evaluated their performance. Moreover, as experimentation is the most popular and accepted method in IR and IIR and, despite the focus on users and interaction on

early discussions of IR evaluation, the research efforts took different paths focusing on IR component evaluation (system-centered) and interaction evaluation (user-centered). Nevertheless, despite great advances, IIR is still considered a recent field with no prescribed experimental methods. Therefore, and reasonably, it relies on a broad menu of evaluation protocols and measures. It may be a consequence of the complexity of evaluating the user behavior and interactive interfaces simultaneously. This wide variety of evaluation tools was evidenced in the systematic review in [110] and also in the recent works discussed here, specially considering evaluation measures and statistical analysis methods.

Considering IIR research, Thomee & Lew (2012) [194] suggested that for evolving the benchmarking and evaluation materials, the community has been working on constructing large and freely distributed databases, as well as proposing new evaluation measures, which are expected to be more adequate for the evaluation of interactive systems. Additionally, it is reported a great effort on conducting proper use simulation. Nevertheless, although such issues have been addressed in several works, there still remains a great room for improvement towards building better evaluation resources and protocols.

In spite of a great effort on formulating theoretical [14, 69, 110] and practical [69, 98, 104] foundations for interactive information retrieval, there is still no well-established understanding, modeling, and evaluation standards. Therefore, IIR evaluation is still conducted with non-standard collections, target subjects, and diverse sets of measures for supporting multiple task variations and research objectives, which makes it very difficult to extend and compare different studies [69].

2.6.1 Evaluation Protocols

In general, IIR evaluation studies aim at mimicking real-world scenarios, which require the modeling and simulation of several interactive patterns and capture and analyze multiple response signals. From datasets to user behavior and statistical data analysis, there is a vast amount of choices and their proper usage depends on the study objectives and available resources. Therefore a common IIR evaluation work includes the definition of several parameters such as: result evaluation protocol, training and testing datasets (Section 2.6.2), effectiveness measures (Sections 2.6.3, 2.6.4, and 2.6.5), search type (target, category, achievement-based, etc.), number of queries/topics, number of items retrieved per iteration, number of feedback samples, relevance assessment grades, user approach (real or modeled), among others. In active learning studies, it is also important to establish the number of learning iterations before the user has access to the final results and how the samples are selected for user assessment.

User modeling

Considering the user complexity presented in Section 2.5, we observed that IIR works are still conducted with non-standard modeling but some groups of approaches can be highlighted: Perfect user simulation (with classes/categories information or relevance assessments [9, 27, 220]; Probabilistic modeling [19, 87, 244]; Click model [79]; Log analysis [15, 163, 225]; and Real users [71, 98, 125, 184, 221].

Search Task

Even when considering text, image, or video search works individually, recent works have evaluated several search/task formulations for interactive retrieval, e.g., ad hoc search [15], target search [112], conceptual search [46, 243], category search [67, 136, 219], and the not so common, here named achievement-based search. As an example of the latter, a search session continues until at least a given number of relevant items are found in the same iteration [184].

Interactive Result Processing

The handling or aggregation of the results obtained throughout an interactive session plays an important role on the effectiveness evaluation and the mapping of retrieval results and items' relevance into a measure of success. Surprisingly, it is very often not explicitly described in the literature, which introduces analysis weakness and harms reproducibility.

Quite frequently, the experiments are conducted using the rank-shift [99] procedure in which the relevant items previously found are shifted to the top of the ranking in future iterations biasing and artificially increasing effectiveness values. This bias is known as "ranking effect" [167]. Alternatively, with the collection reranking procedure, all items in the target dataset are reranked in future iterations. In turn, with a residual collection strategy [73], only the items not previously seen are presented in further iterations, no matter if they were judged relevant or not. Differently, the freezing approach [111] keeps the relevant items in the same rank positions they were firstly retrieved. As a variation, the full freezing protocol [224] holds every item in the same position they were retrieved, and consequently a final ranking can be constructed by appending the results from each iteration.

Williamson (1968) [224] describes the feedback process as being either fluid or frozen. Fluid feedback is suggested when the user has to judge the relevance of items by analyzing only item surrogates and thus the item itself is only examined after the search is finished. In this approach, the entire collection is re-ordered according to the modified query. Differently, in a frozen approach, items (content) are examined by the user during the search so the original order is not changed for the next iterations. The freezing approach seems to be more suitable for environments in which the user is able to inspect the items while providing relevance feedback. The authors also present a different approach named "re-ranked original order." In this approach, the collection is just reranked by moving judged relevant items to the top of the ranking while moving judged non-relevant (or already seen non-judged) items to the end of the list. This approach suits the case when user just examines surrogates but no feedback is used by the systems for collection re-ordering. The focus of this approach is the impact of the effort of user feedback without any explicit result refinement by the system.

Another evaluation protocol makes use of a second collection (feedback collection) for query reformulation and the reformulated query is run over a different (target) collection [73]. This approach in turn uses a "training" collection that is different from the target one and may not have representative relevant items as the target collection. Therefore, both residual collection and feedback collection techniques may be fair approaches

for systems/techniques comparison but are not always practical in real environments.

Each of these approaches may be appropriate for different retrieval tasks and consequently refer to a different effectiveness evaluation protocol. At the same time, each technique brings some experimental drawback that should be carefully considered. For instance, since judged relevant items tend to be or are explicitly placed at the top of the ranking the usage of fluid or rank-shifting approaches may mask the improvement of the rank position of unseen relevant items.

These protocols allow capturing different user interaction effort and system effectiveness signals. It is worth mentioning that while there is no established guideline, the impact of the different protocols may lead to completely different understanding of the user interaction outcomes and system behavior. These protocols were found in recent literature, such as: rank-shift [123, 166], collection reranking [228, 244], residual collection [8, 161, 243], and full freezing [27].

2.6.2 Datasets

A summary and brief description of the datasets used in recent interactive retrieval works are presented in Table 2.2. It is important to notice that some collections were used in multiple works described here, whereas several works have explored only subsets of their content. Moreover, several works conducted experiments on customized or manually constructed collections, which are not necessarily available for future work.

As observed from Table 2.2, even when considering text-only or image-only evaluations, recent works have relied on a wide variety of test collections, which were actually not constructed for interactive experiments and sometimes do not provide all the required simulation resources. As traditionally used in IR experiments, most of these interactive retrieval works rely on category information and relevance assessments for user modeling and simulation.

In [237], the authors discussed about the drawbacks of traditional image collections considering several user-related characteristics. As described, such collections do not represent the vagueness of user queries. They are constructed based on documents (images), and do not properly represent personal photo collections. For most traditional collections, relevance assessments are only binary, which is considered not adequate as they do not provide a definite judgment but just an estimated probability of relevance, specially when obtained via relevance models for multimedia information retrieval.

For allowing better user-centered evaluation, Zellhofer (2012) [237] proposed a new collection, built with image samples from real photographers with focus on representing real off-line user collections, which include duplicates, variance in quality, and noise. This new collection, named Phytia Image Collection v1 (PICv1), was constructed for allowing more adequate user-centered evaluation as an alternative or complement to traditionally used collections such as Caltech 101 [54] and 256 [65], MIR Flickr [90], MSRA-MM [216], and Social Event Detection Task [150] (extended in [149]). None of these collections has all the characteristics of PICv1, which are: real user data (without image preprocessing steps except for scaling and anonymization), real user queries (event-based search), real user assessments (including graded levels), extensibility for new users and features.

Table 2.2: Datasets explored in recent IIR works. *(Main) data type. (**)Number of classes/concepts/tags.

Type*	**Dataset**	**Size** (**)
Text	Letor 3.0[160] and 4.0[159]	Multiple datasets
Text	TREC 1 [74] and 2 [75] Ad hoc tracks	742,611 docs
Text	TREC 3 Ad hoc track [76]	741,856 docs
Text	TREC 6 Ad hoc track [214]	556,077 docs
Text	TREC 7 [211] and 8 [212] Ad hoc tracks	528,155 docs
Text	TREC 9 [213] and 10 [215] Ad hoc tracks	Multiple datasets
Text	TREC 9 Query Track [26]	510,000 docs
Text	TREC Filtering Track 2002 [165]	800,000 docs
Text	TREC HARD Track 2005 [3]	1,033,461 docs
Text	TREC 6 [141], 7 [142], and 8 [78] Interactive Tracks	210,158 articles
Text	TREC Microblog Track 2012 [182]	16mi tweets
Text	TREC Microblog Track 2013 [128]	243mi tweets
Text	TREC Robust Topics 2005 [210]	1,033,461 docs
Text	ClueWeb09[3]	1.04bi web pages
Image	Aerial orthoimagery [234]	600 (6)
Image	Brodatz [24]	1,776 (111)
Image	Caltech-101 [54]	8,677 (101)
Image	Caltech-256 [65]	30,607 (256)
Image	Coil-100 [138]	7,200 (100)
Image	Corel [137]	circa 80,000 (800)
Image	ImageCLEF Photographic Retrieval Task 2007 [66] and 2008 [10]	20,000
Image	IRMA (Medical Collection)[4]	Multiple datasets
Image	MIRFlickr[89]	25,000 (1386)
Image	MPEG- 7 Part B [120]	1,400 (70)
Image	MSRCORID[5]	4,320 (20)
Image	NUS-WIDE [37]	269,648 (81)
Image	Oxford Flower17 [139]	8,189 (103)
Image	PASCAL VOC 2006 [51]	2,618 (10)
Image	PASCAL VOC 2007 [49]	9,963 (20)
Image	PASCAL VOC 2012 [50]	11,530 (20)
Image	University of Washington[6]	1,109 (20)
Video	MediaEval Video Genre Tagging Task 2012 [174]	15,000 (26)
Video	TRECVID 2005 [146]	169 hours of video
Video	TRECVID 2006 [147]	328 hours of video
Video	TRECVID 2007 [143]	200 hours of video
Video	TRECVID 2008 [145]	253 hours of video
Video	TRECVID 2009 [144]	410 hours of video

[3]http://lemurproject.org/clueweb09/ (As of October 6, 2015).
[4]http://research.microsoft.com/en-us/projects/objectclassrecognition/ (As of October 6, 2015).
[5]http://www.irma-project.org/ (As of October 6, 2015).
[6]http://imagedatabase.cs.washington.edu/ (As of October 6, 2015).

The author suggests two main applications for PICv1: (a) search for sharpen images (including duplicate removal) or visual variations (e.g., using clustering) and (b) event-based retrieval (61 event-based topics). In summary, the PICv1 collection includes:

- 5555 personal photos from 19 photographers;
- Demographic metadata of the photographers and assessors which allows persona creations for user simulation;
- EXIF data, GPS coordinates (automatically or manually included), and city or country names;
- Tags: indoor/outdoor, day/night, altered, blurred etc.;
- Number of people in the photo;
- Event information/ground-truth using WordNet [135];
- 130 fully assessed topics from different domains;
- 32 topics with graded relevance assessments (0 – irrelevant to 3 – fully relevant);
- Ideal DCG curves [97]; and
- 18 low-level visual features.

Unfortunately, since PICv1 collection does not provide any diversity related information it is not suitable for the experimental analysis proposed in this thesis. In our experiments, we rely on the collections from the MediaEval Retrieving Diverse Social Images Task [92, 94] and ImageCLEF Photographic Retrieval Task [10]. These collections provide all the experimental materials for our study on diversity-oriented multimodal and interactive retrieval.

2.6.3 Effectiveness measures

The historical analysis in [110] revealed that even though classic measures were modified in several ways, none of those actually became a standard choice and the system-centered measures were accepted as part of the evaluation paradigm for IIR systems. Moreover, although there was a clear distinction between user-centered and system-centered evaluation approaches, most user-oriented evaluation works examined also carried system-centric evaluation characteristics using research models quite similar to the traditional Cranfield [17] and TREC-like[7] paradigm that only incorporated instruments and measures for handling interactions data and assessing user experience. The most commonly reported measures were grouped and are presented in Table 2.3.

Table 2.3: Most commonly reported measures.

Performance measures	recall, precision, accuracy, and variations
Process measures	number of clicks, number of queries, number of documents viewed, and time-based measures
Usability measures	usefulness of the system, user-friendliness, and satisfaction

[7]http://trec.nist.gov (As of October 6, 2015).

In the recent literature, authors have conducted effectiveness evaluation with many different measures. The most common measures reported are the traditional relevance-based, such as: Average Precision, Mean Average Precision, Precision@N, Recall@N, Precision X Recall, and NDCG. Several works have computed these measures in a per-iteration basis, e.g., Recall@N X Iteration. Alternatively, several studies applied not so common measures such as: R-Precision (in [123]), BPREF, and GMAP (in [27]), and the number of relevant items per iteration (in [59]). Some measures were also reported for evaluating results' diversity, such as intent-aware measures (in [23]) and Cluster Recall (in [27]). Moreover, and quite rarely, some studies introduced different success estimation measures such as the cumulative percentage of successful sessions in [184] and session time in [71].

Some measures related to learning-to-rank and session-based retrieval are discussed in Section 2.6.4 and Section 2.6.5, respectively.

2.6.4 Evaluation and Measures for Learning-to-Rank Methods

When machine learning techniques are used for constructing search engines, their optimization processes often rely on finding optimal settings that consequently produce high values in terms of an effectiveness measure. This metric is usually taken for representing the user satisfaction and may have different purposes, reflecting different aspects of the retrieval effectiveness. Moreover, these measures may evaluate the (user-oriented) effectiveness on the top of the ranking (e.g., precision at rank 10) or the (system-oriented) overall ranking quality (e.g., MAP) [235].

Although a common belief, based on the *empirical risk minimization*, suggests optimizing the final evaluation measure using the training set for maximizing the test set effectiveness, the work in [235] has experimentally shown that, under certain circumstances, it is not the case. The authors in [235] proposed considering the informativeness characteristic of a measure for the learning process assessment and that optimizing the search system for a more informative measure can lead to better performance in the actual final evaluation using a less informative measure. The informativeness concept of a measure is related to: (i) the sensitivity to rank quality changes or items flip; and (ii) the importance of different parts of the ranking (e.g., discount functions). The work in [235] has also shown that optimizing a more informative training measure implicitly optimizes the less informative one. It occurs because reaching the local optimum of the former leads to more likely reaching the local optimum of the latter in comparison to training and testing with the same measure.

We can notice that the optimization of IR and IIR systems may be directly affected by the target evaluation measures and therefore developing sensitive, informative, learn-to-rank suitable measures is still an open and promising field.

The evaluation of learning-to-rank methods using implicit feedback (e.g., click data) is becoming a more frequent alternative to traditional evaluation models based on explicit relevance information. This fact is also interesting for implicit feedback, which is a natural product of user-system interaction with little cost and reflecting real user experience [80].

2.6.5 Session-based effectiveness

As stated in [112], real users usually search using short queries and try to improve the search by reformulating and issuing several queries in a session or examining more documents. Such behavior has been shown to compensate for poorly, broadly, or ambiguously defined queries. However, it is quite different from the traditional Cranfield-like evaluation activities that commonly explore longer queries for optimizing a single search. While some works conducted session-based evaluation on the results of the final query [105], these methods did not capture the information of whether the user engaged in the session, e.g., because she received poor or incomplete results or just changed the search aspect after finding some satisfactory results [47, 202]. Therefore, as pointed out in [18], the session-based evaluation demands specialized modeling and evaluation measures.

The effectiveness evaluation procedures with real users and multiple query sessions are difficult to analyze because of the necessity of monitoring different variables, which are strictly dependent on testing settings. Moreover, traditional effectiveness metrics require special evaluation protocols, usually not properly reflecting the user interaction effort. Although real interactive search users usually issue multiple queries, for instance providing relevance feedback or conducting query reformulation, several works in the literature and most IR evaluation measures consider only a unique query for each retrieval session. As one cannot assume a retrieval system provides independent results for each query in a session, the results of each query should not be independently evaluated and aggregated for representing the session effectiveness.

The authors in [98] argued that traditional measures in general provide insufficient information for evaluating searcher's interaction effort and proposed a new effectiveness measure claimed to be more adequate for session-based evaluation, the Session-based DCG ($sDCG$), defined as:

$$sDCG(q) = (1 + log_{bq}^{q})^{-1}) \times DCG, \tag{2.4}$$

where bq is the base for query discount and q is the position of the query. The discount vector $sDCG(q)$ of a query q can be normalized and concatenated to represent the whole session ($nsDCG$).

Extended from the Discounted Cumulated Gain [97], $sDCG$ is a metric for evaluation tasks with multiple query sessions, graded relevance assessments, and adapted to different search stop user criteria. Moreover, $sDCG$, by handling query sequences, allows additional discount of relevant items retrieved after each user interaction effort. As discussed in [98], this new measure is considered more suitable for session-based evaluation for:

- considering items in equivalent rank position more relevant when returned for an earlier query;
- using smooth discount for document-based gains and query sequence effectiveness importance; and
- being configured with parameters directly related to search and session characteristics.

In a usual IIR scenario, the user examines a ranked list of results and at any moment can interact with the system by reformulating the query or even finishing the session. This behavior can be captured by observation or inferred using the last clicked document. However, the evaluation materials for batch experimental simulation of static sessions do not include these reformulation and stopping points. The authors in [104] argued that using an interactive evaluation paradigm can better assess the real user experience but previously proposed measures, e.g., instance recall [142] and $nsDCG$ [98], are not able to properly capture the high degrees of freedom of user interactions and also result in an expensive process for requiring many test subjects. Moreover, since $nsDCG$ does not model the early abandonment of a session and requires a fixed reformulation point, it does not capture different user behavior in response to different retrieval results.

For allowing the evaluation of retrieval systems using static multi-query session, model-free, and model-based measures were proposed in [104]. The model-free family of measures, inspired by the interpolated precision, does not include the user's behavior on the formulation (reformulation points), whereas the model-based family is constructed for a simple user interaction model. The formulations of the two families allowed generalizing traditional evaluation measures for multi-query session evaluation. These formulations are defined over the concept of interaction path. Each path is a set of actions including: (i) moving down on ranking; (ii) reformulating and starting at the top of a new ranking; and (iii) abandoning/ending the search. For instance, a generalized model-free version of the precision measure for multi-query session (sP) is represented in Equation 2.5:

$$sP = \frac{rR@j,k}{k}, \tag{2.5}$$

where $rR@j,k$ is the set of counts of relevant documents for all possible paths of size k that end at reformulation j. The recall measure is similar to Equation 2.5 but dividing $rR@j,k$ per R (the total number of relevant items).

Assuming a simple model in which the user examines a ranked list of documents until some point, it is possible to derive probabilistic (model-based) measures instead of assuming the user will receive optimal results as the model-free measures. Therefore, the work in [104] also formulated the session-based measures according to the expected retrieval effectiveness (Eq. 2.6) and not the maximum values, as used for interpolated measures:

$$esM = \sum_{w \in W} P(w) M_w, \tag{2.6}$$

where $P(w)$ is the probability of a path w and M_w is a measure for the path w. For a detailed description and thorough formulation of the session-based measures the reader is directed to the original work in [104].

For effectiveness prediction, by describing session-based features for queries, the authors in [119] have shown that it was possible to improve query performance prediction. The proposed method combined click-based features with session-based features (the information grouped from all sessions containing a given query q). Among the session-based features, we can highlight the mean reciprocal rank of all first clicks in queries co-occurring

in one session, the number of sessions, average number of queries per session, average distance of the query position to the initial and terminal queries of the session, and time-based statistics. Additionally, the authors have also computed aggregated features for all queries co-occurring in a session with q with at most k queries of distance.

Finally, as some search tasks may be fulfilled with different query sessions, which is named cross-session search, recent works have studied the experimental characteristics, evaluation methods, and user models for this context. A deeper discussion on cross-session search is out of the scope of this thesis and for more information the reader is directed to the works in [109, 115, 202].

Significance Analysis

For strict result analysis and the construction of an adequate comparison between different retrieval systems or even variations of the same systems, it is common to explore statistical analysis methods. The well-know k-fold cross-validation strategy has been successfully applied in the IR literature, for instance, in the recent works in [79, 87, 243]. Additionally, for significance definition, several statistical methods and coefficients have been applied, such as: standard deviation, confidence intervals, student's t-test, Friedman's test, Post hoc Holm's test, Wilcoxon's signed rank test, Levene's test, Kendall's Tau, among others.

As observed in recent work, there is still no well-established choice and the selection of the test to be used is rarely properly augmented. For the interested reader, an experimental comparison of several statistical significance tests for IR evaluation can be found in [180].

2.7 Multimedia Retrieval and Applications

In the works described in this thesis, most of the interactive methods were proposed for document retrieval and visual image retrieval. However, several multimodal and multimedia retrieval experiments have been conducted on other media applications such as audio and video retrieval.

In the image retrieval context, most of the methods focus on general photo collections, such as the Caltech-256 [65], Corel [137], and Pascal VOC [50] datasets. Nevertheless, some interesting works on interactive retrieval have been conducted for medical images [87], remote sensing images [44], soccer teams [151], fish images [59], and flowers [232].

In [194], several interactive retrieval applications have been highlighted such as search over the Internet, 2D and 3D medical repositories (MRI, X-ray, CT scans, ultrasound, and electron microscopy), computer-aided diagnosis, and digital libraries. In [43, 44], interactive strategies with active learning were proposed for remote sensing images retrieval on earth observation data archives.

Wei & Yang (2013) [221] highlighted some important and interdependent factors related to interactive video retrieval: (i) the exploration-exploitation dilemma (see Section 2.4.1); (ii) prior vs posterior knowledge; and (iii) domain adaptation. The exploitation is achieved with the posterior knowledge about data distribution, e.g., with user

feedback, and exploration guides the search out of local optima using the prior knowledge, e.g., according to labeled data distribution. In turn, the domain adaptation is achieved by combining and enhancing prior and posterior knowledge.

In the multimedia context, the work in [221] proposed an integrated framework for video retrieval with relevance feedback based on an active learning model (see Section 2.2.2) using both prior and posterior knowledge. Moreover, the active learning and posterior knowledge is enhanced by selecting semantically constructed data groups whose distribution is similar to the labeled samples.

As we can notice, the work in [221] integrates several research alternatives described in the previous sections for enhancing retrieval effectiveness. Another interesting alternative, intrinsically related to video retrieval is the combination of multiple features and also multiple information modalities (see Section 2.4.5). For instance, Mironica et al. (2013) [136] proposed a RF method with the combination of several visual, audio, and textual features from videos.

2.8 Summary and Considerations

In this chapter, we reviewed many aspects related to interactive learning-to-rank for information retrieval. From theoretic foundations to practical resources, we have described remarkable efforts on leveraging more effective and efficient interactive retrieval systems. We have shown that while the research community achieved important advances in the last decades and specially in the latest years, some important questions still impose great challenges. As an intrinsically multidisciplinary field, IIR has evolved over the years by integrating novel components from several research areas. At the same time, the increasing importance of information access on the day-to-day life and the ever increasing amount and variety of the information generated and stored demanded retrieval engines to adapt towards better answering complex user needs.

As we presented, in the last years, IIR research has been directed to integrate as much information as possible, fusing multiple data sources and analytical methods which allowed targeting customized user experience. Moreover, extracting as much information as possible from user interactions was important to enhance learning strategies that evolved from intra-query approaches, to session-based, collaborative long-term learning, and hybrid methods.

While it is one of the most important factors of the interaction loop, user understanding is still a complex task given the absence of standard frameworks and experimental materials. Moreover, with the wide spectrum of applications and scenarios, standard evaluation protocols are difficult to be established and consequently require further research efforts. Nevertheless, while an important obstacle to the research development, it opens opportunity and imposes the need for the proposal and validation of new evaluation criteria.

In order to best explore advanced learning techniques, researchers have proposed using many different boosting clues, such as unlabeled data and multimodal evidences. Moreover, it has been demonstrated the effectiveness of smart procedures for maximizing

the user-system information transferring with implicit feedback, active learning, diversity promotion, and exploitation-exploration balancing.

By integrating historical advances and novel methods, this review works as an introduction to IIR ground concepts and also presents a deep and broad view of the state-of-the-art. Finally, we hope the compiled challenges and directions may guide and foster new research proposals and the development of more advanced IIR methods.

Chapter 3

Multimodal Diversity Promotion

This chapter presents our proposals and experimental analysis on relevance improvement and diversity promotion using multimodal data, such as: image visual features, metadata, and user credibility information. The remainder of this chapter is organized as follows. Section 3.1 presents our research questions associated with the integration of multimodality and diversity promotion and our experimental objectives and contributions. Section 3.3 presents our first proposal on relevance-diversity trade-off balancing, which is based on multimodal features combination. In turn, Section 3.4 presents our second proposal, which explores a learning approach for multimodal rank aggregation. Finally, Section 3.5 presents a summary and final considerations.

3.1 Research Questions and Proposals

Inspired by the successful approaches described in Chapter 2 for combining multiple sources of information and the promising application of diversity promotion for answering difficult queries, including ambiguous or underspecified information need, our experimental work was guided by the following research questions:

- Does multimodal data integration contribute for optimizing the relevance-diversity trade-off?
- Is it possible to improve diversity with small or even no impact on relevance?

Considering these questions, the relevance-diversity trade-off optimization problem, and the complementary nature of multimodal information, in the following sections we present two proposals (Sections 3.3 and 3.4), which integrate relevance enhancement and diversity promotion steps for relevance-diversity trade-off enhancement. The proposed pipelines are evaluated considering single modality image retrieval and the corresponding extension to multimodal retrieval. The multiple modalities are considered not only for relevance-based reranking, but also for filtering out non-relevant images and also on the diversity promotion step.

Our experimental analysis revealed the effectiveness of combining multiple modalities for the relevance-diversity trade-off balancing. By integrating several sources of information it was possible to significantly improve retrieval diversity, keep equivalent relevance values, and finally, produce superior trade-off balancing.

Figure 3.1: The representative images for the location *Arc de Triomphe* from the MediaEval Retrieving Diverse Social Images Task 2015.

3.2 Experimental Context and Data

The investigation of our research questions and the experimental evaluation of our proposals were conducted in the context of the Retrieving Diverse Social Images Task at MediaEval 2014 [94] and 2015 [92].

In these tasks, the experiments are conducted assuming a retrieval scenario in which a certain tourist (search engine user) aims at receiving a visual overview of a place she is interested in visiting. The experimental image collection for the 2014 edition is divided into development and test sets with images from 30 and 123 locations, respectively. In turn, for the 2014 edition, the collection is divided into development and test sets with images from 153 and 139 locations, respectively.

For each point of interest there are circa 300 images retrieved from Flickr and the objective is to generate a summary set with 50 images that are at the same time relevant and diverse.[1] For each location there is a set of up to six representative images extracted from Wikipedia[2], which may be used as additional input information for queries. For instance, Figure 3.1 presents the representative images for the location *Arc de Triomphe*. Some pictures from the original list for this location were presented in Figure 2.2.

Along with visual and textual features, there are also several credibility descriptors related to user/owner of the images. The credibility scores represent the quality (correctness) of the user's annotation tags and the quality of the tag-image relationships. These scores were computed in relation to all images in the collection of a specific user and therefore work as evidence of the probability of this user to share relevant images.

For evaluating the results, three official measures are considered: Precision@N (P@N), CR@N (Equation 2.2), and F1@N (Equation 2.3). The official measures are computed for the first 20 images, but we present our results for N = 5, 10, 20, 30, 40, 50.

[1]http://www.flickr.com (As of October 6, 2015).
[2]http://www.wikipedia.org (As of October 6, 2015).

3.3 Diversity Promotion - Part 1: Exploring Multimodality

This section presents our first pipeline proposal for multimodal diversity promotion in the context of the *Retrieving Diverse Social Images Task at MediaEval 2014* [94]. For this task, we developed and evaluated a summarization and diversification approach for social photo retrieval. Our approach is based on irrelevant image filtering, image reranking, and diversity promotion by clustering. We have used visual and textual features, including image metadata and user credibility information.

For tackling the task previously described, our approach follows the general pipeline presented in Figure 3.2. At first, two filtering steps are conducted in order to reduce the amount of irrelevant images. Afterwards, reranking steps are applied for improving image rank positions according to two different relevance aspects (visual similarity and user credibility). Finally, clustering is performed and followed by representative and diverse images selection. Specific combinations of the proposed steps were set for each type of ranking (Section 3.3.5).

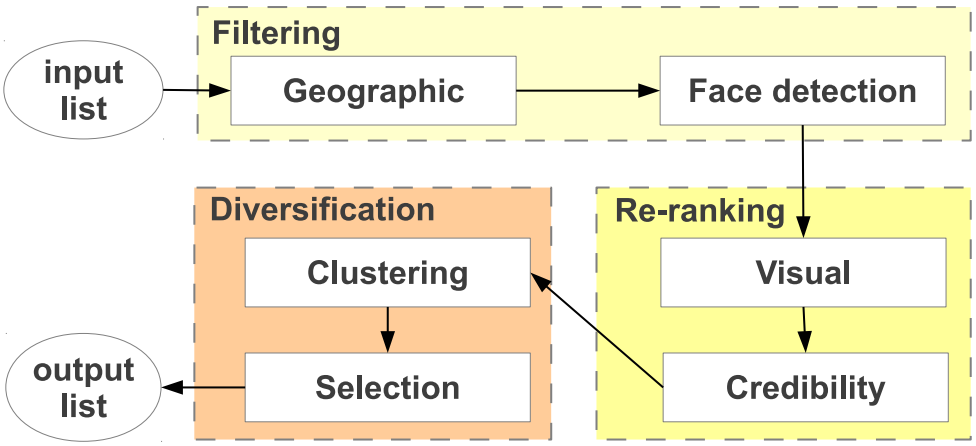

Figure 3.2: Overview of first multimodal proposed approach.

The pipeline proposed includes several steps related to improving the initial non-diversified ranked list received as input. Notice that the diversification methods are prone to promote diverse but non-relevant items, which is exactly what our filtering and reranking procedures are intended to attenuate. Therefore, our hypothesis is that by reducing the amount of non-relevant images and constructing improved ranked lists we may trigger better results after the diversification step, specially by reducing the amount of noisy (non-relevant) information that the diversification module has to process.

3.3.1 Filtering Step

In order to reduce the number of non-relevant images, we applied two filtering strategies: Geographic filtering and Face filtering. Eliminating non-relevant images allows higher effectiveness in terms of final relevance and boosts the diversification procedure. This is a consequence of fewer non-relevant items as candidates for the final diversified list.

The geographic filtering (named GeoFilter) takes the reference lat/long of each location and then eliminates all images located farther than a given range. In this case, only geo-tagged images are assessed.

Since, following the task constraints, images containing a person or crowds in the foreground are considered non-relevant,[3] we used a face detection module of Face++[4] for filtering. For all images, we computed the features: a) number of faces; b) biggest face size; c) smallest face size; d) average face size; e) total face size. The size values were computed as a fraction of the image spatial domain.

Our first face-based filtering approach (named NumFacesFilter) eliminates all images with a number of faces superior to a predefined threshold. The second approach (named FaceClassifierFilter) used a kNN classifier based on the described features and considering all development images as training instances.

3.3.2 Reranking

Since the original lists may present redundant and non-relevant items, their positions may not be optimal for their relevance. Even after the filtering procedures, some non-relevant images may remain and therefore we proposed two reranking strategies: visual-based and user credibility-based.

The visual reranking was conducted using as queries the location's representative images obtained from Wikipedia. Figure 3.3 presents an illustration of the ranking score for a given feature. The original lists were reranked according to the distance in relation to the representative set. The visual distance (d) from each image (img_x) in a list to the corresponding representative set (rep) was computed as the minimum distance value between img_x and each representative image.

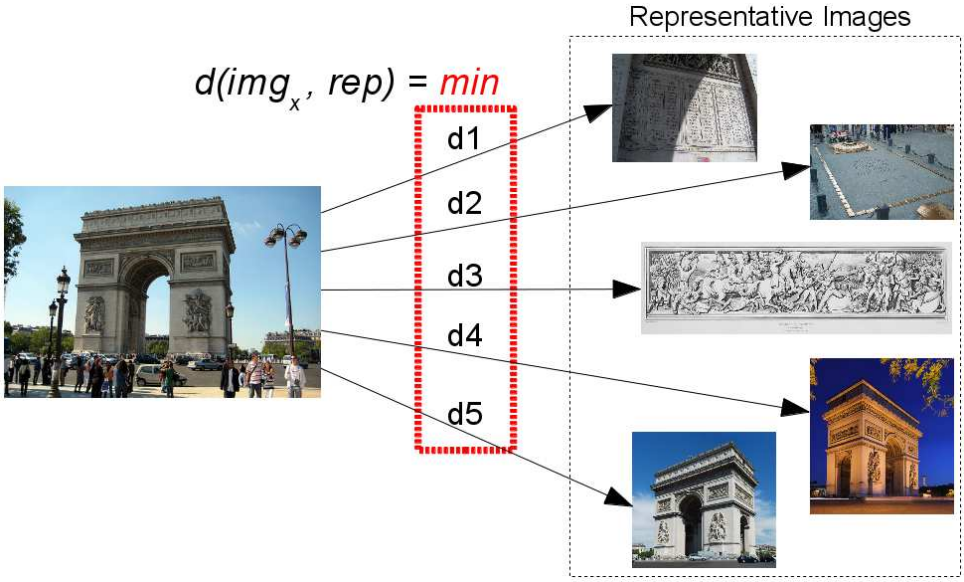

Figure 3.3: Distance from one image of a location to the whole representative set.

[3]These constraints were initially followed on relevance and diversity judgments by human assessors for the construction of the ground-truth.

[4]http://www.faceplusplus.com (As of October 6, 2015).

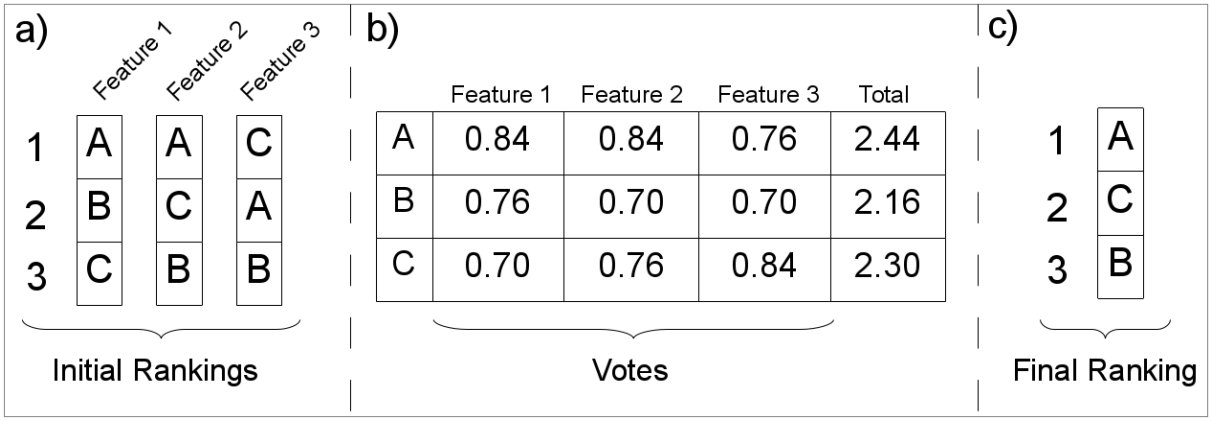

Figure 3.4: Illustrative example of the Borda Count rank aggregation algorithm.

For multiple feature fusion, we used a smoothed version of the Borda Count rank aggregation algorithm [236]. In our version, the vote (relevance score) for the n^{th} image in a ranking was computed as $\frac{1}{\sqrt[4]{n+1}}$. An illustrative example of this algorithm is presented in Figure 3.4 for three images (A, B, and C) and three features (rankings). The rankings are created according to each feature (Step a) and the votes for each image are computed and summed up (Step b). Finally, the aggregated ranking is built with the images ranked according to the total sum of votes (Step c).

We also exploited a different reranking strategy with the user-credibility descriptors provided with the data. Hence, we combined a relevance-based score (*relScore*) with another score based on credibility (*credScore*). The *relScore* of each image was computed according to its position in the list as described for the visual reranking. The *credScore* was computed as the product of three credibility features: *visualScore*,[5] *faceProportion*,[6] and *tagSpecificity*.[7] The final reranking score was computed as $relScore \times credScore$.

3.3.3 Diversification Method

After the filtering and reranking procedures, the next step consists of the actual summarization and diversification. We evaluated two diversification methods: MMR [31] (described in Section 2.4.6) and a clustering technique based-on k-Medoids [107].

The k-Medoids clustering technique is divided into two main steps: the definition of medoids and the construction of clusters. The initial centroids were defined in an offset fashion. The offset value was computed by dividing the list size by the predefined number of clusters (k). The centroids were then defined as the images in the positions $i \times$offset, with $0 \leq i < k$. Hence, the initial medoids were picked throughout the list from the top to the bottom. After the clusters are constructed, the process iterates until there is no further transition between the clusters. At each iteration, the new medoids were defined as the best connected images (average distance to all images in the cluster).

[5]The visual score represents the general relevance of the images of the user.

[6]The face proportion represents the percentage of the user images with faces.

[7]The specificity of a tag is the percentage of users which used that tag (computed in a larger Flickr dataset). Therefore, the tag specificity score of a given user is the average specificity for all tags she has used.

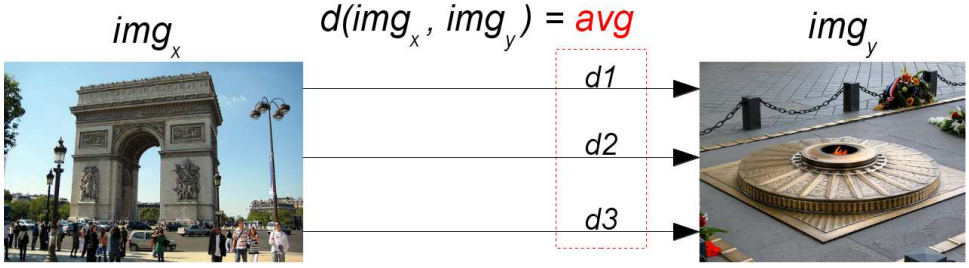

Figure 3.5: Illustrative example of the distance computed between two images for diversity assessment.

The distance between two images is computed as the average of their distances computed for each feature. Figure 3.5 presents an illustrative example of the distance between two images for three features and consequently, three distance values ($d1$, $d2$, and $d3$). Notice that each distance may correspond to a feature from a different modality, e.g., visual or textual.

Finally, the images in each cluster are ranked according to their positions in the original non-clustered list. The final output list is composed of the most relevant (top ranked) item from each cluster.

3.3.4 Experimental Setup

For the geographic filtering, according to the results on the development set, a 10km range limit from the reference point was a good threshold choice for presenting the best filtering accuracy among several values (5, 10, 15, and 20km).

In the filtering step based on the number of faces, following the best results on the development set, we eliminated all images with more than one face. Similarly, the face classifier method was configured as a 1-NN classifier and all images predicted as non-relevant were eliminated.

In the diversification module, given the consistent superiority[8] of k-Medoids over MMR on the development set, we used the k-Medoids clustering for the evaluation. For the sake of simplicity these extensive preliminary experiments are not described here in details. Since we were supposed to return 50 representative images, the algorithm was set to create 50 clusters.

Features for Reranking and Diversification

For the textual and multimodal approaches, we evaluated the TF-IDF,[9] BM25, and Cosine measures. In the development set, the best results were achieved using the Cosine measure. To enable the combination with other distance measures, the Cosine similarity values were converted by subtracting it from 1.0, therefore all distance measures are equally represented in the [0,1] interval, with 0 representing the most similar and 1 the most distant images.

[8]The methods were evaluated using each feature in separate executions.
[9]The sum of TF-IDF scores for each common term in the descriptions of two images.

For visual approaches, besides the features provided along with the image collection,[10] we also extracted two global descriptors (BIC [183] and LAS [187]) and two bag-of-visual-words (BoVW) [203] descriptors based on dense (6 pixels) or sparse (Harris-Laplace detector) SIFT, with 1000 visual words (randomly selected), soft assignment ($\sigma = 150$), and max pooling. These new features and settings have been chosen based on their successful application in previous works [30, 151].

3.3.5 Ranking Configurations

We evaluated five types of rankings and their configurations are presented in Table 3.1. The features used in each configuration and each step were selected according to the best results on the development set. For the definition of the settings in Table 3.1, we combined the best performing features according to independent executions on the development set. For the sake of simplicity, these extensive preliminary experiments on the development set are not presented here.

Table 3.1: Diversity Promotion Ranking Configurations for the first pipeline.

Ranking	**Filtering**	**Reranking**	**Diversity**
1 - Visual	GeoFilter and NumFacesFilter	-	k-Medoids ($BoVW^{sparse}_{max}$ + HOG)
2 - Textual	GeoFilter and NumFacesFilter	-	k-Medoids (Cosine)
3 - Multimodal	GeoFilter and NumFacesFilter	Visual reranking (CM3x3 + HOG + BIC)	k-Medoids ($BoVW^{sparse}_{max}$ + HOG + Cosine)
4 - Multimodal	GeoFilter and NumFacesFilter	Visual reranking (CM3x3 + HOG + BIC) and Credibility reranking	k-Medoids (CN3x3)
5 - Multimodal	GeoFilter and FaceClassifierFilter	Visual reranking (CM3x3 + HOG + BIC) and Credibility reranking	k-Medoids (CN3x3)

3.3.6 Results and Discussion

In Table 3.2, we present the official evaluation measures for the five rankings (using the 123 test queries). The best results (for all measures) were achieved when the proposed full pipeline was applied (Rankings 4 and 5). Ranking 2 (purely textual) slightly outperformed Ranking 1 (purely visual) in terms of diversity. The multimodal combination (Ranking 3) slightly outperformed Rankings 1 and 2 on CR@20 and F1@20. However when the credibility reranking was applied (Ranking 4) the best results were achieved by the visual approach with reasonable improvement on all effectiveness measures. Notice that when the face-based filtering used the classifier (Ranking 5), the results were lower than using the face number threshold (Ranking 4) but still superior to Rankings 1 to 3 on F1@20.

[10] Color histograms, histogram of oriented gradients, color moments, local binary patterns, MPEG-7 color structure descriptor, run-length matrix statistics, and spatial pyramid representation of these descriptors [93].

Table 3.2: Ranking Effectiveness - Official Measures.

Ranking	**P@20**	**CR@20**	**F1@20**
1	0.7130	0.4030	0.5077
2	0.6976	0.4139	0.5133
3	0.7016	0.4177	0.5168
4	**0.7598**	**0.4288**	**0.5423**
5	0.7407	0.4076	0.5206

Figures 3.6, 3.7, and 3.8 present a comparative of the diversified results generated with our algorithm and the baseline original ranked list retrieved from Flickr. Figures 3.6 and 3.7 present for the first 50 images, the precision and cluster recall curves, respectively. We can notice that the proposed method was able to improve the average diversity for all modalities and also for all cutoff points. However, the diversification procedure was not able to keep the precision effectiveness. Nevertheless, considering the overall precision and diversity trade-off represented with the F1-measure in Figure 3.8, all retrieval modalities achieved superior effectiveness, given that the diversity gains were high enough to compensate for the precision loss.

For a stricter analysis we computed the confidence intervals (95% confidence) for each cutoff point of the curves in Figures 3.6, 3.7, and 3.8, considering the results for 123 test queries. For clarity purposes, the error bars are not presented in the graphs, but we describe the findings in the following. In terms of Precision@N, we had statistically significant loss with Rankings 1-3, while Rankings 4-5 achieved equivalent values. In turn, analyzing the confidence intervals for CR@N, all the rankings achieved equivalent values for N=5, and statistically superior values for all the other cutoff points. Finally, for F1@N, only Rankings 4-5 allowed statistically significant superiority over the baseline (N=20 and N=30). For all other cutoff points, all rankings were considered equivalent to the baseline.

This analysis supports the statement of superior effectiveness of the proposed method against the baseline. Moreover, it highlights the importance of integrating multiple modalities, e.g., considering that the best results were achieved by Rankings 4-5 that combined geographic information, visual features, and user credibility descriptors. Beyond it, considering the relevance-diversity trade-off (assessed with the F1 measure), all modalities, and all cutoff points, the proposed method was able to achieve equivalent or superior effectiveness in relation to the baseline.

Considering our research questions and the experimental results described, we conclude that the application of filtering and reranking strategies were indeed beneficial to the whole summarization process and the combination of multiple features was actually decisive on allowing equivalent or superior effectiveness in terms of relevance and diversity.

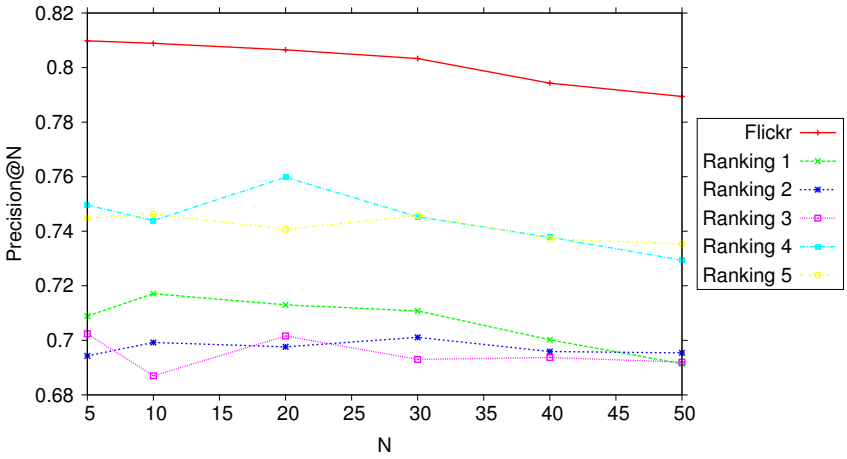

Figure 3.6: Comparative to the Flickr ranking with the Precision curve.

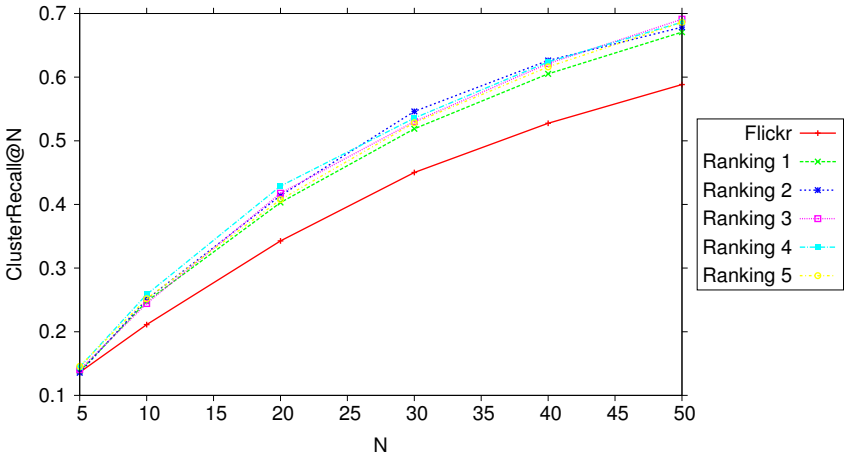

Figure 3.7: Comparative to the Flickr ranking with the Cluster Recall curve.

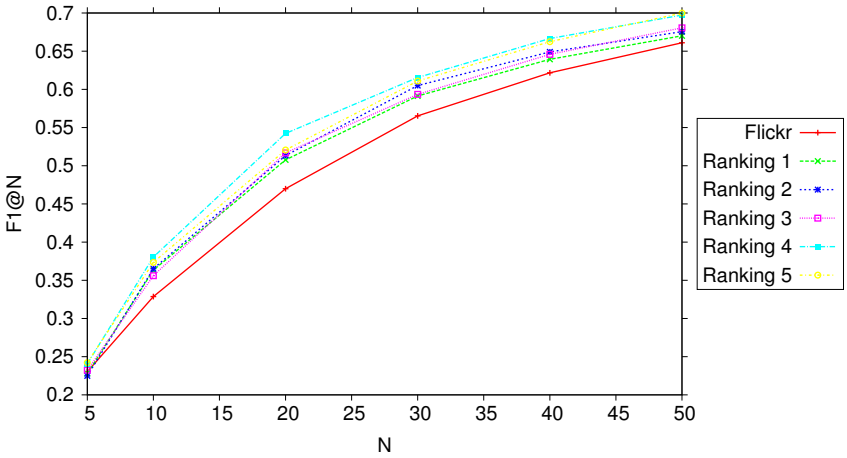

Figure 3.8: Comparative to the Flickr ranking with the F1 curve.

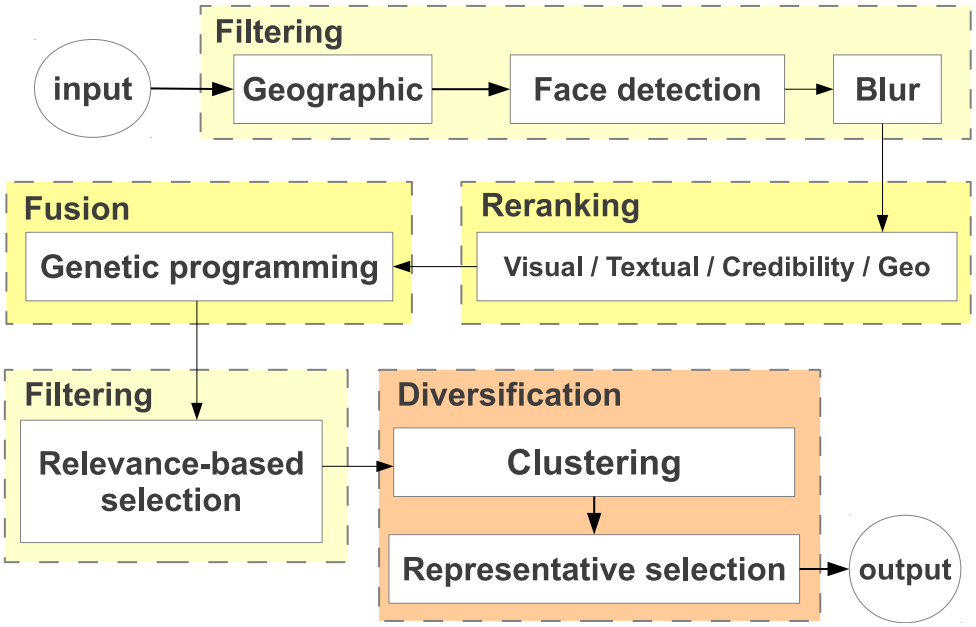

Figure 3.9: Overview of second multimodal proposed approach.

3.4 Diversity Promotion - Part 2: Exploring Multimodal Rank Fusion

This section presents our second pipeline for multimodal diversity promotion in the context of the *Retrieving Diverse Social Images Task at MediaEval 2015* [92]. This proposal is based on irrelevant image filtering, reranking, rank aggregation, and diversity promotion. We extended the previous pipeline which applied a multimodal approach and exploited image metadata and user credibility information. This new solution is described next.

The experiments were conducted with the same scenario from Section 3.3 but over an extended version of the dataset with more locations. The updated experimental image collection is divided into development and test sets with images from 153 and 139 locations, respectively. The test locations are already divided into one-concept and multi-concept queries. The one-concept queries are formulated, e.g., in relation to the name of a location whereas multi-concept queries are related to events or states associated with locations.

As a relevance enhancement step, our second approach includes irrelevant image filtering, multimodal image reranking, and rank aggregation. Image filtering was conducted according to face detection data and geographic location of the images. We also evaluated several different visual features and text similarity measures. For reranking the original retrieval list, we exploited textual, visual, geographic, and credibility information. Moreover, as an additional step, the reranked lists were also aggregated with a Genetic Programming (GP) [117] approach. Our proposal follows the general work-flow presented in Figure 3.9. Each of these steps is described next.

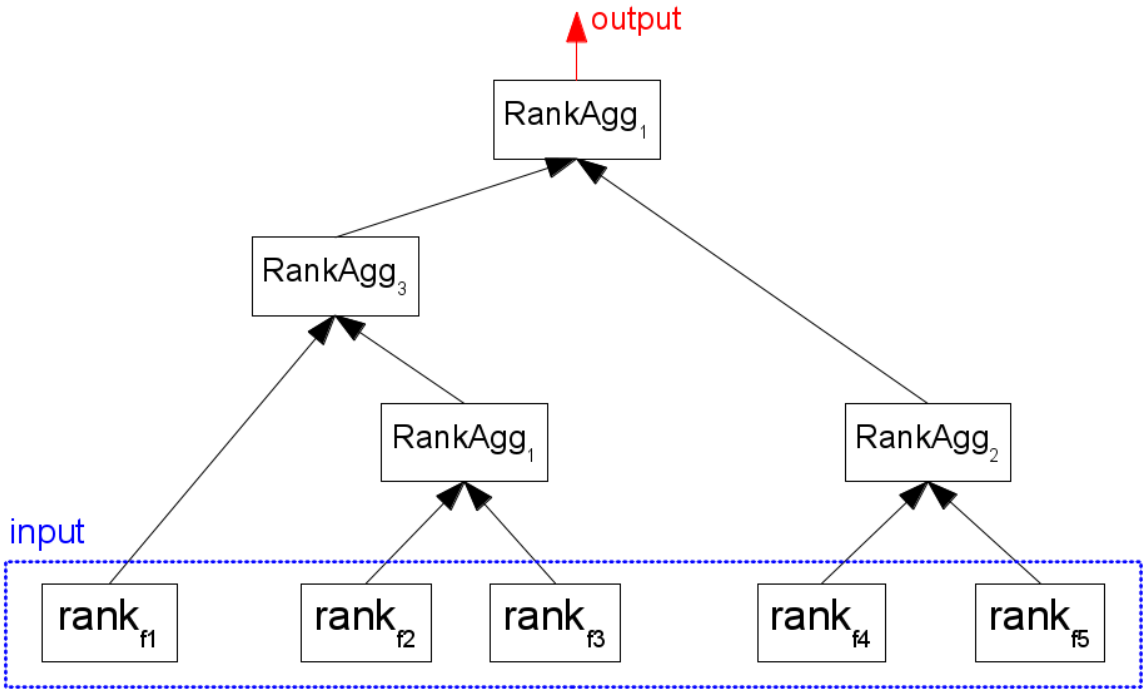

Figure 3.10: Illustrative example of the GP-based rank fusion method.

3.4.1 Filtering Step

For irrelevant images removal, we used three filtering strategies: geographic-based, face-based, and blur-based. The geographic filter eliminated all images located farther than a given range in relation to the reference location lat/long. In turn, the face-based procedure was used to filter out images containing people as the main subject. We used only the filter with number of faces from Section 3.3. Finally, for the blur-based filtering, all images considered as out-of-focus (using the method from [196]) were eliminated.

3.4.2 Reranking and Aggregation

For improving the original list ranking, we explored visual, textual, credibility, and geographic ranking. These features were individually used for the construction of reranked lists. For text-only and multimodal reranking, the text-based scores were computed as the similarity between the text vectors associated with the images and the localities' text vectors using several measures. The visual method reranked the original list according to the similarity in relation to the location's representative Wikipedia images. The visual distance from each image to the representative set was computed as the minimum distance to each representative image. All credibility scores were individually used for reranking. Additionally, lat/long data were used to rank images according to the Haversine distance to the reference point.

For feature fusion, the reranked lists were combined using the GP approach from [206] which uses several rank aggregation methods. This approach constructs optimized hierarchical fusion chains modeled as trees in which the leaf nodes (input) are ranked lists, e.g., obtained using different features or methods, and the inner nodes are ranking fusion methods. A toy example of a fusion tree is depicted in Figure 3.10. This figure depicts as fusion

combination tree for the ranking generated with five features ($rank_{f1}$, $rank_{f2}$, $rank_{f3}$, $rank_{f4}$, and $rank_{f5}$) and using three rank aggregation methods ($RankAgg_1$, $RankAgg_2$, and $RankAgg_3$).

After the optimized aggregated list is generated, a final filtering strategy was applied by eliminating the images in the bottom of the ranking. This procedure works as a relevance-based filtering and avoids that deep ranked images, here considered non-relevant, being eligible to the diversification procedure.

3.4.3 Diversification Method

After filtering, reranking, and aggregation steps, the improved relevance-based lists were submitted to diversification. We evaluated four methods: clustering-based (k-Medoids and agglomerative) and reranking-based (MMR [31], MSD [62]). In all cases, the k-Medoids method achieved significantly superior results on the development set and was used in the evaluation analysis.

In the clustering step, the initial medoids were selected in an offset fashion for equally sampling from the top to the bottom of the ranked list. The medoids updating procedure just selected the best connected image of the cluster, using the average distance to all other images in the cluster. The process iterates until no intercluster image transition occurs or up to 50 iterations.

For Rankings 1, 2, 3, and 5, the clusters were ranked according to their sizes in descending order and intra-cluster sorting was applied using average connectivity. For the credibility-based ranking (Ranking 4), the images were clustered according to their owner (user) and the clusters were ranked according to the users' credibility computed as a linear combination of all the scores used (see Section 3.4.4).

After the clusters were constructed, the representative images were selected in a round robin fashion from the final clusters.

3.4.4 Experimental Setup

Following the result on the development set, the radius for the geographic filter was set to 10 km. Similar to our first proposal in Section 3.3, we eliminated all images with more than one face. Finally, for the blur-based filtering, the out-of-focus images were eliminated with blur threshold set to 0.8.

For visual similarity, besides the features provided along with the image collection,[11] we also extracted: (i) two general purpose global descriptors (BIC [183] and GIST [140]); (ii) a bag of visual words (BoVW) descriptor, based on sparse (Harris-Laplace detector) SIFT, with 512 visual words (randomly selected), soft assignment ($\sigma = 150$), and max pooling or using Word Spatial Arrangement (WSA) [155] for encoding the spatial arrangement of visual words; and (iii) fifteen features available in the Lire package [131].[12]

[11]It includes all features used in Part 1 (see Section 3.3.4) and two convolutional neural network based descriptors [92].

[12]CEDD, FCTH, OpponentHistogram, JointHistogram, AutoColorCorrelogram, ColorLayout, EdgeHistogram, Gabor, JCD, JpegCoefficientHistogram, ScalableColor, SimpleColorHistogram, Tamura, LuminanceLayout, and PHOG. Available at: http://www.lire-project.net/ (As of October 6, 2015).

For text-only and multimodal rankings, we used the Cosine, BM25, Dice, Jaccard, and TF-IDF measures which were computed using the TF, DF, and TF-IDF vectors available with the image collection.

For user credibility, we used all scores applied in Part 1 (visualScore, faceProportion, and tagSpecificity) and also other scores. The scores used were: uploadFrequency (the average time between two consecutive uploads), meanTagRank (the mean rank of tags of a user in a list in which the tags are ranked in descending order considering the number of appearances in a large subsample of Flickr images), meanImageTagClarity (an adaptation of the Image Tag Clarity from [185] using as individual tag language model the tf/idf), photoCount (the total number of images a user contributed), and locationSimilarity (the average similarity between the geotagged photos from a user and a probabilistic model of a surrounding cell). These credibility scores were be best performing ones on the preliminary experiments with the development set.

For the combination of the multiple features, the GP-based rank fusion method was trained using the development data and combined order-based (MRA [52], RRF [40], and Borda Count [236]) and score-based (CombMIN, CombMAX, CombSUM, ComMED, CombANZ [177], and RLSim [152]) rank fusion methods.

As a relevance-based filtering criterion, from the final aggregated list, up to 150-top ranked images were selected as the input list for the summarization procedure.

For diversification, the number of clusters was defined as 30 for Rankings 1 and 3, and 40 for Rankings 2 and 5. These were the best configurations on the development set for the selected features.

3.4.5 Ranking Configurations

We evaluated five types of rankings (Table 3.3). The features used for diversification were selected according to the best results in extensive experiments with the development set. For the reranking procedure, all the features were used for the construction of independent rankings which were then aggregated by the fusion module.

In all rankings, the geographic filtering and reranking were only applied for one-topic queries since multi-topic queries do not have reference geo-location. Additionally, in Ranking 5, the face-based filters were also only applied to one-topic queries since multi-topic queries have a different relevance constraint in relation to people in the foreground. Finally, in Rankings 1, 3, and 5 no visual reranking was applied for multi-topic queries.

Table 3.3: Ranking Configurations (*only for one-topic queries).

Ranking	**Filtering**	**Reranking**	**Diversity**
1 - Visual	Geo*, face, blur	Visual*	BIC
2 - Textual	Geo*, face, blur	Textual	Cosine + Jaccard
3 - Multimodal	Geo*, face, blur	Visual*, textual	Jaccard
4 - Credibility	Geo*, face, blur	Credibility	Users
5 - Multimodal	Geo*, face*, blur	Visual*, textual, credibility, geo*	Jaccard

Table 3.4: DevSet and TestSet Results.

	DevSet			TestSet		
Ranking	**P@20**	**CR@20**	**F1@20**	**P@20**	**CR@20**	**F1@20**
1	0.7487	0.4336	0.5409	0.7129	0.4111	0.5063
2	0.8013	0.4514	0.5694	0.6996	0.4248	0.5101
3	0.7837	0.4436	0.5592	0.7058	0.3881	0.4883
4	0.7644	0.4446	0.5532	0.7198	0.4309	**0.5219**
5	0.8190	0.4637	**0.5853**	0.7324	0.4123	0.5084

Table 3.5: TestSet Results: One-topic and Multi-topic.

	One-topic			Multi-topic		
Ranking	**P@20**	**CR@20**	**F1@20**	**P@20**	**CR@20**	**F1@20**
1	0.6906	0.4000	0.4991	0.7350	0.4221	**0.5133**
2	0.7130	0.4316	0.5205	0.6864	0.4181	0.4998
3	0.6942	0.3982	0.4970	0.7171	0.3782	0.4798
4	0.7630	0.4301	**0.5390**	0.6771	0.4318	0.5051
5	0.7290	0.4286	0.5228	0.7357	0.3963	0.4942

3.4.6 Results and Discussion

Table 3.4 presents the effectiveness results for the five rankings for the development (123 queries) and test set (139 queries). The best results (F1@20) on the development set were achieved by Ranking 5, followed by Rankings 2 and 3, in which textual information was used. However, these were the rankings with the greatest effectiveness difference when comparing development and test queries, specially considering the multi-topic queries.

Table 3.5 presents the effectiveness results for one-topic (69 queries) and multi-topic (70 queries) test queries. As we can observe, even with no visual reranking, the visual-only ranking allowed slightly superior results for multi-topic queries considering all the ranking types and also comparing to one-topic queries. All other ranking types achieved superior effectiveness on one-topic queries, specially when the credibility information was used (Rankings 4 and 5).

Figures 3.11, 3.12, and 3.13 present a comparative of the diversified results generated with our algorithm and the original ranked list retrieved from Flickr. Differently from the results presented in Section 3.3.6, this new approach allowed in general superior diversity effectiveness (Figure 3.12), while it was also possible to improve the precision of the diversified result (Figure 3.11). Consequently, all modalities also achieved superior relevance-diversity trade-off (Figure 3.13).

As presented in Section 3.3.6, for our second proposal, we also computed the confidence intervals, this time considering the 139 available queries. Again, for clarity purpose, the error bars are not presented in the graphs. Here, for the Precision@N curves, all rankings produced great variation of precision values for the queries, therefore, for all rankings and all cutoff points our method was considered equivalent to the baseline. This result is

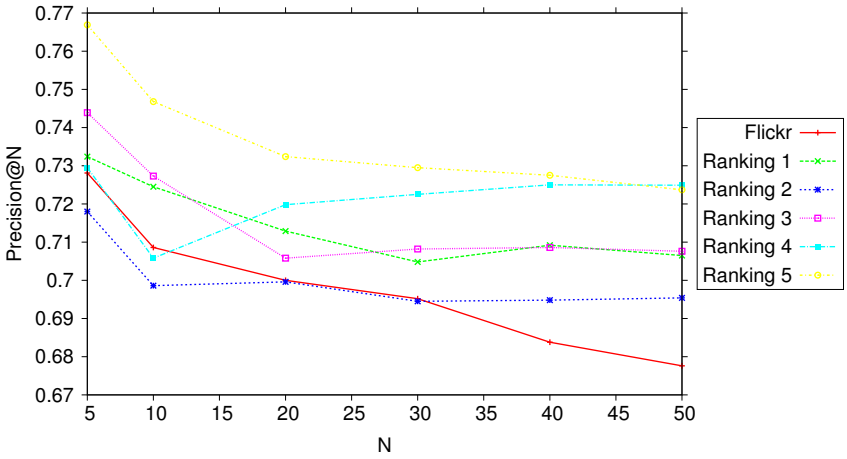

Figure 3.11: Comparative to the Flickr ranking with the Precision curve.

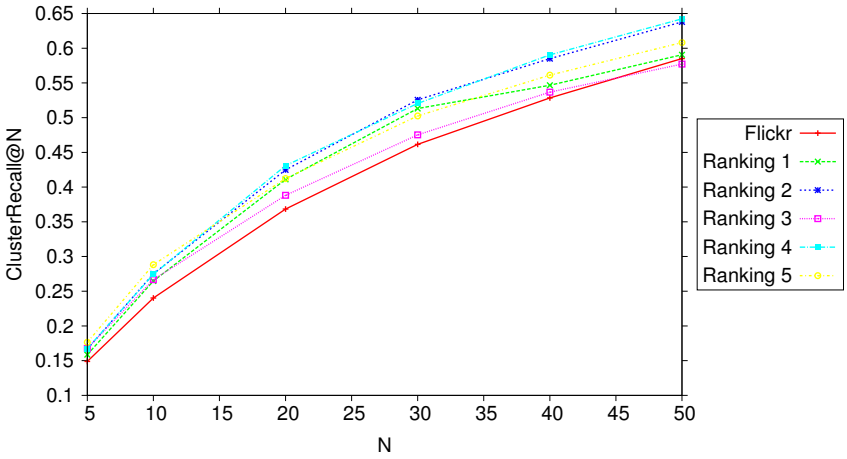

Figure 3.12: Comparative to the Flickr ranking with the Cluster Recall curve.

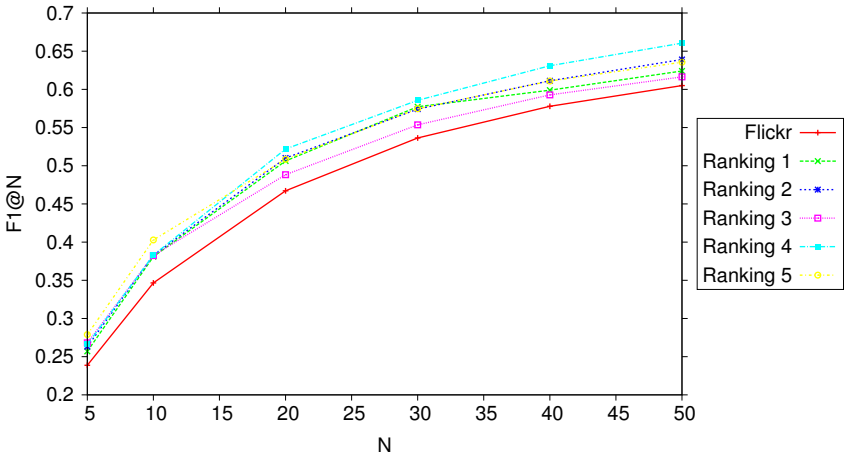

Figure 3.13: Comparative to the Flickr ranking with the F1 curve.

promising, considering that our previous proposal, for Rankings 1-3, achieved statistically inferior precision values. For the CR@N results, our methods achieved equivalent or superior diversity. As important factors, we highlight the statistically superior diversity achieved by Rankings 2 and 4 for N=20-40. In turn, Ranking 5 achieved superior diversity for N=5-10, which may be a consequence of the integration of more sources of information than in Rankings 1-4. When considering the relevance-diversity trade-off, Ranking 4 achieved superior values for F1@20, while Ranking 5 achieved superior balance for F1@5 and F1@10. These results are also a consequence of the greater average precision and superior diversity effectiveness for such cutoff points.

3.5 Summary and Considerations

For balancing the relevance-diversity trade-off, we proposed and evaluated two methods which jointly exploited several types of information, including, visual, textual, geographic, and user credibility. The methods and results described in this chapter were partially published in [28] and [29].

Our first proposal was a multimodal approach with the use of filtering and reranking approaches in conjunction with a clustering technique for diversification. Our best results were achieved with image reranking by combining their relevance score and user credibility information. In Appendix D, we present a set of visual examples of the results for the queries evaluated in the experimental analyses.

With an extended pipeline, we proposed filtering strategies and the combination of multiples features with a learning-to-rank fusion method. These improved ranked lists were used as input for a clustering-based summarization method. Our experiments suggest that different summarization alternatives may result in different effectiveness for one-topic and multi-topic queries. Moreover, our results suggest that visual features are important when considering multi-topic queries while the textual information seems more suitable for one-topic queries. Nevertheless, such behavior deserves more focused experiments and additional analysis for proper validation.

In this chapter, we have seen that combining multiple sources of evidence is effective for relevance-diversity enhancement and balancing. Moreover, both proposed methods were able to statistically outperform the results from the baseline retrieval with the standard Flickr algorithm.

Considering these findings, in Chapter 4, we evaluate the potential of diversity promotion on aiding learning-to-rank methods in an interactive retrieval scenario. Moreover, we consider a different evaluation environment and several methods for diversity-oriented retrieval sessions.

Chapter 4

Diversity-driven Interactive Learning

In interactive retrieval systems, one of the main objectives is to maximize the user information gain throughout search cycles. Retrieving many relevant items is quite important for this process but it does not necessarily satisfy user needs completely. When only relevant near-duplicate items are retrieved, the amount of different concepts that the users are able to extract from the target collection is very limited. Therefore, broadening the number of concepts represented in a result set may improve the search experience. Diversifying item concepts in the retrieved set is one of the methods for increasing the information gain in a single search iteration, maximizing the probability of including at least some relevant items for each interpretation of ambiguous or underspecified queries. Moreover, at the same time, relevance feedback approaches may take advantage of diverse results and improve internal machine learning models.

In this chapter, considering the scenario of interactive learning with diversity, we revisit our research question about the possibility of improving the diversity with small or even no impact on relevance. Moreover, we experimentally investigate whether it is possible to boost interactive learning with diversity. Therefore, we introduce a new genetic programming approach for enhancing the user search experience based on relevance feedback over results produced by a multimodal image retrieval technique with diversity promotion. We have studied maximal marginal relevance reranking methods for result diversification and their impacts on the overall retrieval effectiveness. We show that the learning process using diverse results may improve user experience in terms of both the number of relevant items retrieved and subtopic coverage.

In this chapter, we propose an approach for enhancing the user experience by interactively learning with user feedback over diversified results produced by a multimodal image retrieval technique. In other words, we propose an interactive image retrieval method, which exploits, at the same time, user relevance feedback and diversity promotion on top of a multimodal genetic programming (GP) framework. In such method, relevance feedback (RF) and diversification are completely integrated in the sense that: (i) the user feedback over diversified results is incorporated by the GP-based learning; and (ii) the diversification also exploits relevance information from the user feedback. As far as we know, our method is the first to simultaneously exploit user relevance feedback, diversity promotion, and multimodal learning into an single integrated framework. Our experiments demonstrate that learning with diverse items helps improving overall diver-

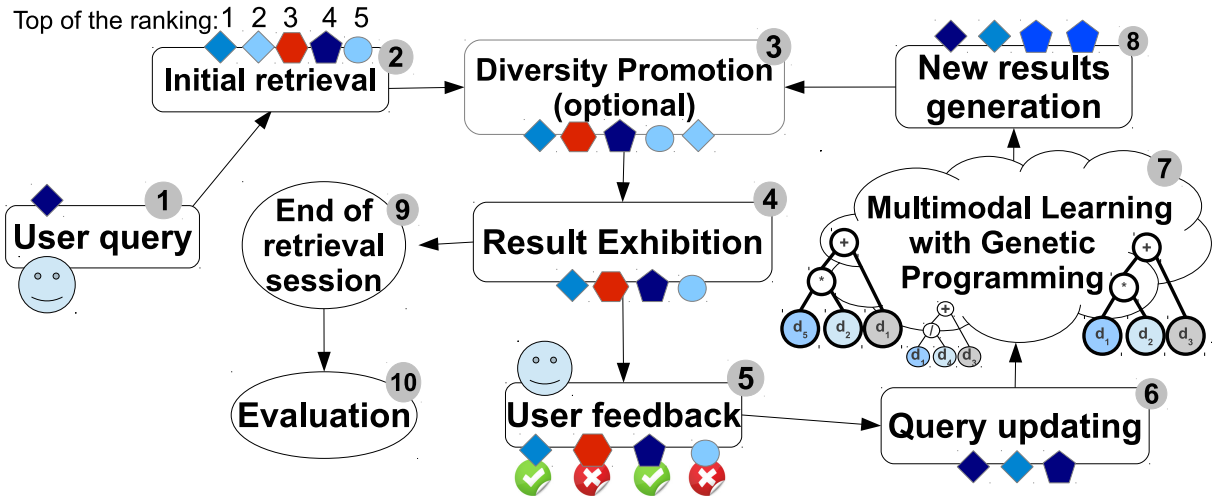

Figure 4.1: Overview of the RF framework with diversity promotion. Source: [27].

sity during a search session while simultaneously retrieving more relevant items in fewer feedback iterations.

Different from previous work described in Chapters 2 and 3, we propose an integrated interactive method with user relevance feedback over diversified results produced using a multimodal learning framework and an MMR-like reranking method.

The remainder of this chapter is organized as follows. Section 4.1 describes our diversity-driven learning method. Section 4.2 introduces the experimental settings for the evaluation. Section 4.3 presents our empirical results and discussion. Finally, Section 4.4 presents a summary and considerations.

4.1 Diversity-driven Learning with RF

In this section we describe our baseline framework and the characteristics of a relevance feedback session (Section 4.1.1), the method proposed for learning with diversified results (Section 4.1.2), and the interactive diversification approaches we have evaluated by means of the simulation of user relevance feedback sessions (Section 4.1.3).

4.1.1 Relevance Feedback Framework

We use the RF engine proposed in [30] as basis for our experiments. It is a framework for multimodal image retrieval that uses genetic programming. The learning approach is used for the optimization of multimodal similarity functions based on user feedback. We have chosen this method as it presents state-of-the-art effectiveness (in terms of relevance) for multimodal retrieval.

An overview of the relevance feedback session is presented in Figure 4.1, in which the small geometric forms represent images while their different shapes and colors refer to different content in terms of visual or textual features.

Relevance feedback workflow

In order to start a session, the user provides a query that we consider potentially multimodal. Different query features may be used and the retrieval modalities studied are described in Section 4.2.1. The user may provide a query using a textual sentence and/or some images (Step 1). An initial result set is constructed based on the average distance for the set of textual and/or visual measures (Step 2). The top items in this result are eligible to be reranked for diversity promotion (Step 3), and are presented to the user (Step 4). The user then provides relevance feedback for the initial result, e.g., by clicking/marking on the desired images in the result page (Step 5). Using the feedback, the system updates the query by adding all relevant images and consequently their visual and textual features (Step 6). With this new expanded query pattern, a GP-based learning method is used at runtime[1] for finding the best combination functions for the distance measures (textual and visual) that more properly encodes the characteristics of the new query (Step 7). Using the best functions produced, the residual collection is reranked (Step 8). This ranked list is generated through a voting scheme in which for each selected ranking function, the collection is sorted and each image receives a vote inversely proportional to its position in the ranking ($\frac{1}{k}$, where k is the image position in the ranked list). The new ranking score for each image is the sum of the votes given by the ranking functions. The images are sorted according to the votes and are eligible to diversity promotion (back to Step 3). The top of the reranked list is then selected for exhibition (Step 4). The user analyzes the new result and decides to end the retrieval session if she is satisfied (Step 9) or start a new learning iteration by providing feedback (back to Step 5).

Effectiveness evaluation is performed at the end of the session (Step 10). Notice that only the residual collection is reranked after each feedback iteration which means no previously seen image is ever presented again. For more details on the evaluation protocol please refer to Section 4.2.4.

Diversity promotion of the retrieved images is performed in different interactive approaches that are described in Section 4.1.3. The details on the textual and visual measures we used are presented in Section 4.2.3.

Genetic Programming learning system

The GP framework (Step 7 in Figure 4.1) is used for the combination of textual and visual measures into a single multimodal ranking function (GP individual). The individuals are modeled as trees with similarity measures being the leaves (input) and arithmetic operators the inner nodes. An example of individual is presented in Figure 4.2. In this example, the complex similarity combination function $\frac{m_1 \times m_2}{tfidf - m_3} + \sqrt{m_4}$ is used to combine similarity measures defined by visual (m_1, m_2, m_3, and m_4) and textual features ($tfidf$).

[1]The execution cost of the method is directly dependent on the genetic programming configuration, e.g., the size of the population and the number of generations. Another important factor is the amount of images presented per iteration, which is in general very limited and, in this case, defines how much query data the system has to process. Therefore, the responsiveness of the process may be adjusted according to the application needs. Moreover, the genetic programming process is naturally parallelizable which makes this method suitable for modern computing architectures.

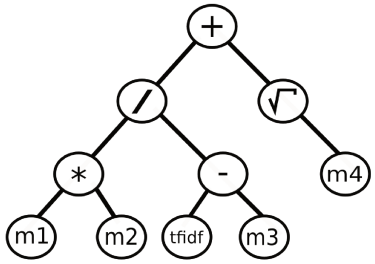

Figure 4.2: Example of GP individual: $m1$, $m2$, $m3$, and $m4$ are visual measures and $tfidf$ is a textual similarity measure.

1: Generate the initial population of individuals
2: **for** N generations **do**
3: Compute the fitness of the individuals
4: Select the individuals for genetic operations
5: Apply reproduction
6: Apply crossover
7: Apply mutation
8: **end for**
9: **return** the best individuals found

Figure 4.3: General genetic programing algorithm.

A general algorithm for the GP process is presented in Figure 4.3. In the learning process, a group of candidate individuals are iteratively evolved using the genetic operators such as reproduction, crossover, and mutation.

During reproduction, "adapted" individuals are directly copied to the next generation. This is done to preserve "good genotypes." Mutation occurs by probabilistically replacing some nodes of an individual (distance measures or arithmetic operations). It is used to reduce some biases and avoid the solution to "get trapped" in some local minimum. Crossover combines pieces of different functions, which means exchanging parts among functions, trying to build upon good solutions.

For optimizing and selecting the best functions at each feedback iteration, a training set is created using the query images, positive examples provided as feedback, negative examples (images not marked by the user as relevant) and random images from the residual collection. This training set is used for the evaluation and selection of the new functions that are going be used for ranking the residual collection.

More formally, let Q be the query pattern, T be the training set, T_d the desired number of training objects, and P the set of objects presented to the user per iteration. At each RF iteration, assuming we have $|Q|$ objects in the query pattern, the training set T is composed as the following:

- If $|Q| \leq |P|$, all objects in Q are added to T and $|P| - |Q|$ non-labeled objects are randomly selected from the ones the user has already seen and have not marked as relevant. These randomly selected items are considered non-relevant;

- If $|Q| > |P|$, $|P|$ objects are randomly copied from the query pattern to the training set; and

- In order to fulfill the training set, $T_d - |P|$ objects are randomly picked from the residual collection and assumed as non-relevant.

To compute the fitness of a given candidate individual (Υ), a ranking ($\Re$) is created for the training set T in relation to the updated query Q using Υ as ranking function. The similarity of a given training set item to Q is computed as the average similarity to each item in Q. The quality score (Φ) for $\Re$ is computed according to the ranking positions of the relevant images using Equation 4.1.

$$\Phi(\Re) = \gamma \times \sum_{p=1}^{|P|} \Omega(p) \times \Gamma(\Re(T, Q, \Upsilon)[p]) \tag{4.1}$$

where γ is a constant, Ω is a decreasing function, $\Re[p]$ is the p-th object in the ranking $\Re$, and Γ indicates the relevance of the object (1 for images marked as relevant or 0 otherwise).

The GP-based learning method was chosen as it provides an automatic way of tuning the importance of the textual and visual features available. Therefore, better ranking functions are expected to be discovered during the retrieval iterations according to the information obtained with user feedback. For more details regarding the GP framework, see [30].

4.1.2 Diversity Promotion

Although intent-aware methods have been reported to be more effective than non-intent oriented ones, defining query intents or even detecting multi-intent queries is still an open issue [20, 106, 172]. Additionally, as far as we know, there is no established CBIR benchmark incorporating ambiguous queries with explicit sets of subtopics in a multimodal task.

Among the most common non-intent oriented diversification methods, the clustering-based and the dissimilarity-based greedy selection algorithms are reported to be the more effective for different tasks. Hence, for learning with diversity (Step 3 in Figure 4.1), our work applied an approach inspired in the method from Carbonell & Goldstein (1998) [31]. It works as a reranking procedure over the top-K most relevant candidate items. The diversity promotion is performed by reranking the first K images of the ranking where, in our case, K is proportional to the number of images that will actually be shown to the user.

This is conducted by incrementally selecting, from the list of candidates, the most similar to the query that is at the same time the most diverse (different) from the previously selected ones. For this, a combined score is computed using relevance and diversity scores. The first image of the original list is selected as the first one in the reranked list S. After that, the image that will be selected for the next position in S is the one that minimizes the new relevance-diversity combined distance score. The process then iterates for the remaining candidate images until the final number of images that will be shown to user is achieved.

Provided we have a set D including visual descriptors and textual distance measures, the relevance scores are computed as the distance between the candidate images and the query. Since, in this context, the query is interactively updated, based on feedback, by adding positive examples, the relevance score ($rel(c_k, Q)$) between a candidate image (c_k, $k \leq K$) and the whole query (Q) is the minimum value, corresponding to the minimum average distance to the query images ($q_n \in Q$) using all available distance measures $\theta_j \in D$. This score is represented in Equation 4.2.

$$rel(c_k, Q) = min\left(\frac{\sum_{i=1}^{|D|} \theta_i(c_k, q_n)}{|D|}\right) \mid q_n \in Q \tag{4.2}$$

In turn, the diversity score between a candidate image and the images previously selected ($s_m \in S$) is the minimum value computed as follows:

$$div(c_k, S) = min\left(\frac{\sum_{i=1}^{|D|} \theta_i(c_k, s_m)}{|D|}\right) \mid s_m \in S \tag{4.3}$$

Similar to Equation 2.1, the new ranking score (Equation 4.4) is obtained with a linear combination between the relevance and diversity scores.

$$score(c_k, Q, S) = (1 - \lambda) \times rel(c_k, Q) + \lambda \times (1 - div(c_k, S)) \tag{4.4}$$

After the diversity promotion the top of the reranked list is selected for user assessment. For fusion purposes, the distance values for all measures were normalized to [0,1].

4.1.3 Interactive Diversification Approaches

To assess the diversity impact on the learning process throughout the feedback iterations, we analyze four alternative methods:

(i) NoDiv: No diversity promotion is performed at all;

(ii) InitDiv: Diversification is performed only for the first result set presented to user;

(iii) FullDiv: Diversification is performed for all result sets, i.e., at each feedback iteration; and

(iv) NoInitDiv: Diversification is performed for all result sets, except for the first feedback iteration.

Considering the natural drawbacks of the diversification methods regarding the deterioration of the relevance of the diversified results and the few training samples that are available at the beginning of a search session, the NoInitDiv approach was introduced. By not diversifying the first result set, the system allows the user to receive more relevant items in the first iteration, expand the amount of training information, which in turn strengths the learning method against the relevance decrease. This rationale will be better discussed in Section 4.3.

4.2 Experimental Setup

For the proper understanding of the behavior of the retrieval approaches, it is important to have a precise benchmark configuration. In this section, we describe the dataset, features, interaction model, diversification, learning, and evaluation resources we used in the experiments.

4.2.1 Dataset and User Model

We use the *ImageCLEF Photographic Retrieval Task* collection [190] with 20,000 images. Each image is associated with textual metadata such as title, description, and date. The 39 topics of the task were used for the evaluation, being considered proper for diversification [11], as they simulate multi-interpretation queries with diversity requirements. For the task, each query is composed of a small text fragment and three sample images.

For the experimental analysis, we simulated a task in which the search begins with only textual data and the system can use visual features after the first feedback iteration. This corresponds to a scenario in which the user starts the search using a text fragment as query and proceed for some iterations providing feedback after visually assessing the retrieved images.

For user feedback simulation over a retrieval session and effectiveness evaluation of the results, we used the relevance assessments provided with the collection. Following [27], we also assume an ideal user model, who labels all the expected relevant images for each topic as positive examples. Similarly, the configuration of the relevance feedback session is as follows: (i) 10 feedback iterations; (ii) 20 images shown to the user per iteration; and (iii) 4 voting (best) GP individuals per iteration (using Borda Count rank fusion [236]). The configuration for the baseline GP process was defined according to Table 4.1, based on the parametric search conducted in [30].

Table 4.1: Genetic programming settings (Source [30]).

Parameter	**Value**
Population size	60
Number of generations	20
Crossover rate	0.8
Reproduction rate	0.0
Mutation rate	0.2
Minimum initial individual depth	2 [53]
Maximum initial individual depth	5
Individuals selection threshold (α)	0.999
Training set size	55
Constant of the fitness function	2 [53]
Fitness decreasing function	$g(l) = \log_{10}(\frac{1000}{l})$
Operators	+, *, /, and $\sqrt{\ }$

4.2.2 Diversification Parameters

For the expected items for each query, the collection includes cluster ground-truth in which the relevant images are grouped according to some concepts, e.g., by location or vehicle type. In the experiments we have evaluated all combinations of values for the diversity importance factor (λ in Equation 4.4) and reranking depth (d).

The depth is the number of images from the top of the similarity-based input rank that are reranked for diversity promotion. The reranking depths were picked proportionally to the number of images the user sees at each iteration. The depths used were 1.5, 2, 3, 4, and 10, corresponding to 30, 40, 60, 80, and 200 images, respectively. Preliminary exploratory experiments revealed that using deeper reranking for diversification of the relevance-based list ended up with too much relevance harm by considering images not well ranked and consequently noisier information.

The tested values for the diversity factor were 0.1, 0.3, 0.5, 0.7, 0.9, and 0.99. Moreover, we evaluated the effectiveness for all the combinations of the reranking depth and diversity factor.

4.2.3 Visual Features and Text Processing

The experiments used eight global visual descriptors: GCH [186], BIC [183], ACC [85], and JAC [223] as color features; CCOM [116], LAS [187], QCCH [84], and HTD [227] as texture features. For the textual modality, the following similarity measures between the query text and the metadata of the images were computed: Cosine [17], BM25 [17], Dice [122], Jaccard [122], tf-idf-sum [168], and Bag-of-words (normalized terms intersection). As textual feature only the description field of the images and the title field of the topics were used. Stop words removal and stemming were applied. Only the English metadata was considered.

4.2.4 Evaluation Protocol

The experiments were performed with the full-freezing evaluation approach [111] (see Section 2.6.1) in which all items are frozen at the positions they are retrieved and the results of further iterations are appended to the results of the previous ones. Consequently, the highest rank position of a new item is defined by the number of items previously judged plus one [224]. This way we avoid artificial improvements from the so called "ranking effect" [167] and properly assess the system effectiveness in relation to the real user effort on providing feedback.

For evaluation purposes, the system is expected to retrieve the most relevant and the most diverse items at every iteration. Hence, for relevance-based effectiveness, we used traditional measures such as MAP, GMAP, BPREF, and Recall@N (R@N).We have chosen to use the Recall instead of Precision for better representing the evolution of the amount of relevant elements found. Since the recall measure does not take into account the rank position directly in its computation, it does not get biased (smoothed) by the depth of evaluation. Therefore, it makes the Recall more suitable for comparing the results from different iterations, as well as properly representing the effectiveness evolution throughout

a session.[2] In turn, for diversity measurement, we used CR@N. With the full-freezing protocol, the evaluation measures were computed for the 200 images in the exact order they were seen by the user, assessing how early relevant items were found. In this context, R@200 measures the amount of relevant items retrieved at this depth and CR@200 the diversity of the group of items the user gets exposed to at the end of the session. All statistical analysis were performed using the Wilcoxon's test with p-value < 0.05.

4.2.5 Baselines

Besides NoDiv, we have proposed two other baselines: BordaMMR and MinAvgMMR to evaluate the different retrieval modalities. The goal is to assess the learning impact of the GP-based framework in the overall process in comparison to simple rank fusion strategies. These baselines are considered as supervised fusion-based techniques in the relevance portion of the relevance-diversity combination given that at each iteration the algorithms process, as reference for ranking, the query that was updated with user feedback.

The BordaMMR is a variation of the Borda Count algorithm (described in Section 3.3.2) followed by the diversification approach described in Section 4.1.2. In BordaMMR, each textual or visual feature was used as a voter with each one voting for the top-20 images in its corresponding ranking. For the ranking using each feature, the distance between a given image from the collection and the query pattern was taken as the minimum distance between the collection image and each image in the query. The MinAvgMMR function-based baseline uses Equation 4.2 (Section 4.1.2) for ranking and the same diversification approach.

4.3 Results and Discussions

We executed simulations for each depth and diversity weight combination and conducted two rounds of experiments. In the first round, the configurations that yielded the best results in Recall@N, CR@N, and Precision X Recall for the proposed methods were compared against our baselines (Table 4.2). Statistical significance was computed with Wilcoxon's test for p-value < 0.05. For all measures, MinAvgMMR performed consistently better than BordaMMR, with statistical superiority in the initial results and after the 4th iteration. For Recall@N, we can notice that the proposed learning methods outperformed the baselines right after the first feedback. The inferiority of InitDiv and FullDiv in recall and diversity in the first result set is probably a consequence of the diversification, harmed by the small amount of information available (only the query items). This effect was reduced and extinguished over the iterations with more labeled data from feedback.

Regarding relevance, InitDiv and FullDiv (that include diversification since the first result set) performed quite similar to NoDiv, with statistical superiority only in late iterations. In turn, NoInitDiv (that does not suffer from the consequences of the initial diversification) performed quite better after the 4th feedback iteration. Using the diversity

[2]For this purpose, one could just rely no the number of relevant items found but it would not carry any information related to the difficulty of the query in relation to the amount of relevant elements available in the target collection.

Table 4.2: First Round Result Summary. Parameters d and λ refer to the diversification depth and the linear combination factor, respectively. Wilcoxon's test: Bold face means superiority over the BordaMMR. △ means superiority over the MinAvgMMR. ▽ means inferiority over the MinAvgMMR. ▲ means superiority over the NoDiv.▼ means inferiority over the NoDiv.

	Recall@N																							
Method	**d**	λ	**R20**		**R40**		**R60**		**R80**		**R100**		**R120**		**R140**		**R160**		**R180**		**R200**			
BordaMMR	10	0.3	0.0825		0.1478		0.2091		0.2502		0.2757		0.2868		0.3055		0.3235		0.3388		0.3505			
MinAvgMMR	1.5	0.99	**0.0959**		0.1558		0.2304		0.3140		**0.3964**		**0.4422**		0.4935		**0.5418**		**0.5819**		**0.6155**			
NoDiv	-	-	**0.0959**		**0.1900**	△	**0.2798**	△	**0.3538**	△	**0.4292**		**0.4938**	△	**0.5463**	△	**0.5880**	△	**0.6146**		**0.6365**			
InitDiv	1.5	0.99	0.0817	▽▼	**0.1850**	△	**0.2737**	△	**0.3451**	△	**0.4172**		**0.4765**		0.5285	△	**0.5847**	△	**0.6336**	△	**0.6591**	△▲		
FullDiv	4	0.3	0.0758	▽▼	**0.1797**	△	**0.2759**	△	**0.3660**	△	**0.4482**	△	**0.5054**	△	**0.5532**	△	**0.5991**	△	**0.6428**	△▲	**0.6754**	△▲		
NoInitDiv	4	0.3	**0.0959**		**0.1898**	△	**0.2947**	△	**0.3858**	△▲	**0.4722**	△▲	**0.5356**	△▲	**0.5832**	△▲	**0.6224**	△▲	**0.6609**	△▲	**0.6978**	△▲		
	CR@N																							
Method	**d**	λ	**CR20**		**CR40**		**CR60**		**CR80**		**CR100**		**CR120**		**CR140**		**CR160**		**CR180**		**CR200**			
BordaMMR	10	0.3	0.2767		0.3922		0.4779		0.5152		0.5245		0.5386		0.5575		0.5811		0.5992		0.6123			
MinAvgMMR	1.5	0.9	0.2709		0.3857		0.4721		0.5319		**0.6163**		**0.6502**		**0.6673**		**0.6958**		**0.7231**		**0.7491**			
NoDiv	-	-	0.2709		0.3945		0.4916		0.5603		**0.6020**		**0.6535**		**0.6891**		**0.6947**		**0.7147**		**0.7201**			
InitDiv	4	0.5	0.2229	▽▼	0.4178		0.5165		**0.6000**	△	**0.6628**	▲	**0.6993**		**0.7322**	△	**0.7476**		**0.7598**		**0.7644**			
FullDiv	4	0.7	0.1933	▽▼	0.3727		0.4718		0.5276		0.5939		**0.6558**		**0.6958**		**0.7362**		**0.7491**		**0.7812**			
NoInitDiv	4	0.5	0.2709		0.3967		0.5071		0.5900		**0.6525**		**0.6779**		**0.7366**	△	**0.7583**	△▲	**0.7867**	△▲	**0.8027**	△▲		
	Precision@Recall																							
Method	**d**	λ	**R0.0**		**R0.1**		**R0.2**		**R0.3**		**R0.4**		**R0.5**		**R0.6**		**R0.7**		**R0.8**		**R0.9**		**R1.0**	
BordaMMR	10	0.1	0.4885		0.3328		0.2837		0.1680		0.1180		0.0746		0.0539		0.0414		0.0335		0.0228		0.0000	
MinAvgMMR	1.5	0.99	0.4853		0.3852		0.3431		**0.2976**		**0.2441**		**0.2188**		**0.1887**		**0.1442**		**0.1004**		0.0518		0.0102	
NoDiv	-	-	0.5017		**0.4146**		**0.3677**		**0.3413**		**0.2759**		**0.2505**		**0.2226**		**0.1906**		**0.1354**		**0.0837**		0.0347	
InitDiv	1.5	0.1	0.4932		**0.3860**	▼	**0.3579**		**0.3278**	△	**0.2792**	△	**0.2553**	△	**0.2150**	△	**0.1809**	△	**0.1233**		**0.0774**	△	0.0229	
FullDiv	3	0.1	0.5097		**0.4000**		**0.3716**	△	**0.3493**	△	**0.3045**	△▲	**0.2638**	△	**0.2449**	△	**0.2176**	△	**0.1405**		**0.0972**	△	0.0312	
NoInitDiv	4	0.3	0.5160	△	**0.4318**	△	**0.3911**	△ ▲	**0.3602**	△	**0.3117**	△▲	**0.2804**	△	**0.2523**	△	**0.2326**	△	**0.1515**	△	**0.0930**	△	0.0258	

measurements we can see a similar behaviour of the methods throughout iterations except for NoInitDiv that achieved statistically superior results in later iterations, explained by the larger amount of relevant images retrieved throughout the iterations.

In the second round, for each retrieval method, we selected the three best rankings for comparison against NoDiv. These three rankings were selected using the best values of MAP, R@200, and CR@200, respectively. Additionally to the original ranking, we used other ten different random seeds and computed average values for each measure. Statistical significance was assessed with confidence intervals for $\alpha < 0.05$. A summary of the results is presented in Table 4.3 and statistically significant improvements are highlighted in bold face. The best rankings were the ones with diversification depth of 3 to 4 times the amount of images shown to the user, meaning that the algorithm reranked 60 to 80 images, achieving equivalent or superior results than the NoDiv. The best results were achieved by Rankings 8 and 9. Ranking 8 achieved significant effectiveness gain in all measures. This ranking achieved relative gains of roughly 6.5% in MAP, 19% in GMAP, 6% in BPREF, 8% in R@200, and 2.4% in CR@ 200. Ranking 9 achieved gains of roughly 4% in R@200 and 5% in CR@200.

From Table 4.3, we can observe that FullDiv rankings were outperformed by the NoInitDiv ones (in R@200 and CR@200), meaning that avoiding diversification in the initial result set may allow finding more relevant items to the query in the first iteration. Therefore, this larger set of items allowed the improvement of the diversity by attracting similar items in further iterations which in turn provided more information for increasing the diversity in the novel result sets, e.g., by reaching novel items through different modalities/features. This can be corroborated by analyzing the precision-based measures for the FullDiv rankings. We can observe that the first initial diversification step using only initial query information (no user feedback yet) ended up with worse precision-based effectiveness and worse diversity scores.

Analyzing the InitDiv rankings, specifically 2 and 3, we observe better effectiveness than the FullDiv rankings. We can consider this as a consequence of the diversification in the first result set using just a shallow reranking depth which in turn had little impact in

Table 4.3: Second Round Result Summary. Parameters d and λ refer to the diversification depth and the linear combination factor.

Ranking	**Method**	**d**	λ	**MAP**	**GMAP**	**BPREF**	**R@200**	**CR@200**
1	NoDiv	-	-	0.2243	0.0950	0.2665	0.6288	0.7294
2	InitDiv	1.5	0.1	0.2255	0.0979	0.2718	0.6394	0.7324
3	InitDiv	1.5	0.99	0.2012	0.0921	0.2519	0.6385	0.7481
4	InitDiv	4	0.5	0.1820	0.0905	0.2537	0.6362	0.7448
5	FullDiv	3	0.1	0.1769	0.0672	0.2376	0.5846	0.6911
6	FullDiv	4	0.3	0.1675	0.0661	0.2443	0.5984	0.7026
7	FullDiv	4	0.7	0.0979	0.0341	0.1689	0.4921	0.6851
8	NoInitDiv	4	0.3	**0.2387**	**0.1129**	**0.2814**	**0.6789**	**0.7468**
9	NoInitDiv	4	0.5	0.2163	0.0986	0.2607	**0.6574**	**0.7679**
10	NoInitDiv	3	0.7	0.1935	0.0811	0.2377	0.6018	0.7503

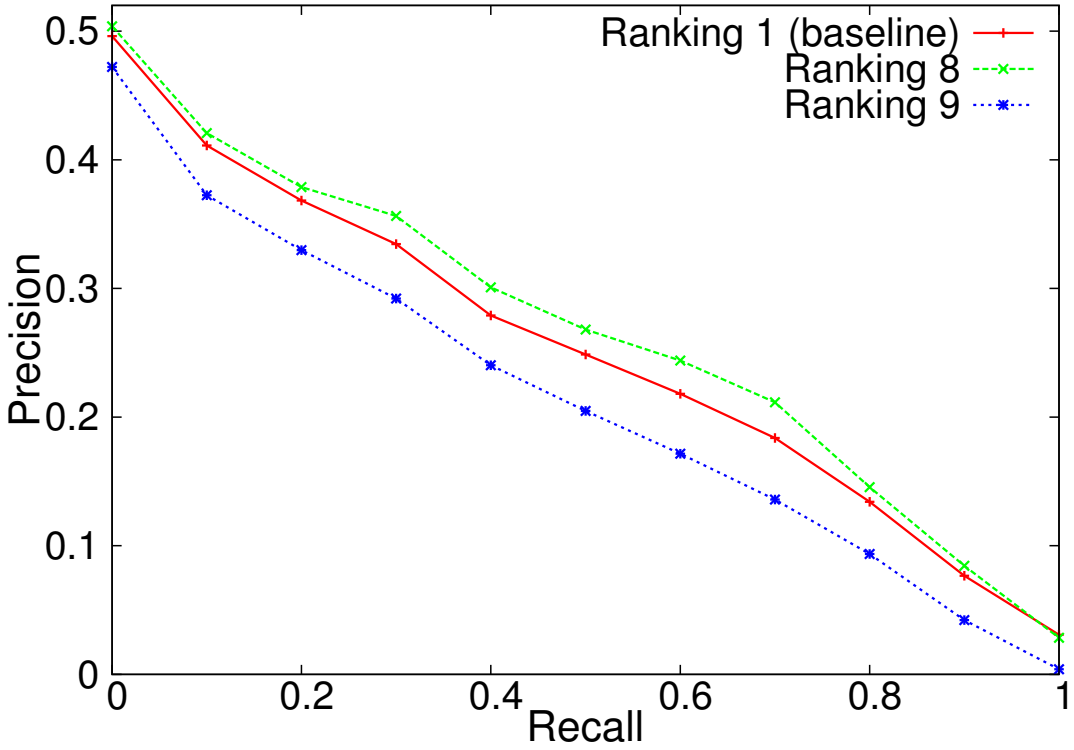

Figure 4.4: Precision x Recall Curves of the best rankings and the baseline.

overall effectiveness. Moreover these rankings ended up to be equivalent to the NoDiv with similar average measures. This can be explained by the fact that the initial diversification with only a shallow reranking had only a small negative impact on the following iterations. Furthermore, in Ranking 4, in which a deeper reranking was performed when compared to Rankings 2 and 3, the harm on effectiveness was even more significant. This corroborates the previous discussion comparing the FullDiv and NoInitDiv methods.

In the context of the same interactive retrieval method, we can observe the impact of the relevance-diversity combination factor (λ). Comparing Rankings 2 and 3 (both used the same diversification depth (1.5), we can see that the larger λ value of Ranking 3 produced higher diversity at the end of the session even using a shallow reranking depth. The same can be observed when comparing the NoInitDiv rankings 8 and 9 (both with $d = 4$) in which the bigger λ value of Ranking 9 allowed higher diversity at the end of the session with an impact on the amount of items retrieved. This situation is not true for FullDiv rankings 6 and 7, which were penalized by the initial diversification. A comprehensive analysis of the impact of d and λ on the diversity-oriented sessions is presented in Chapter 5 (Section 5.2.1).

Figure 4.4 presents precision-recall curves for the baseline (Ranking 1) and the best diversification rankings. The precision-recall curve for Ranking 8 is statistically superior to the baseline for recall values smaller than 0.9 and statistically equivalent for the remaining cutoff points. This means that diversification allowed retrieving more relevant items along the interactive session and also in earlier iterations. Beyond it, with the diversification it was possible to achieve significant improvements on overall precision-based retrieval effectiveness.

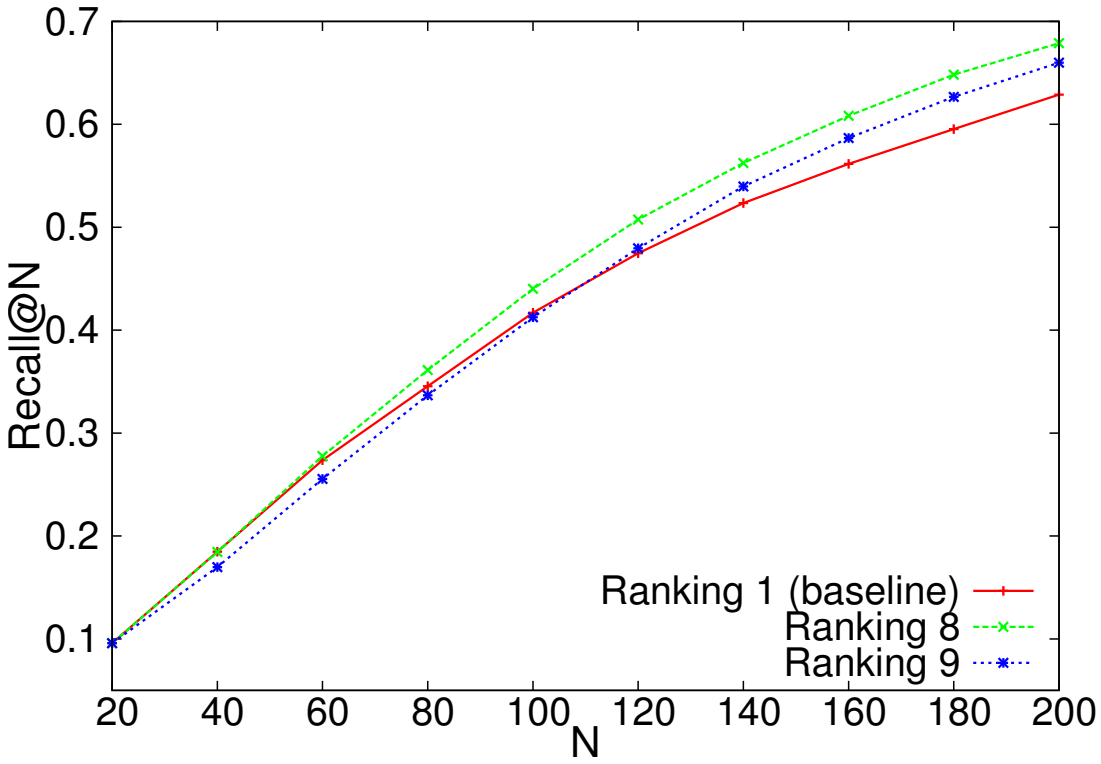

Figure 4.5: Recall Curves of the best rankings and the baseline.

Figure 4.5 presents recall evolution curves throughout feedback iterations (equivalent to Recall@N) for the baseline and the best diversification rankings. The recall curve for Ranking 8 is statistically superior to the baseline for recall values over Recall@100 and statistically equivalent for the remaining. Our assumption is that the amount of diverse information accumulated in early iterations helped approximating the query set to relevant items in the collection. Since the distance of a collection image to the query set is computed using the smallest value from the image to each image in the query set, improving the diversity and consequently generating a broader coverage of the feature space in the query helped increasing the possibility of finding new relevant items. This is a common outcome of query expansion techniques but in this case the query enrichment was boosted by the diversity.

Figure 4.6 presents diversity evolution curves throughout the feedback iterations (equivalent to CR@N) for the baseline and the best diversification rankings. The curve for Ranking 8 is statistically superior to the baseline for CR@200 and statistically equivalent for the remaining. While not statistically improving diversity for most of the iterations the diversification allowed at least achieving significant gain at the end of the session. Notice that although the diversification is not properly increasing subtopic coverage, it was able to improve the dispersion of the relevant items on the feature space. Since we used a greedy algorithm, we also have considered the possibility that, during the reranking step, relevant images on top positions in the ranking can be replaced by other relevant ones from the candidate set (possibly relevant to same subtopic) that has higher dissimilarity to the already selected images. Consequently it produces no improvement on subtopic coverage in the reranked list.

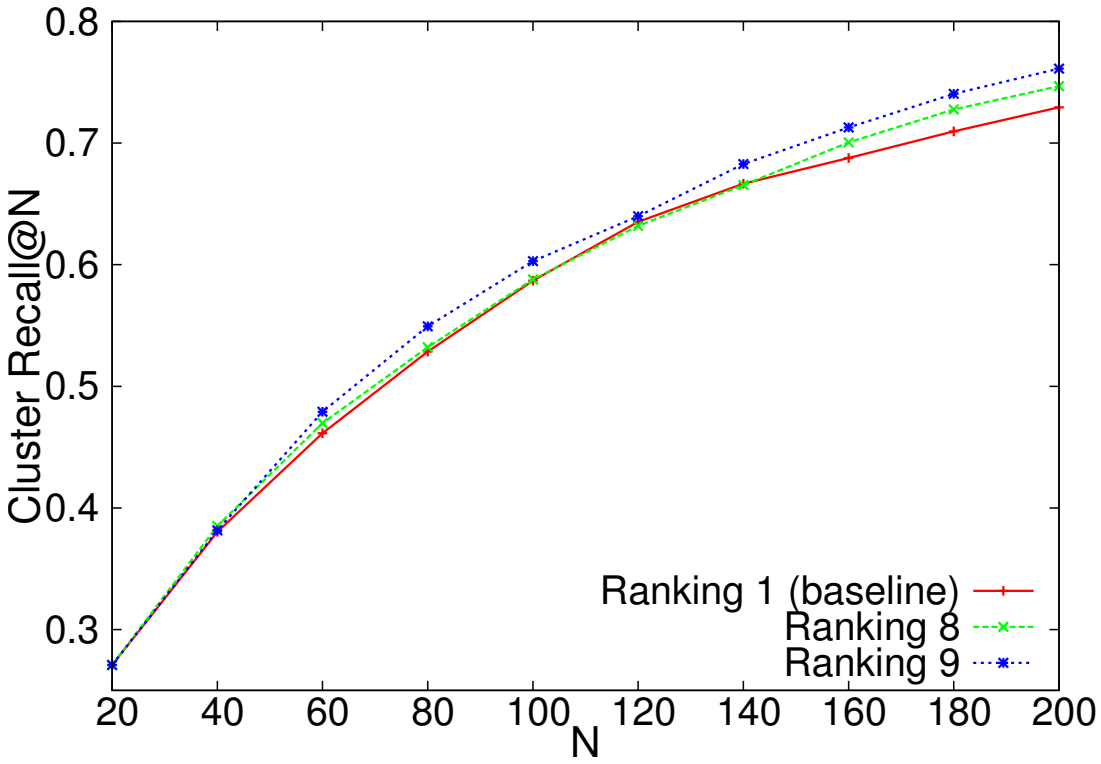

Figure 4.6: Cluster Recall Curves of the best rankings and the baseline.

Finally, we present an example of multimodal ranking functions discovered by the evolutionary process to compute the distance between two images (x and y) in Equations 4.5, 4.6, and 4.7. Notice that the GP framework was able to combine distance scores associated with both textual and visual features. These functions were randomly selected from the ones generated at the tenth iteration of Ranking 9. To highlight the multimodal characteristic of these functions, textual and visual measures are highlighted in blue and red, respectively.

$$dist(x,y) = cos(x,y) + \sqrt{qcch(x,y) + cos(x,y)} \tag{4.5}$$

$$dist(x,y) = cos(x,y) + las(x,y) + \sqrt{bic(x,y)} + (jac(x,y) \times dice(x,y)) \tag{4.6}$$

$$dist(x,y) = \frac{bm25(x,y)^2 \times \sqrt{jaccard(x,y)}}{jaccard(x,y)} \times \sqrt{las(x,y) + bm25(x,y) + (bow(x,y)^2 \times (jac(x,y) + jaccard(x,y)))} \tag{4.7}$$

4.4 Summary and Considerations

The demand for result diversification may arise in several different scenarios such as tackling underspecified or ambiguous queries, avoiding redundancy in ranked lists, answering naturally diverse information needs, improving interactive learning and even reducing the semantic gap. Several research efforts have shown that addressing these issues may improve the overall user search experience. In fact, some studies highlight the positive impact of diversification methods in the overall retrieval effectiveness. However, improving search diversity without significant relevance loss is not an easy task and is still an open research problem.

In this chapter, we revisited the research question and our investigation about the possibility of improving diversity with small or even no impact on relevant. Different from Chapter 3, we conducted the experiments in an interactive search scenario with explicit user feedback. We have studied the impact of diversification over an interactive multimodal image retrieval method with explicit user relevance feedback. Our experiments demonstrated the possibility of improving diversity over search iterations while keeping or even improving the amount of relevant items retrieved.

Another research question we investigated regards the possibility of improving the interactive learning with diversity-driven feedback. Hence, ou evaluation have also shown that more relevant items were retrieved in earlier iterations (higher recall) when the diversification was applied, overall resulting in an interesting approach for improving image retrieval methods.

The difficulty in balancing relevance and diversity has been reported in several works [27, 45, 58], since trying to improve one of them usually harms the other. Introducing an implicit diversity promotion reranking step on the framework pipeline improved the effectiveness of the baseline method. Consequently, our experiments produced important outcomes by statistically improving diversity while preserving or even increasing overall relevance.

Considering the experimental results and findings from this chapter, Chapter 5 significantly extends our investigation on interactive learning with diversity. Therefore, we consider new research questions over the impact of the diversification on the learning process and the relationship between diversity promotion and the different retrieval modalities.

Chapter 5

Diversity-based Interactive Learning meets Multimodality

Chapter 4 has experimentally analyzed the impact on the diversity promotion method in an interactive learning-to-rank context. However, as observed in Chapter 3 and several previous works [12, 30], query processing using different retrieval modalities may result in different retrieval effectiveness.

While the experiments described in Chapter 4 have shown the benefits of capturing feedback information on diverse results and its impact on the interactive learning procedure, it examined only a single multimodal environment, simulating a textual query followed by multimodal learning-to-rank from user feedback. Therefore, in Chapter 5, we go further on the analysis and investigate new research questions.

In this context, this chapter presents a comprehensive analysis of several multimodal image retrieval approaches built over a online learning-to-rank framework for relevance feedback on diversified results. We refer to online learning-to-rank as including no off-line training procedure. In fact, the learning technique is executed between the iterations of a search session, when the training (feedback) information is actually captured.

Our experimental analysis shows that different retrieval modalities are positively impacted by diversity but achieve best retrieval effectiveness with diversification applied at different moments of a search session. Moreover, the best results are achieved with a query-by-example approach using multimodal information obtained from feedback. In summary, we demonstrate that learning with diversity is an effective alternative for boosting multimodal interactive learning approaches.

The remainder of this chapter is organized as follows. Section 5.1 formally introduces our research questions. Section 5.2 presents our experimental results,[1] analysis, and discussions. Finally, Section 5.3 presents a summary and final considerations.

[1]The experimental setup for these new experiments was the same as described in Chapter 4.

5.1 Proposed Analysis

In Chapters 3 and 4, the main research questions were related to the relevance-diversity trade-off and whether diverse results could improve interactive learning. In these previous analyses, we verified that:

- Multimodal data integration contributes for optimizing the relevance-diversity trade-off;
- It is possible to improve diversity with small or even no impact on relevance; and
- It is possible to boost interactive learning with diversity.

The experiments demonstrated that learning with diverse items allowed enhancing diversity in a search session while simultaneously retrieving more relevant items in earlier feedback iterations. In this chapter, we investigate, in a much more detailed way, issues regarding the combination of interactive multimodal retrieval with diversity and present a more comprehensive experimental analysis. More specifically, we investigate new research questions including:

- How sensitive is the learning method to the diversification parameters?
- Are there any recurrent patterns regarding the effectiveness impact?
- How the diversification approaches behave with different retrieval modalities and vice versa?
- What is the effect of diversification on machine learning and fusion approaches for multimodal relevance feedback?

Furthermore, we analyze the stability of the process considering different diversification approaches and several evaluation measures (see Section 5.2.5). Finally, we also analyze the overall best results for the combination of all studied diversification approaches and retrieval modalities (see Section 5.2.6).

For this new experimental study, we analyze the different querying scenarios shown in Table 5.1. In the table, *mm* means multimodal information (textual and visual), *txt* means textual similarity only, and *vis* means visual similarity only. For instance, the mm-vis retrieval modality is a multimodal session with a visual example as query. Therefore, it indicates that the initial set of objects is built only according to the visual similarity between the query image and the images from the collection, thus simulating a query-by-example beginning followed by a relevance feedback process which uses both textual and visual information for learning.

Table 5.1: Retrieval Modalities.

Approach	**Relevance Feedback**	**Initial Set**
Multimodal Retrieval (mm-mm)	mm	mm (txt + vis)
Multimodal Retrieval with textual initiation (mm-txt)	mm	txt
Multimodal Retrieval with visual initiation (mm-vis)	mm	vis
Textual Retrieval (txt)	txt	txt
Visual Retrieval (vis)	vis	vis

5.2 Results and Discussion

In order to address the research questions presented in Section 5.1, we conducted a set of experiments described and discussed in Sections 5.2.1, 5.2.2, 5.2.3, and 5.2.4. Section 5.2.5 presents a study case using the best retrieval modality chosen according to the experiments. Section 5.2.6, in turn, presents the discussion of the best combinations of retrieval modalities and diversification methods. As in Chapter 4, statistical significance was computed using Wilcoxon's test with p-value < 0.05.

5.2.1 Diversity Method Parameters Impact

The proposed diversification method uses two main parameters: the reranking depth (d) and the diversity importance factor (λ). For understanding the impact of these parameters in the overall session effectiveness, we explore several combinations of them (see Section 4.2.2) and assess their effectiveness using different measures.

Figure 5.1 presents the results for the different retrieval modalities grouped by each evaluation measure and using heat maps. It depicts the effectiveness behavior for MAP, Recall@200, CR@200, and F1@200, respectively. F1@200 is taken as the harmonic mean between Recall@200 and CR@200. Notice that the goal is not to directly compare different modality maps using absolute effectiveness values. Thus, we use appropriate scales for each map in order to highlight the system behavior considering the diversity factor and reranking depth effects for each modality.

From Figure 5.1, we can notice a generally similar behavior for all retrieval modalities and evaluation measures which shows that the parameter combination produced the same effectiveness behavior despite the retrieval modality. For all measures the impact is similar for all modalities with a small variation among the modalities only in terms of diversity which did not significantly affect F1 values.

We can also see that using very high values for the diversity factor or a too deep reranking harms the relevance of the results (MAP and Recall@N). At the same time too deep reranking also damaged the final diversity, which may be a consequence of the precision loss and consequent decrease on cluster representation. Therefore, optimizing the diversity of the results depends on a proper balance of the reranking depth and diversity importance factor no matter which retrieval modality is used.

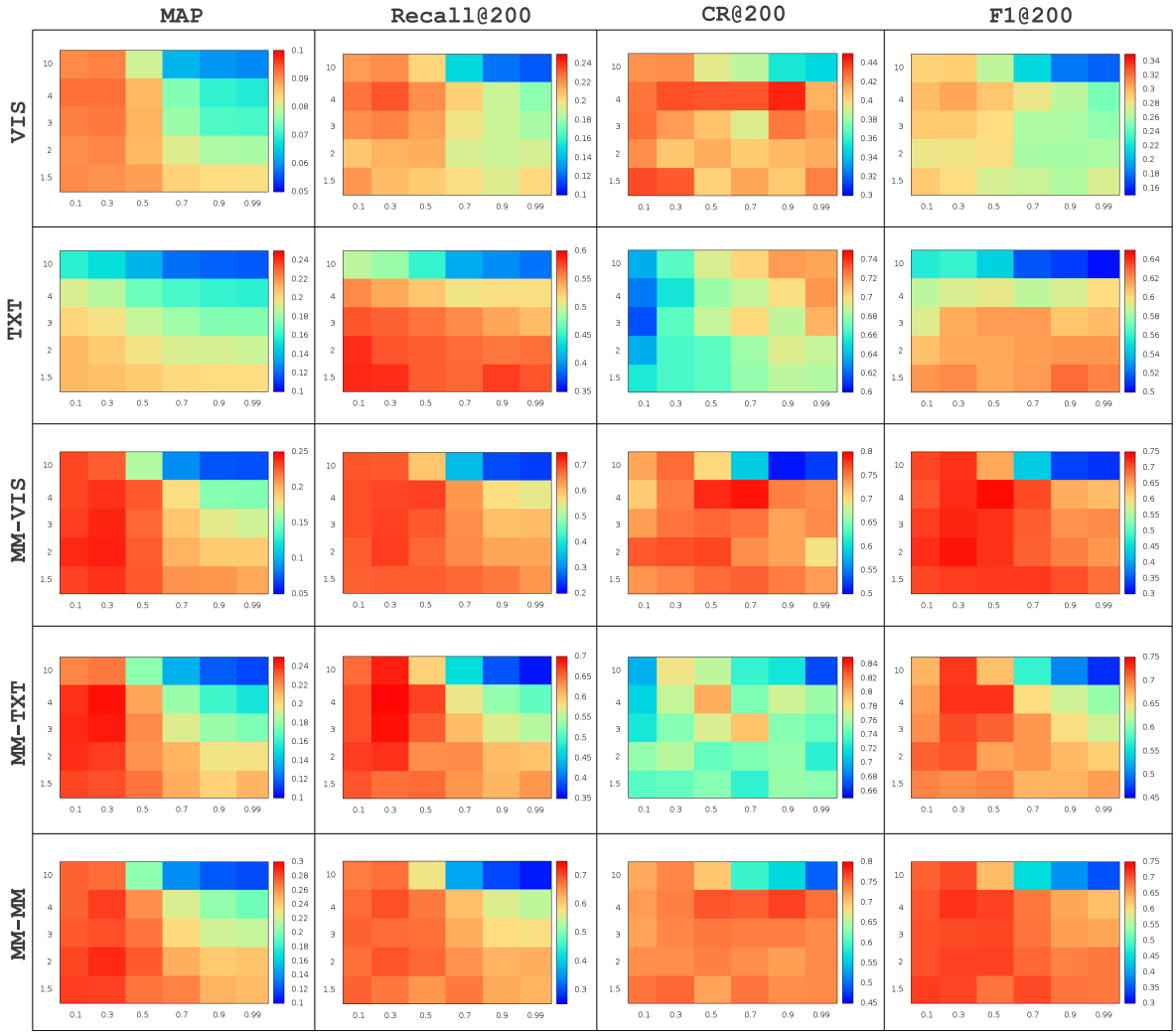

Figure 5.1: Impact of different combinations of the diversification parameters. In the maps, the X axis corresponds to the diversity factor, the Y axis corresponds to the reranking depth, and the graded bar on the right depicts the effectiveness scale.

In general, for all modalities the experiments have shown that in terms of Recall@N the best results were obtained with medium to deep reranking ($2 \leq d \leq 4$) and small diversity factors ($0.1 \leq \lambda \leq 0.5$). The exception was the textual (*txt*) approach that demanded only shallow reranking ($d = 1.5$) for higher recall values, which is even cheaper in terms of computational costs. Considering diversity effectiveness, the best results were obtained with deep reranking ($3 \leq d \leq 4$) and great diversity emphasis ($0.5 \leq \lambda \leq 0.99$).

As expected, the best relevance-based results were obtained by the diversification methods with just small diversity promotion, which was enough for allowing the construction of more diverse result sets while keeping or improving the overall effectiveness. In sum, we can say that promoting diversity, at some extent, may improve not only the overall diversity of the evaluated methods, but also their relevance-based effectiveness.

Moreover, disregarding relevance, the diversity was also significantly improved by using higher diversity factor which again also allowed keeping or improving the overall effectiveness (F1). Nevertheless, there is something else we have to consider: retrieving more relevant items may or may not trigger diversity improvement since additional relevant items from the same cluster do not change diversity values.

5.2.2 Per-Modality Diversity Analysis

For each retrieval modality, we compared the different diversification approaches. From Figure 5.2, we can see that the FullDiv and NoInitDiv achieved *slightly* superior effectiveness for F1@200. As expected, the text-only approach outperformed the visual-only approach given the high visual heterogeneity of the dataset images and the conceptual richness of the queries. We can also notice that all multimodal approaches allowed statistically superior effectiveness when compared to the single-modality ones (VIS and TXT). These results suggest that using diversification approaches throughout the retrieval session (FullDiv and NoInitDiv) may allow a better search experience regarding effectiveness. Furthermore, the higher F1 values achieved by these methods suggest that the effectiveness can be improved by properly combining relevance and diversity.

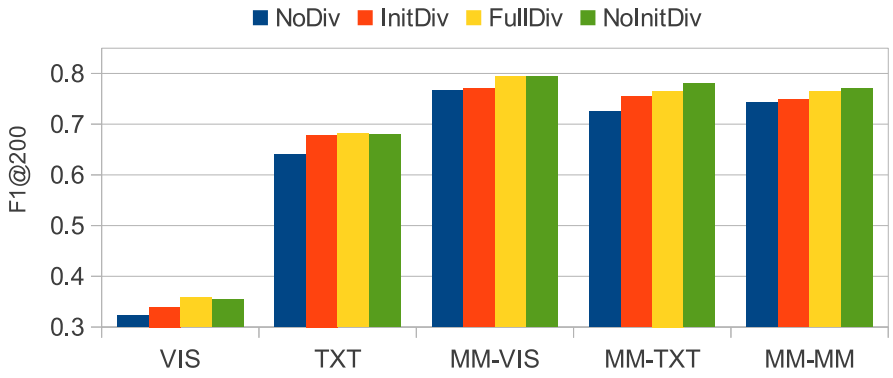

Figure 5.2: Best diversity approach for each retrieval modality.

5.2.3 Modality Analysis for Diversification Approaches

For each diversification approach, Figure 5.3 presents the best results for each retrieval modality. Comparing the different modalities, the MM-VIS produced statistically superior effectiveness for Recall@200, CR@200, and F1@200 when compared to the single modality methods regardless of the diversification approach. Similarly, the MM-VIS approach achieved equivalent or superior effectiveness considering the other multimodal approaches. For MAP, the MM-MM retrieval modality achieved superior average effectiveness which may be a consequence of using all the modalities from the beginning of the session, which allowed retrieving more relevant items in the very first result set as also observed in Figure 5.10. Nevertheless, the multimodal approaches were considered statistically equivalent.

From a different point-of-view, when comparing the best diversification results to the no diversified sessions, these results have also shown that the greatest relative gains[2] were achieved on visual-only sessions (the worst effective modality). These results highlight the fact that introducing diversity may significantly enhance retrieval results even for low performance features. On the other hand, by analyzing absolute gains, the greatest improvements were achieved by the MM-TXT approach,[3] in this case boosted by learning

[2]In terms of Recall (16.2%), Cluster Recall (13%), and F1 (13.2%).
[3]In terms of Recall (6.2%), Cluster Recall (8.3%), and F1 (5,6%).

with multimodal data. Additionally, as previously reported, in terms of MAP, the best performing approach was the MM-MM, which also achieved superior relative and absolute gains for that measure.

Finally, Figure 5.3 shows the superiority of the multimodal approaches, which suggest that, when diversity is also considered, using multiple sources of information is an effective way for improving retrieval effectiveness as observed with several evaluation measures.

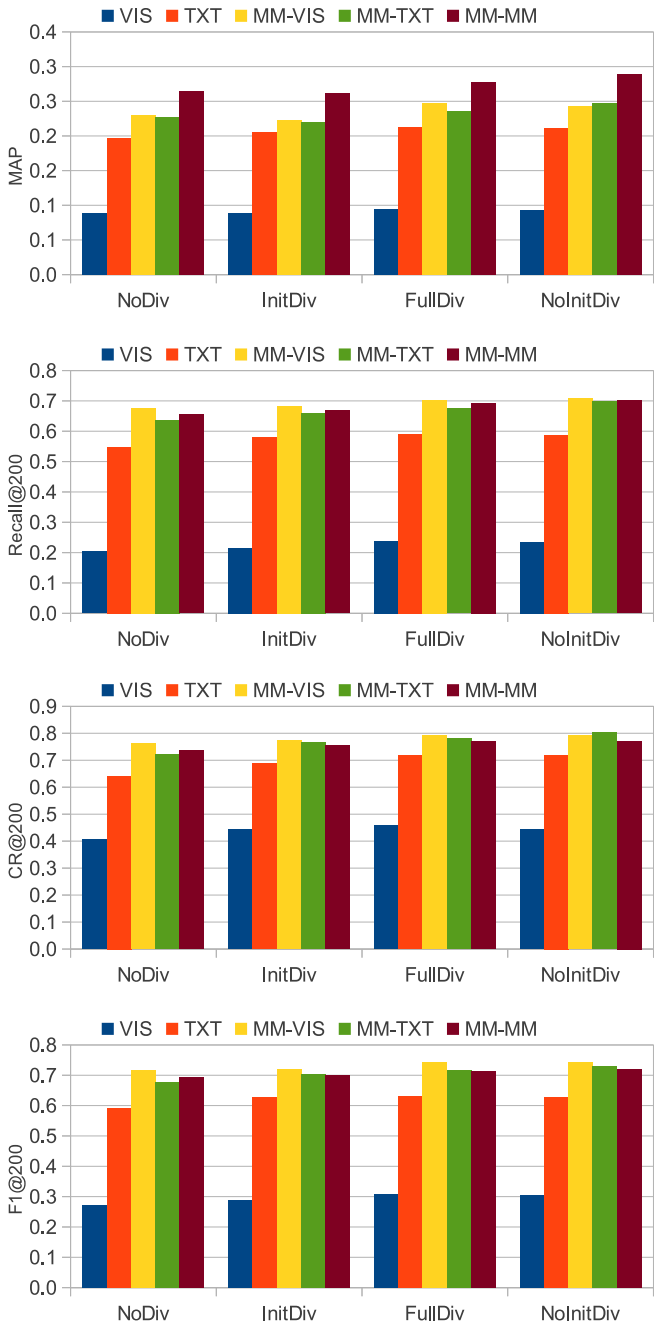

Figure 5.3: Best results per evaluation measure.

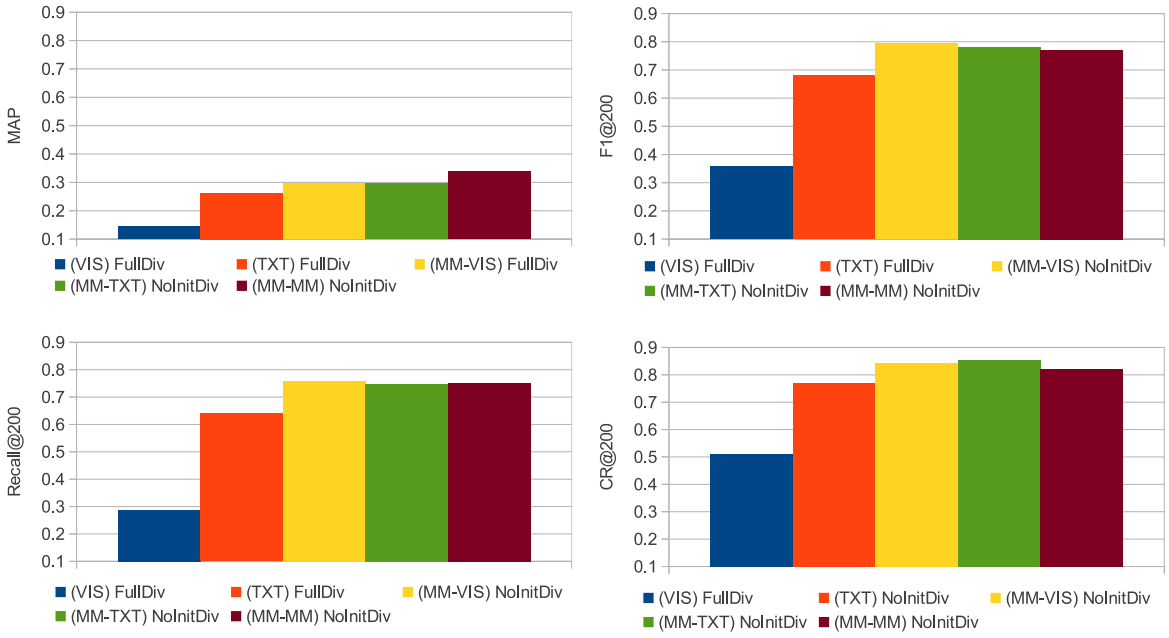

Figure 5.4: Overall best results.

5.2.4 Overall Effectiveness Analysis

Figure 5.4 presents a comparison of the different retrieval modalities in terms of their best diversification approaches. We can see that for all measures the text-only approach outperforms the visual-only approach and the multimodal retrieval allows superior effectiveness on all measures. The only exception was MM-VIS, which was considered statistically equivalent to TXT in terms of diversity.

From Figure 5.4, we can see that in general the single modality approaches were most successful with the FullDiv method. The only exception was the text-only retrieval in terms of diversity that achieved the best results with the NoInitDiv method. On the other hand, the multimodal approach achieved the best results with the NoInitDiv method and were also superior to the single modality retrieval. For the multimodal approaches, the only exception was the MM-VIS that achieved superior MAP with the FullDiv.

These results suggest that when using single modality retrieval approaches, the methods are less harmed by the initial diversification. On the other hand, it suggests that using multimodal information reduces the problem of retrieving non-relevant items in the first iteration which in turn helps the system to produce better results by using more information from the relevant items obtained without diversification.

Additionally, as we can see from Figures 5.2 and 5.4, the TXT approach has achieved better effectiveness than the VIS one. It is important to mention that the visual heterogeneity of the images in the collection and the fact that image annotations are composed of detailed textual descriptions harden the task when only visual information is used. However, in many cases, textual information is not available along with the images. Even in these cases, our proposed approach can be used with satisfactory results, as we have seen.

5.2.5 Comparison with the Baselines

As observed in Section 5.2.3, the best results for the proposed diversification approaches were achieved using the MM-VIS retrieval modality. Therefore we have chosen this modality for further analysis and comparison to the different proposed methods and baselines.

Comparison with fusion-based methods

It is important to understand the behavior of the diversification approaches throughout the retrieval session. Figures 5.5 and 5.6 present the effectiveness evaluation at each relevance feedback iteration in terms of relevance (Recall@N) and diversity (CR@N).

Figure 5.5 presents the evolution throughout the retrieval session in terms of the amount of relevant items retrieved in relation to the number of relevant items that actually exist in the collection. It is clear that the GP-based learning approaches outperform the fusion-based ones. All learning approaches were statistically superior to BordaMMR for $N \geq 60$ and superior to MinAvgMMR for $N \geq 100$.

Figure 5.6 presents the evolution throughout the session in terms of diversity. All learning approaches were statistically superior to BordaMMR for $N \geq 80$. In this case, the best learning approach (NoInitDiv) was superior to MinAvgMMR for $N \geq 100$.

Figure 5.7 presents the Precision@Recall curves for the different retrieval approaches in comparison to the fusion-based baseline methods. All learning methods were superior to BordaMMR for $0.1 \leq Recall \leq 0.8$. In relation to MinAvgMMR, the InitDiv method was in general statistically equivalent while the FullDiv and NoInitDiv methods achieved statistically superior effectiveness for $0.1 \leq Recall \leq 0.8$.

As we can see, the fusion-based methods were statistically outperformed by the GP approach that produced a more suitable combination of the different features, consequently making the diversity-based feedback more useful.

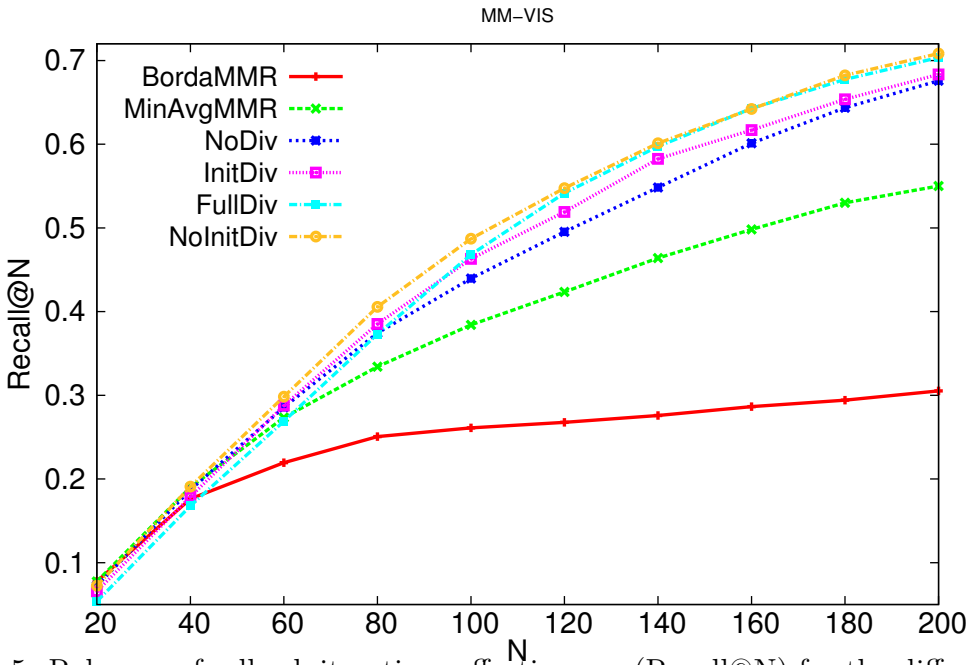

Figure 5.5: Relevance feedback iterations effectiveness (Recall@N) for the different diversification approaches.

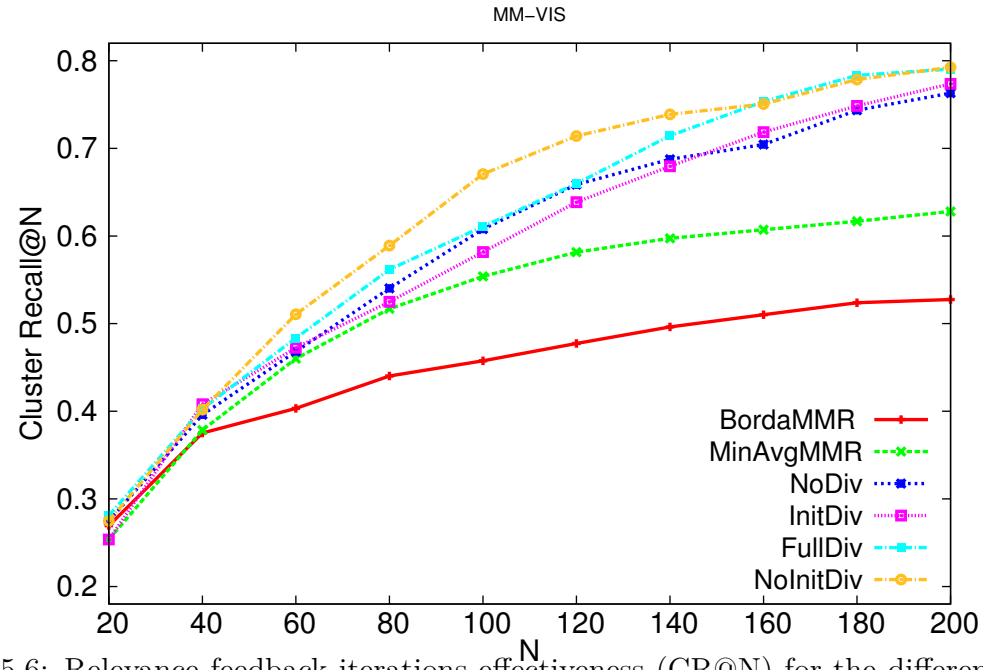

Figure 5.6: Relevance feedback iterations effectiveness (CR@N) for the different diversification approaches.

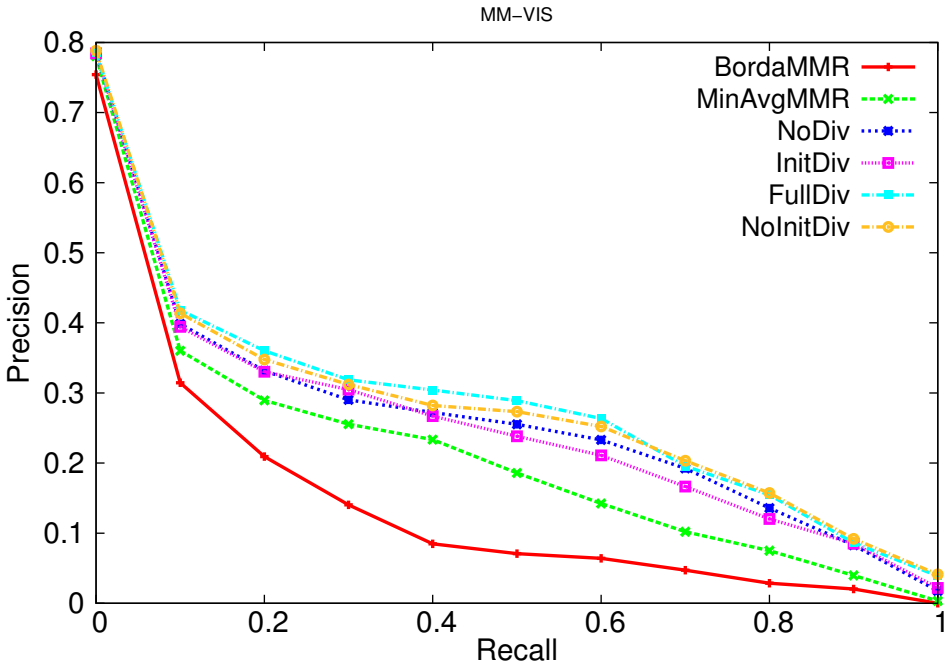

Figure 5.7: Relevance feedback iterations effectiveness (Precision x Recall) for the different diversification approaches.

Stability Analysis

For estimating the stability of the proposed diversification approaches, we conducted extra experiments using ten different random seeds for the genetic programming framework and computed confidence intervals for $\alpha < 0.05$. Figure 5.8 presents the results for three relevance-based measures: MAP, GMAP, and BPREF. As we can see, the NoInitDiv approach produced statistically superior effectiveness for all reported measures.

Figure 5.9 presents the comparison in terms of relevance (Recall@200), diversity (CR@200), and their combination (F1@200). The FullDiv and NoInitDiv methods achieved statistically superior effectiveness when compared to the no diversification baseline (NoDiv). Notice that FullDiv and NoInitDiv alternate positions for Recall@200 and CR@200 but are statistically equivalent in terms of F1@200.

Here we can notice that the diversification approaches performed consistently better in the executions with the different random seeds. This highlights the fact that the reported effectiveness was not affected by the randomness intrinsic to the GP method.

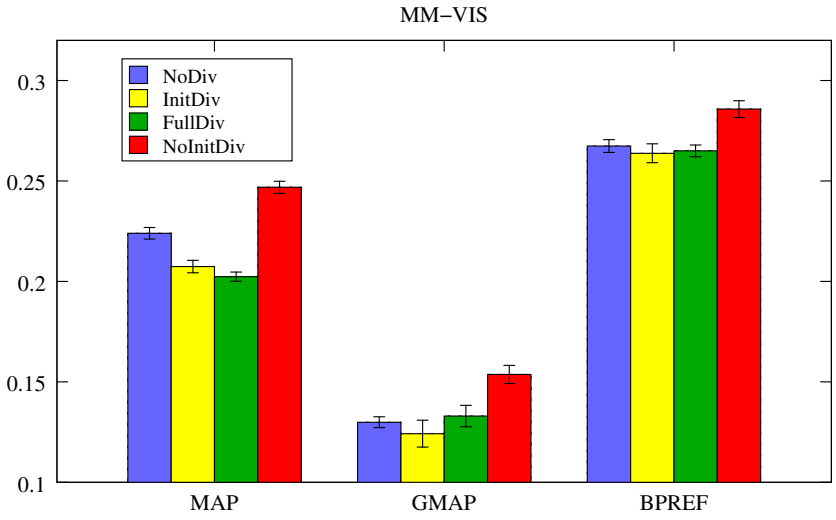

Figure 5.8: Random seed variation results for MM-VIS: Precision-based measures.

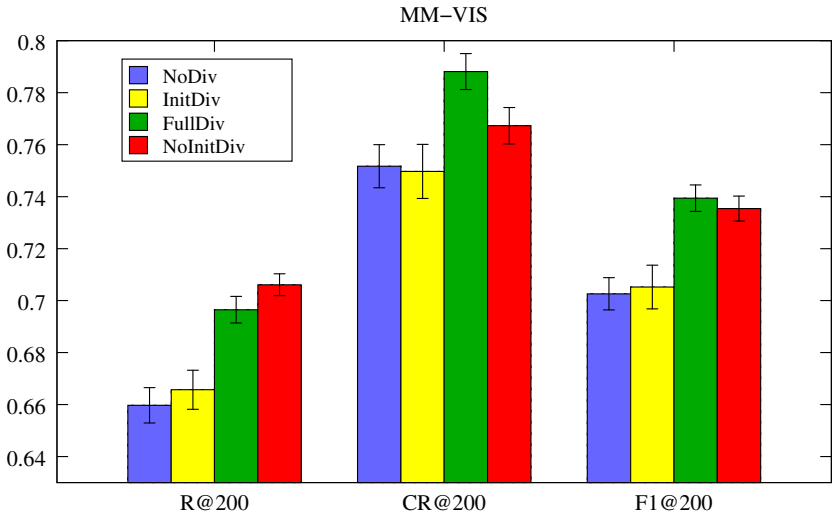

Figure 5.9: Random seed variation results for MM-VIS: Recall x Diversity.

5.2.6 Session Effectiveness Analysis for the Best Alternatives

Best Results per Retrieval Modality

In this section, we analyze and compare the overall best results for all modalities and diversification approaches with ours and the baseline approaches. As before, we present effectiveness results throughout the relevance feedback session. Figures 5.10 and 5.11 present the best combination of retrieval modality and diversification methods in terms of the relevance of the retrieved items.

From Figures 5.10 and 5.11, we can notice the best results were achieved by the MM-MM (NoInitDiv) method, which is a completely multimodal session with diversification in all iterations but the first.

Similarly, Figure 5.12 presents the best diversity combination. We can notice the superiority of the multimodal approaches compared to the single modality ones. The multimodal approaches alternate positions at different moments of the session.

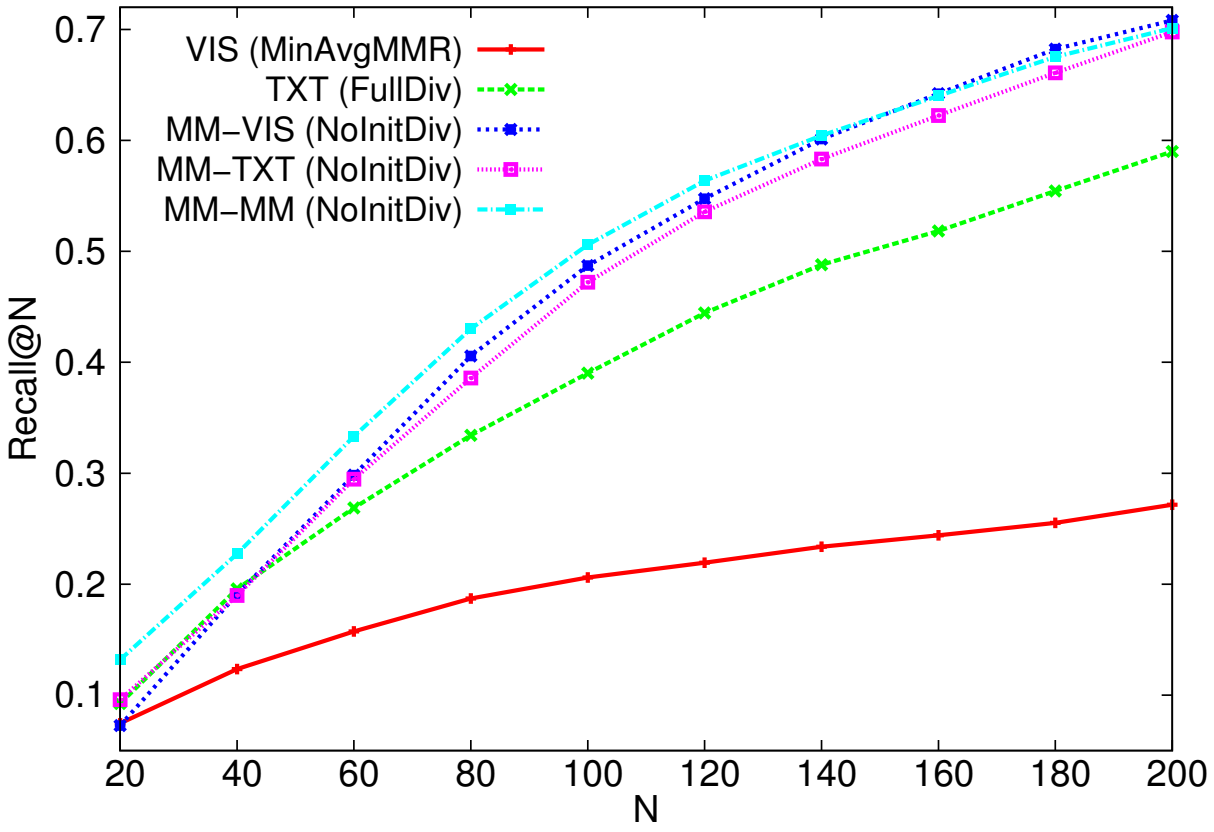

Figure 5.10: Best Recall results for the different retrieval modalities.

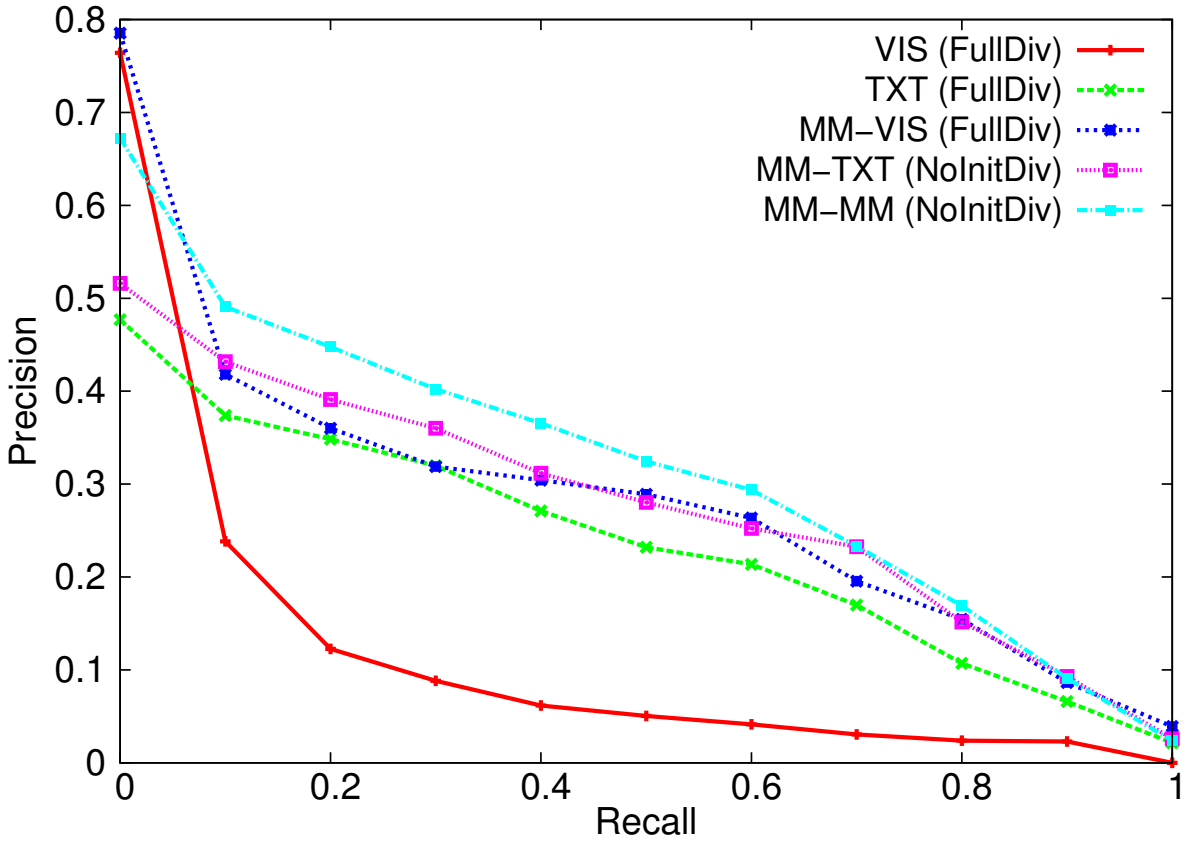

Figure 5.11: Best Precision x Recall results for the different retrieval modalities.

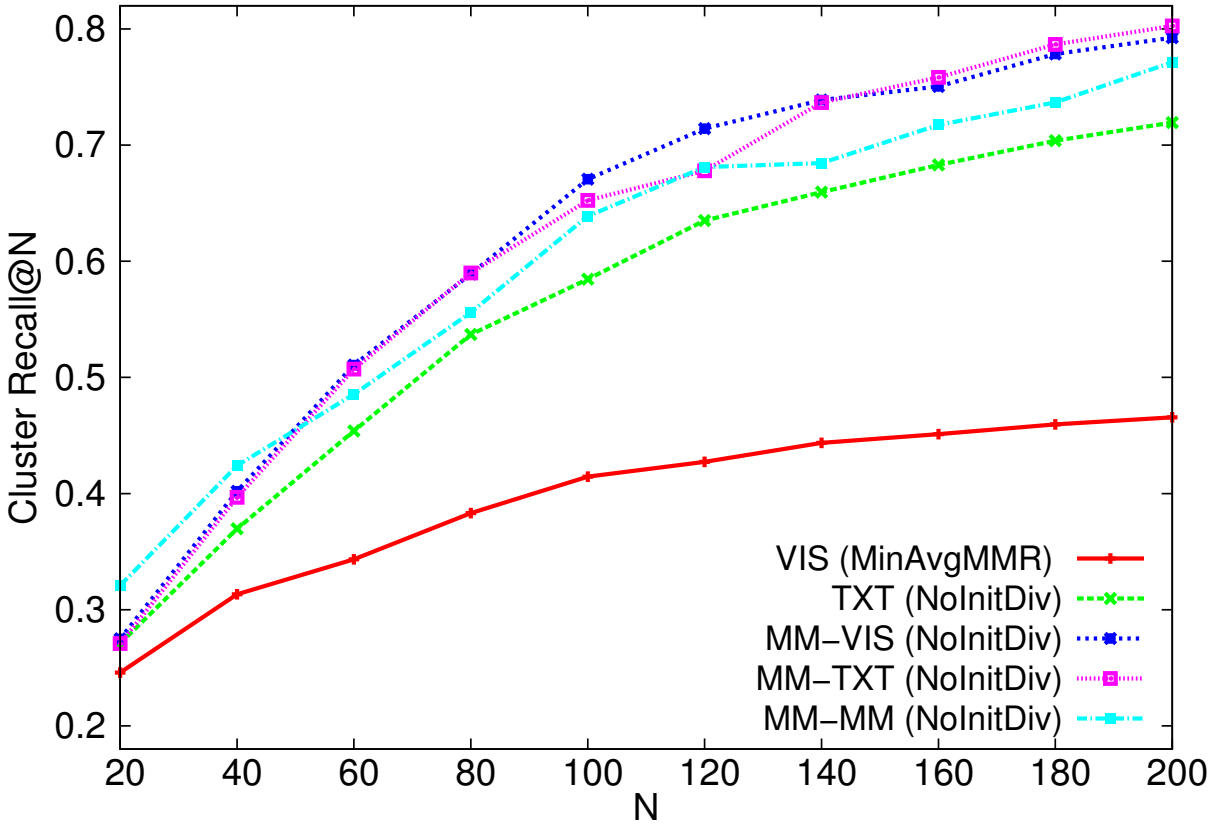

Figure 5.12: Best diversity results for the different retrieval modalities.

Best Results per Diversification Method

Considering the Recall effectiveness throughout the retrieval sessions (Figure 5.13), although the rank aggregation methods achieved the best results with MM-MM (BordaMMR) and MM-TXT (MinAvgMMR) all GP-based methods were more effective with the MM-VIS approach. We consider this as a consequence of the visual assessment of the initial results by the user which helped improving the GP individuals by detecting visually similar but non-relevant images and boosting the results on the following iterations that also used the textual information.

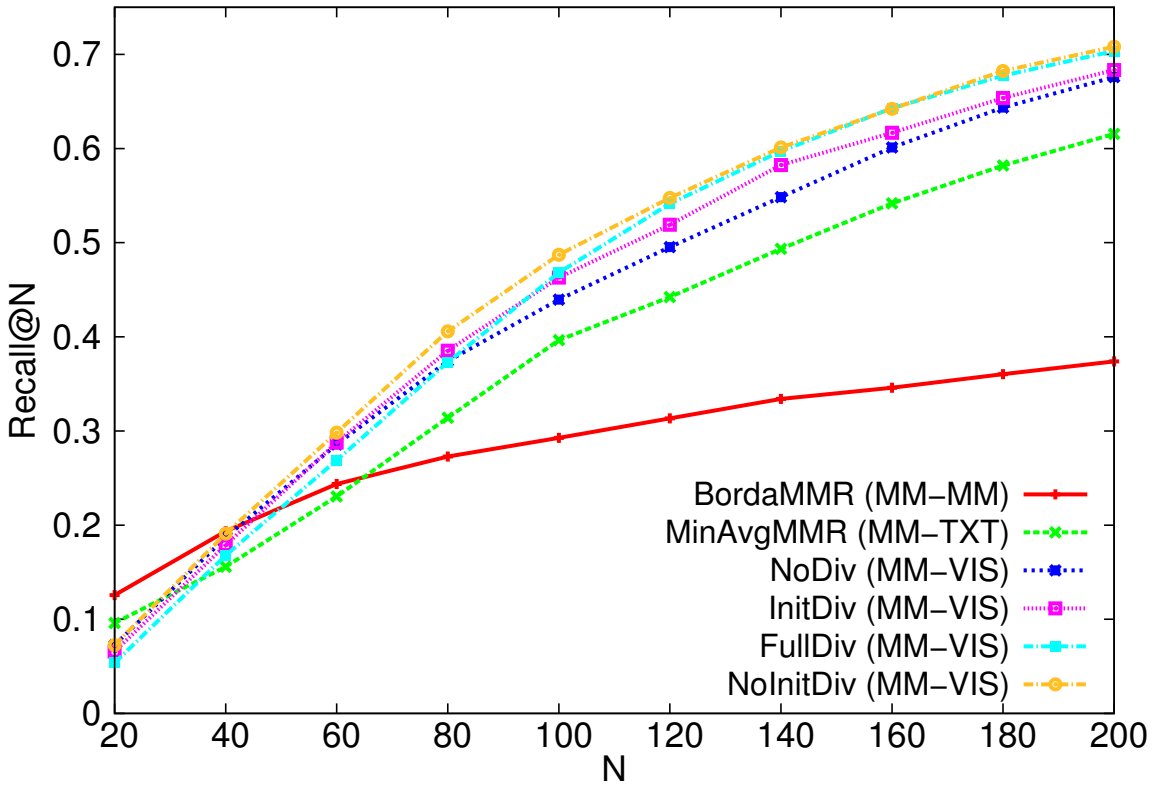

Figure 5.13: Best Recall@N curves for all evaluated methods.

Regarding diversity (Figure 5.14), both rank aggregation methods were more effective with MM-TXT. Similar to the recall behavior, almost all GP-based methods achieved better results with MM-VIS. The exception was the NoInitDiv method that was more effective with the MM-TXT approach. Here we can notice again that the diversification procedure is more harmful to the visual approach than the textual one which may be directly related to the relevance-based effectiveness of these modalities. In other words, the less effective a method is in terms of relevance measures, the higher is the harm caused by the diversification procedure on relevance.

Finally, as we can see from Figure 5.15, regarding the Precision@Recall evaluation (using MAP values as comparison measure) all the rank aggregation and GP-based methods achieved the highest effectiveness with the completely multimodal approach (MM-MM). This may be understood by the multimodal combination power on separating relevant and non-relevant images which helped retrieving more relevant items in top rank positions.

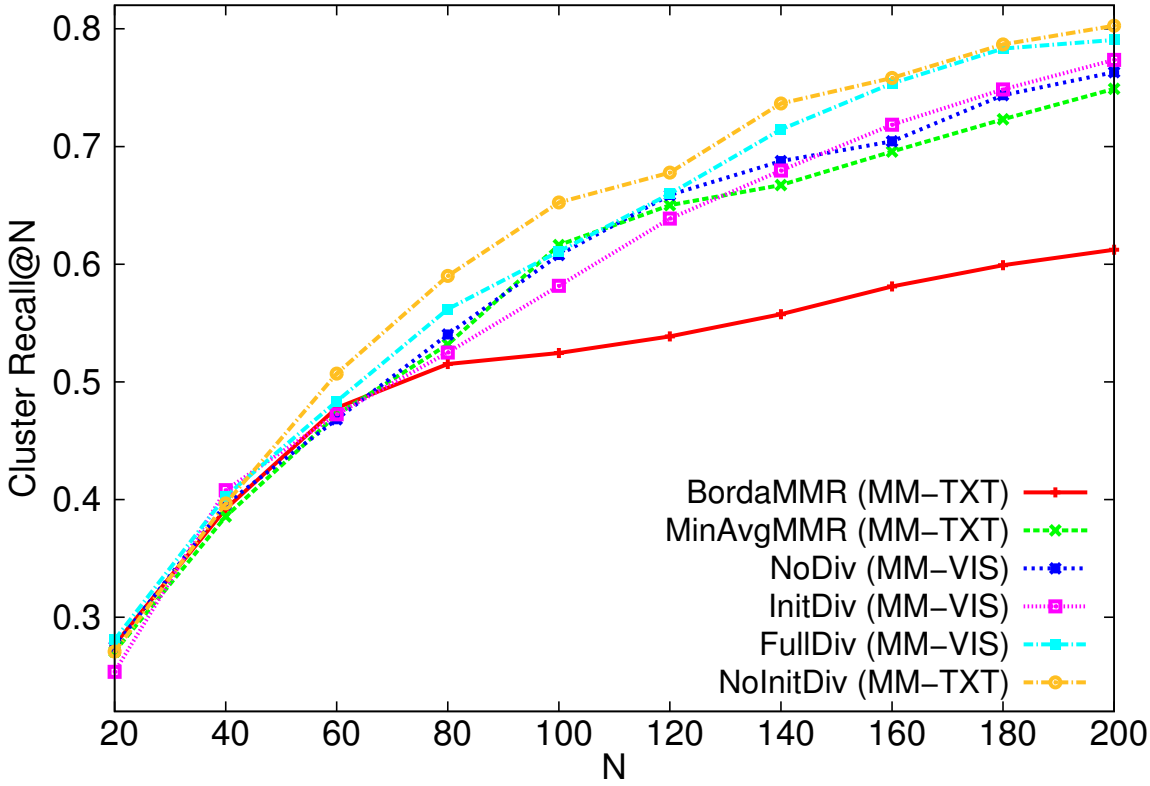

Figure 5.14: Best CR@N curves for all evaluated methods.

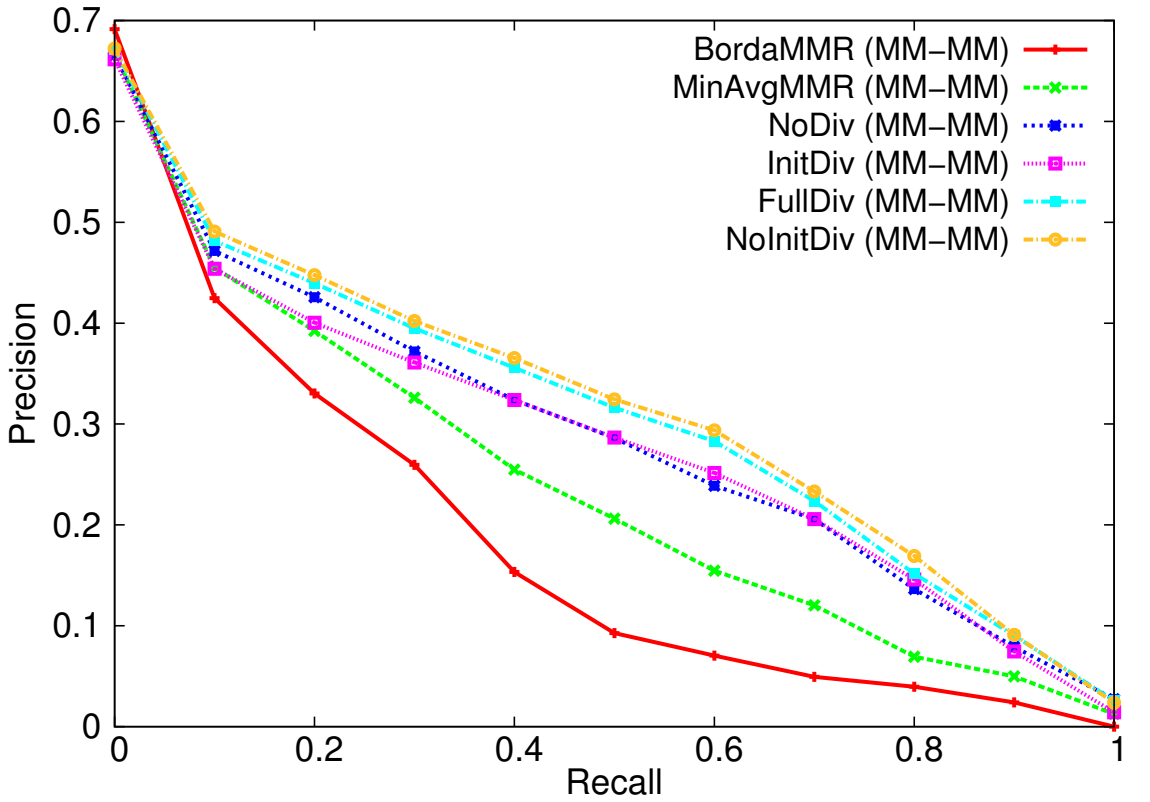

Figure 5.15: Best Precision@Recall curves for all evaluated methods.

5.3 Summary and Considerations

In this chapter, we have analyzed the behavior of a diversity-based learning approach with regard to different retrieval modalities for relevance feedback sessions.

Our initial research question was related to how sensitive is the learning method to the diversification parameters and if there is any recurrent patterns regarding its impact on effectiveness. Our experiments revealed that the diversity-oriented sessions produced similar behavior for all retrieval modalities and similar impact for all effectiveness measures considered.[4] Moreover, considering relevance-based measures, such as MAP and recall, we have seen that applying to much emphasis on diversity or a too deep reranking ended up with worse effectiveness. Beyond it, a too deep reranking also produced inferior effectiveness in terms of diversity. In summary, we have seen that promoting diversity, at some extent, allowed significant improvements on overall relevance and diversity.

In our experimental analysis, we also intended to understand how diversification approaches behave with different retrieval modalities and vice versa. Therefore, we conducted a broad investigation on the use of different diversification alternatives and how they affect and are affected by different retrieval modalities. Our experiments demonstrated that, independently of the retrieval modality, a search session may be significantly impacted by the diversification approach as the final effectiveness is quite sensitive to level of diversity promotion. Nevertheless, using relevance feedback over diversified results was useful for boosting the retrieval effectiveness for the different modalities. We have also seen that, for all diversification strategies studied, the multimodal approaches outperformed the single-modality ones. At the same time, diversity promotion had a higher impact on the multimodal sessions.

The case study using the query-by-example simulation (MM-VIS) has also shown that the retrieval session quality also depends on the retrieval modality as the results presented here outperformed the ones from the baseline work (MM-TXT) from Chapter 4.

In a slightly different direction, we also evaluated the effect of diversification on machine learning and fusion approaches for multimodal relevance feedback. Our results have demonstrated consistent and statistically superior effectiveness of the GP-based learning method when compared to fusion-based methods. In fact, the learning procedure was boosted by the feedback over diverse items and allowed the construction of optimized ranking functions, which more properly integrated the multiple features.

Finally, we also assessed the stability of the diversity promotion methods when compared to the retrieval sessions without diversification. Our findings have shown that the diverse-oriented sessions performed consistently better than the session without diversity. In summary, it shows that the superior performance was not affected by the randomness of the GP algorithm.

[4] A slight variation was detected in relation to diversity measurements, which in fact had no impact on overall performance measured with F1.

Chapter 6

Conclusions

The advances and wide availability of resources for data production, storage, and processing allowed the construction of large data repositories and consequently pushed the IR technologies to evolve towards efficient and effective search engines. By studying several inherent issues and trying to understand the user-information relationship, the IR research community has faced great challenges on fulfilling complex user needs. While great advances have been accomplished so far, as quite often is the case, the further a science field is developed, the greater are the challenges for the next generation. That is precisely the scenario of the IR field.

As highlighted in Chapter 2, IIR has emerged as an alternative for maximizing and/or speeding up the transferring of information from the user to the search engine. These interactive systems were built by the integration of state-of-the-art solutions from several associated disciplines. Nevertheless, with more and more information available from the data and user contexts, building learning models for handling the variety of evidence from the data collections and the myriad of user search tasks (and aspects) is still a hard problem. Therefore, developing more advanced retrieval strategies may be achieved by jointly exploiting the advances from several complementary fields.

As a broad review of several IIR related subjects, Chapter 2 integrated historical and state-of-the-art literature and may introduce an interested reader to the field while providing insights and directions to new research activities.

Integrating different sources of evidence into multimodal retrieval strategies has been theoretically and experimentally shown as a successful approach for effectiveness enhancement. Moreover, several boosting approaches have been proposed for improving interactive learning methods. Considering the processing of ambiguous or underspecified queries, the diversity promotion methods have attracted great attention as a method for, e.g., attenuating redundancy in retrieval results or maximizing the satisfaction for as many search aspects as possible.

Under these assumptions, we hypothesized that since combining multimodal information is effective on enhancing retrieval effectiveness, it should also hold when diversity is considered.

For verifying such statement, we conducted an experimental evaluation guided by the following research questions:

- Does multimodal data integration contribute for optimizing the relevance-diversity trade-off?
- Is it possible to improve diversity with small or even no impact on relevance?

As a result, Chapter 3 introduced two multimodal diversity promotion approaches for image retrieval. We have shown that combining multiple modalities and introducing a diversity promotion technique was able to effectively improve retrieval accuracy and subtopic coverage. Hence, we evolved this study towards evaluating the impact of diversity promotion on interactive search sessions and verifying whether it is possible to boost interactive learning with diversity. Therefore, we drew new hypotheses, such as:

- Learning-to-rank methods usually outperform traditional multimodal retrieval methods and it should also happen when diversity is integrated;
- Diversifying retrieval results covers different search aspects and should improve user-system information exchange and consequently speed up interactive learning; and
- Assuming learning with diversity positively impacts retrieval effectiveness, the retrieval methods using each modality may equally benefit from the diverse information.

Derived from these statements, our analysis was driven by some research questions, for instance:

- Is it possible to boost interactive learning with diversity?
- How the "strength/intensity" of the diversification affects the overall results?
- How the diversification approaches behave when used along with different retrieval modalities and vice versa?
- What is the impact of diversification on our learning approach and on the baseline fusion techniques in the context of multimodal relevance feedback?

Bridging interactive learning-to-rank, multimodal retrieval, and diversity promotion, Chapters 4 and 5 described our proposal and experimental analysis on diversity-oriented learning for multimodal image retrieval. The thorough experimental evaluation has shown how the user interaction over diversified results is able to boost the learning procedure resulting in higher retrieval effectiveness while also speeding up the discovery of relevant items in earlier iterations.

Considering the multiple modalities and multimodal approaches our experiments have shown that combining multiple sources of information consistently outperformed the single modality retrieval. Moreover, as a boosting factor, the diversity promotion methods

also had superior impact on multimodal sessions allowing higher effectiveness gains when compared to the sessions with individual modalities.

Finally, while our interactive learning framework has been instantiated for multimodal image retrieval, the proposed framework may be adapted to any kind of digital objects adherent to a relevance model. For instance, it may be applied for different multimedia data, such as videos, speech, music, 3D objects, etc.

6.1 Publications

As a result of our research activities, we have the following associated publications:

- R. T. Calumby, R. da S. Torres, and M. A. Gonçalves. Diversity-driven learning for multimodal image retrieval with relevance feedback. In Proceedings of the 21st IEEE International Conference on Image Processing, pages 2197–2201, 2014;
- R. T. Calumby, V. P. Santana, F. S. Cordeiro, O. A. B. Penatti, L. T. Li, G. Chiachia, and R. da S. Torres. Recod @ Mediaeval 2014: Diverse social images retrieval. In Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014;
- R. T. Calumby, Iago B. A. do C. Araujo, V. P. Santana, J. A. V. Muñoz, O. A. B. Penatti, L. T. Li, J. Almeida, G. Chiachia, M. A. Gonçalves, and R. da S. Torres. Recod @ Mediaeval 2015: Diverse social images retrieval. In Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015;
- R. T. Calumby, R. da S. Torres, and M. A. Gonçalves. On Interactive Learning-to-Rank for IR: Overview, Recent Advances, Challenges, and Directions [Accepted for publication in Neurocomputing]; and
- R. T. Calumby, R. da S. Torres, and M. A. Gonçalves. Diversity-based Interactive Learning meets Multimodality [Under review].

Finally, during the development of this thesis we also had the following related works published in collaboration:

- D. C. G. Pedronette, R. T. Calumby, R. da S. Torres. A semi-supervised learning algorithm for relevance feedback and collaborative image retrieval. EURASIP Journal on Image and Video Processing, v. 2015:27, 2015;
- D. C. G. Pedronette, R. da S. Torres, R. T. Calumby: Using contextual spaces for image re-ranking and rank aggregation. Multimedia Tools and Applications, 69(3):689-716 (2014);
- L. T. Li, D. C. G. Pedronette, J. Almeida, O. A. B. Penatti, R. T. Calumby, R. da S. Torres: A rank aggregation framework for video multimodal geocoding. Multimedia Tools and Applications, 73(3):1323-1359 (2014);

- L. T. Li, J. A. V. Muñoz, J. Almeida, R. T. Calumby, O. A. B. Penatti, I. C. Dourado, K. Nogueira, P. R. Mendes-Junior, L. A. M. Pereira, D. C. G. Pedronette, J. A. dos Santos, M. A. Gonçalves, R. da S. Torres: RECOD @ Placing Task of MediaEval 2015. MediaEval 2015;

- L. T. Li, O. A. B. Penatti, J. Almeida, G. Chiachia, R. T. Calumby, P. R. Mendes-Junior, D. C. G. Pedronette, R. da S. Torres: Multimedia Geocoding: The RECOD 2014 Approach. MediaEval 2014;

- D. C. G. Pedronette, O. A. B. Penatti, R. T. Calumby, R. da S. Torres: Unsupervised Distance Learning By Reciprocal kNN Distance for Image Retrieval. ICMR 2014, p.345;

- D. C. G. Pedronette, R. T. Calumby, R. da S. Torres: Semi-supervised Learning for Relevance Feedback on Image Retrieval Tasks. SIBGRAPI 2014, p.243-250;

- L. T. Li, J. Almeida, O. A. B. Penatti, R. T. Calumby, D. C. G. Pedronette, M. A. Gonçalves, R. da S. Torres: Multimodal Image Geocoding: The 2013 RECOD's Approach. MediaEval 2013;

- L. T. Li, D. C. G. Pedronette, J. Almeida, O. A. B. Penatti, R. T. Calumby, R. da S. Torres: Multimedia multimodal geocoding. SIGSPATIAL/GIS 2012, p.474-477; and

- F. A. Faria, R. T. Calumby, R. da S. Torres: RECOD at ImageCLEF 2011: Medical Modality Classification using Genetic Programming. CLEF (Notebook Papers/Labs/Workshop) 2011;

6.2 Future Work

With the wide review on IIR concepts and foundations (Chapter 2) and specially considering the most recent works, we have gathered a list of the great challenges and trends for the development of modern IIR systems (Appendix B). Beyond it, we have also provided a compiled list of promising directions in several related fields (Appendix C).

Therefore, these findings may guide and motivate several research paths not only with relation to the research questions and the experimental analysis covered here, but also for other initiatives on the IIR field.

For organizing our foreseen future work, we have grouped them into the following. Section 6.2.1 presents possible future work when considering non-interactive diversity methods which may also be extended to the interactive context. Section 6.2.2 presents the possible extensions and improvements over our interactive learning-to-rank method. Finally, Section 6.2.3 lists several user related research paths that can be pursued in order to better analyze the retrieval engines and also more precisely capture and analyze the user-system relationship and how the different methods may impact the user search experience.

6.2.1 Multimodal Diversity Promotion

Combining multimodal evidences for information retrieval has been extensively shown as an effective way of maximizing search effectiveness. However, when diversity constraints are considered, quite a few works have discussed about the relationship between the different modalities and diversity promotion or how these multiple modalities could be integrated to boost diversity promotion. Therefore, we can highlight future research opportunities, such as:

- Using additional information on the reranking and diversification steps. For instance, integrating data from user's social relationships, conceptual information from thesauri or gazetteers, unlabeled data [151], and data contextual information [154];
- Features: content-based multimedia retrieval benefits from effective feature descriptors which may also directly impact content-based diversity scores. Consequently, using better features may boost the ranking and diversity promotion methods. For instance, diversity promotion methods can take advantage of modern feature descriptors and representations including local feature descriptors, bag-of-visual-features, and deep learning-based features [231]; and
- Fusion: beyond finding better features, effectively integrating multiple features and modalities is still a task open to further development. Considering not only rank fusion strategies but also feature combination for diversity promotion procedures, proposing selection and fusion strategies for the computation of diversity scores may enhance retrieval results, specially if explored in a per query fashion.

Specially considering the multimodal diversity promotion methods from Chapter 3, one may develop new methods for each of the steps in the proposed pipelines. As promising directions, one may consider:

- Filtering: eliminating non-relevant items from the candidate group directly impacts the relevance of the final set, while also prevents introducing outliers on diversity estimations. For improving the relevance filtering one can explore:
  - Geocoding: modern image capturing devices also store geographic information and this kind of device became very popular with the integration of GPS technology into smartphones. However, most of the data currently available and legacy repositories do not possess such information. Therefore, for integrating image with and without geographic information, geocoding methods, such as [126], may be applied and consequently aid the retrieval and diversification approaches;
  - Face based filtering: the filtering strategies may evolve from simple face detection and basic statistics to more advanced methods. For instance, face-based filters may be trained to generate association rules or decision trees in order to assess whether an image is relevant or not; and

  - Consensus: for taking advantage of the multiple evidences of relevance of a given item future work may combine these information, e.g., using decision methods (mentioned in the previous item) or even using traditional ensemble learning methods.

- Reranking: in this thesis we have used linear combination and rank fusion methods for improving ranked lists. As an alternative, other fusion methods may be applied, e.g., the multiple ranking measures (textual, visual, geographic, and user credibility) could be combined using the GP framework describe in Section 4.1.1;

- Diversity methods: several diversity promotion methods have been proposed, including function-based [31, 62], graph-based [127], and clustering methods [28]. These methods may be found adequate for different scenarios and should also be considered. Moreover, the effective-cost trade-off may be balanced by choosing the most suitable method; and

- Clustering: Maximizing diversification effectiveness may depend on the proper combination of a given data domain and the applied diversification method. Moreover, even when considering clustering methods the best method quite often can just be selected with experimental evaluation [96]. Therefore, novel diversity promotion methods should be developed while there is also a great menu of clustering techniques available and which should be evaluated.

6.2.2 Diversity-driven Interactive Learning

In our experimental analysis, we have shown how interactive sessions may be affected by introducing diverse information and feedback into the learning method. Moreover, we have also discussed the natural multidisciplinary nature of interactive systems and the lack of extensive works which exploit diversity methods for learning-to-rank optimization. Therefore, the work described in this thesis opens several opportunities on diversity-oriented IIR research, such as:

- Features: in the experimental analysis in Chapter 4 and 5, we have shown that the retrieval effectiveness is directly impacted by the type o modality used. While choosing and combining multiple modalities is an effective way of improving retrieval success and reducing the semantic gap, it is also important to use good features for each retrieval modalities. Therefore, as mentioned for the multimodal non-interactive method, improving retrieval accuracy and diversification may be facilitated by applying better feature descriptions and representations;

- Learning: our diversity-oriented GP-based learning framework has been shown as an effective way for improving retrieval effectiveness with and without diversity promotion when compared to rank aggregation methods. However, it still suffers from the cold-start problem. We have seen that allowing the user to provide and initial feedback round over a non-diversified result allowed a better learning session. Therefore, introducing more advanced methods for the generation of the first result

may attenuate such problem. For this purpose, one could include a query-adaptive semi-supervised learn to rank method (see Section 2.4.2) or even include an off-line training stage, e.g., using deep learning methods;

- Diversity promotion: given the interactive learning nature and the experimental evaluation protocol, the relevance-diversity trade-off optimization may be achieved by analyzing different diversity promotion methods. Hence, different applications may benefit from different diversification methods. For instance, novel works should extend our analysis in Chapter 5 by studying the impact of the different modalities on the learning process with different diversification methods and also with different learning algorithms, e.g., multi-objective learning;

- Novelty: while somehow the novelty criteria [169] have been captured by measuring diversity variation throughout the session iterations, we did not use any measure for explicitly assessing it. Therefore, the experimental analysis may be extended to compare the methods considering the novelty criteria specifically; and

- Diversity analysis: while relevance assessment measures have been developed for a long time and some standard measures arose from its recurrent usage, diversity evaluation measures still demand further development. While it is possible to extend our analysis by introducing new diversity measures, e.g., intent-aware measures [32], novel diversity measures should be proposed and evaluated not only for non-interactive but also and specially considering interactive scenarios. By considering additional measures, a learning method may dynamically estimate how much diversification a query demands and consequently adjust the relevance-diversity trade-off.

6.2.3 User-centric Aspects

As highlighted in our literature review, conducting experiments with real users, while quite important for system analysis, is an often neglected activity in IIR research. Therefore, for a better evaluation and understanding of the retrieval system, novel research projects may be developed focused on:

- Real user experiments: a possible future work consists in running such kind of experiments not only for evaluating our proposals with real users but also to assess how close are real user interactions to user models when diversity is included in the retrieval process;

- Effectiveness metrics: when putting real user in an experimental retrieval session with diversity constraints, it is important to properly capture response signals and therefore it is necessary to develop and/or validate measures of success for this kind of environment; and

- Diversity-oriented visualization resources: multimedia information retrieval has fostered the development of adaptive visualization structures. However, those interfaces have focused only in providing the user some sense of relevance among the

item presented. In interactive retrieval and specially considering diverse results, using proper displaying structures may help the user in fulfilling her search task and also aid in maximizing the information transfer between the user and the system. Therefore, our system should be evaluated with real users interacting with traditional interfaces or adaptive methods, e.g., the clustered structures from [156, 157].

Bibliography

[1] G. Aggarwal, T. Ashwin, and S. Ghosal. An image retrieval system with automatic query modification. *IEEE Transactions on Multimedia*, 4(2):201–214, Jun 2002.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *ACM International Conference on Web Search and Data Mining*, pages 5–14, 2009.

[3] J. Allan. HARD track overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, pages 24–37, 2003.

[4] T. Amin, M. Zeytinoglu, and L. Guan. Application of laplacian mixture model to image and video retrieval. *IEEE Transactions on Multimedia*, 9(7):1416–1429, Nov 2007.

[5] J. Amores, N. Sebe, P. Radeva, T. Gevers, and A. Smeulders. Boosting contextual information in content-based image retrieval. In *Proceedings of the 6th ACM International Workshop on Multimedia Information Retrieval*, pages 31–38, New York, NY, USA, 2004. ACM.

[6] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada*, pages 561–568, 2002.

[7] M. Arevalillo-Herráez and F. J. Ferri. An improved distance-based relevance feedback strategy for image retrieval. *Image and Vision Computing*, 31(10):704 – 713, 2013.

[8] M. Arevalillo-Herráez, F. J. Ferri, and S. Moreno-Picot. Distance-based relevance feedback using a hybrid interactive genetic algorithm for image retrieval. *Applied Soft Computing*, 11(2):1782–1791, Mar. 2011.

[9] M. Arevalillo-Herráez, F. J. Ferri, and S. Moreno-Picot. Improving distance based image retrieval using non-dominated sorting genetic algorithm. *Pattern Recognition Letters*, 53(0):109 – 117, 2015.

[10] T. Arni, P. Clough, M. Sanderson, and M. Grubinger. Overview of the imageclefphoto 2008 photographic retrieval task. In *Proceedings of the 9th Cross-language*

Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access, pages 500–511, Berlin, Heidelberg, 2009. Springer-Verlag.

[11] T. Arni, J. Tang, M. Sanderson, and P. Clough. Creating a test collection to evaluate diversity in image retrieval. *Beyond binary relevance: preferences, diversity and set-level judgments, SIGIR Workshop*, 2008.

[12] P. Atrey, M. Hossain, A. E. Saddik, and M. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):1–35, 2010.

[13] A. Axenopoulos, S. Manolopoulou, and P. Daras. Optimizing multimedia retrieval using multimodal fusion and relevance feedback techniques. In K. Schoeffmann, B. Merialdo, A. Hauptmann, C.-W. Ngo, Y. Andreopoulos, and C. Breiteneder, editors, *Advances in Multimedia Modeling*, volume 7131 of *Lecture Notes in Computer Science*, pages 716–727. Springer Berlin Heidelberg, 2012.

[14] L. Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15–24, New York, NY, USA, 2011. ACM.

[15] L. Azzopardi. Modelling interaction with economic models of search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, New York, NY, USA, 2014. ACM.

[16] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 23–32, New York, NY, USA, 2013. ACM.

[17] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition.* Pearson Education Ltd., Harlow, England, 2011.

[18] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114, New York, NY, USA, 2012. ACM.

[19] F. Baskaya, H. Keskustalo, and K. Järvelin. Modeling behavioral factors ininteractive information retrieval. In *Proceedings of the 22nd ACM international conference on Conference on information knowledge management*, pages 2297–2302, New York, NY, USA, 2013. ACM.

[20] S. Bhatia, C. Brunk, and P. Mitra. A query classification scheme for diversification. In *Proceedings of the Second International Workshop on Diversity in Document Retrieval*, 2012. Co-Located with WSDM 2012.

[21] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17(2):216–228, Feb 2015.

[22] W. Bian and D. Tao. Biased discriminant euclidean embedding for content-based image retrieval. *IEEE Transactions on Image Processing*, 19(2):545–554, Feb 2010.

[23] C. Brandt, T. Joachims, Y. Yue, and J. Bank. Dynamic ranked retrieval. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 247–256, New York, NY, USA, 2011. ACM.

[24] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover Publications, 1966.

[25] E. Bruno, J. Kludas, and S. Marchand-Maillet. Combining multimodal preferences for multimedia information retrieval. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 71–78, 2007.

[26] C. Buckley. The TREC-9 query track. In *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*, 2000.

[27] R. T. Calumby, R. da S. Torres, and M. A. Gonçalves. Diversity-driven learning for multimodal image retrieval with relevance feedback. In *Proceedings of the 21st IEEE International Conference on Image Processing*, pages 2197–2201, 2014.

[28] R. T. Calumby, I. B. A. do C. Araujo, V. P. Santana, J. A. V. Muñoz, O. A. B. Penatti, L. T. Li, J. Almeida, G. Chiachia, M. A. Gonçalves, and R. da Silva Torres. Recod @ mediaeval 2015: Diverse social images retrieval. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015.*, 2015.

[29] R. T. Calumby, V. P. Santana, F. S. Cordeiro, O. A. B. Penatti, L. T. Li, G. Chiachia, and R. da Silva Torres. Recod @ mediaeval 2014: Diverse social images retrieval. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.*, 2014.

[30] R. T. Calumby, R. d. S. Torres, and M. A. Gonçalves. Multimodal retrieval with relevance feedback based on genetic programming. *Multimedia Tools and Applications*, 69(3):991–1019, 2014.

[31] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

[32] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S. Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

[33] X. Chen, A. Hero, and S. Savarese. Multimodal video indexing and retrieval using directed information. *IEEE Transactions on Multimedia*, 14(1):3–16, Feb 2012.

[34] X. Chen, C. Zhang, S.-C. Chen, and M. Chen. A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval. In *Seventh IEEE International Symposium on Multimedia*, page 8 pp., Dec 2005.

[35] E. Cheng, F. Jing, and L. Zhang. A unified relevance feedback framework for web image retrieval. *IEEE Transactions on Image Processing*, 18(6):1350–1357, June 2009.

[36] C.-C. Chiang, M.-H. Hsieh, Y.-P. Hung, and G. Lee. Region filtering using color and texture features for image retrieval. In W.-K. Leow, M. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. Bakker, editors, *Image and Video Retrieval*, volume 3568 of *Lecture Notes in Computer Science*, pages 487–496. Springer Berlin Heidelberg, 2005.

[37] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 48:1–48:9, New York, NY, USA, 2009. ACM.

[38] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–666, 2008.

[39] M. Cord and P. H. Gosselin. Image retrieval using long-term semantic learning. In *IEEE International Conference on Image Processing*, pages 2909–2912, Oct 2006.

[40] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.

[41] A. T. da Silva, A. X. Falcão, and L. P. Magalhães. Active learning paradigms for cbir systems based on optimum-path forest classification. *Pattern Recognition*, 44(12):2971 – 2978, 2011.

[42] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 2008.

[43] B. Demir and L. Bruzzone. An effective active learning method for interactive content-based retrieval in remote sensing images. In *IEEE International Geoscience and Remote Sensing Symposiums*, pages 4356–4359, July 2013.

[44] B. Demir and L. Bruzzone. A novel active learning method in relevance feedback for content-based remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2323–2334, May 2015.

[45] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*, pages 39:1–39:8, 2009.

[46] L. Duan, W. Li, I. Tsang, and D. Xu. Improving web image search by bag-based reranking. *IEEE Transactions on Image Processing*, 20(11):3280–3290, Nov 2011.

[47] S. Dumais. Whole-session evaluation of interactive information retrieval systems: Compilation of homework. http://research.microsoft.com/en-us/um/people/sdumais/niishonanworkshop-web/NII-Shonan-CompiledHomework_Final.pdf, NII Shonan Workshop, Oct 2012.

[48] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, Nov 2013.

[49] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[50] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[51] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

[52] R. Fagin, R. Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pages 301–312, 2003.

[53] W. Fan, E. A. Fox, P. Pathak, and H. Wu. The effects of fitness functions on genetic programming-based ranking discovery for web search. *Journal of the American Society for Information Science and Technology*, 55(7):628–636, 2004.

[54] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, June 2004.

[55] M. Ferecatu and N. Boujemaa. Interactive remote-sensing image retrieval using active relevance feedback. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):818–826, April 2007.

[56] M. Ferecatu and D. Geman. Interactive search for image categories by mental matching. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

[57] M. Ferecatu and D. Geman. A statistical framework for image category search from a mental picture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1087–1101, 2009.

[58] M. Ferecatu and H. Sahbi. Telecom paristech at imageclefphoto 2008: Bi-modal text and image retrieval with diversity enhancement. In *Working Notes for the Cross Language Evaluation Forum*, 2008.

[59] C. D. Ferreira, J. Santos, R. da S. Torres, M. Gonçalves, R. Rezende, and W. Fan. Relevance feedback based on genetic programming for image retrieval. *Pattern Recognition Letters*, 32(1):27 – 37, 2011. Image Processing, Computer Vision and Pattern Recognition in Latin America.

[60] A. Franco, A. Lumini, and D. Maio. A new approach for relevance feedback through positive and negative samples. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 4, pages 905–908 Vol.4, Aug 2004.

[61] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.

[62] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web*, pages 381–390, 2009.

[63] P. H. Gosselin. Online kernel learning for interactive retrieval in dynamic image databases. In *19th IEEE International Conference on Image Processing*, pages 1921–1924, Sept 2012.

[64] P. H. Gosselin, F. Precioso, and S. Philipp-Foliguet. Incremental kernel learning for active image retrieval without global dictionaries. *Pattern Recognition*, 44(10):2244–2254, 2011.

[65] G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.

[66] M. Grubinger, P. Clough, A. Hanbury, and H. Müller. Overview of the imageclefphoto 2007 photographic retrieval task. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 433–444. Springer, 2008.

[67] E. Guldogan, T. Olsson, E. Lagerstam, and M. Gabbouj. Instance based personalized multi-form image browsing and retrieval. *Multimedia Tools and Applications*, 71(3):1087–1104, 2014.

[68] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 124–131, New York, NY, USA, 2009. ACM.

[69] M. M. Hall and E. G. Toms. Building a common framework for IIR evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*, pages 17–28, 2013.

[70] M. Halvey, P. Punitha, D. Hannah, R. Villa, F. Hopfgartner, A. Goyal, and J. M. Jose. Diversity, assortment, dissimilarity, variety: A study of diversity measures using low level features for video retrieval. In *European Conference on Information Retrieval*, pages 126–137, 2009.

[71] M. Halvey and R. Villa. Evaluating the effort involved in relevance assessments for images. In *Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval*, pages 887–890, New York, NY, USA, 2014. ACM.

[72] J. Han, K. Ngan, M. Li, and H.-J. Zhang. A memory learning framework for effective image retrieval. *IEEE Transactions on Image Processing*, 14(4):511–524, April 2005.

[73] D. Harman. *Information Retrieval: Data Structures & Algorithms*, chapter Relevance Feedback and other query reformulation techniques. Prentice-Hall, 1992.

[74] D. Harman. Overview of the first trec conference. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 36–47, New York, NY, USA, 1993. ACM.

[75] D. Harman. Overview of the second text retrieval conference (trec-2). *Information Processing and Management*, 31(3):271–289, May 1995.

[76] D. K. Harman. *Overview of the Third Text Retrieval Conference (TREC-3)*. DIANE Publishing Company, 1996.

[77] J. He, E. Meij, and M. de Rijke. Result diversification based on query-specific cluster ranking. *Journal of the Association for Information Science and Technology*, 62(3):550–571, 2011.

[78] W. R. Hersh and P. Over. TREC-8 interactive track report. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, 1999.

[79] K. Hofmann, S. Whiteson, and M. de Rijke. Balancing exploration and exploitation in learning to rank online. In *European Conference on Information Retrieval*, pages 251–263, 2011.

[80] K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 249–258, New York, NY, USA, 2011. ACM.

[81] S. Hoi, W. Liu, M. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2072–2078, 2006.

[82] S. Hoi, M. Lyu, and R. Jin. A unified log-based relevance feedback scheme for image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):509–524, April 2006.

[83] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pages 35–44, New York, NY, USA, 2006. ACM.

[84] C. Huang and Q. Liu. An orientation independent texture descriptor for image retireval. In *International Conference on Computational Science*, pages 772–776, 2007.

[85] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[86] R. Huang, Q. Liu, H. Lu, and S. Ma. Solving the small sample size problem of lda. In *Proceedings. 16th International Conference on Pattern Recognition*, volume 3, pages 29–32 vol.3, 2002.

[87] Y. Huang, H. Huang, and J. Zhang. A noisy-smoothing relevance feedback method for content-based medical image retrieval. *Multimedia Tools and Applications*, 73(3):1963–1981, 2014.

[88] M. Huiskes. Image searching and browsing by active aspect-based relevance learning. In H. Sundaram, M. Naphade, J. Smith, and Y. Rui, editors, *Image and Video Retrieval*, volume 4071 of *Lecture Notes in Computer Science*, pages 211–220. Springer Berlin Heidelberg, 2006.

[89] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[90] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pages 527–536, New York, NY, USA, 2010. ACM.

[91] X. Hunag, S.-C. Chen, and M.-L. Shyu. Incorporating real-valued multiple instance learning into relevance feedback for image retrieval. In *Proceedings of the International Conference on Multimedia and Expo*, volume 1, pages I–321–4 vol.1, July 2003.

[92] B. Ionescu, A. Gînscă, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, Wurzen, Setember 14-15 2015. CEUR-WS.org.

[93] B. Ionescu, M. Menéndez, H. Müller, and A. Popescu. Retrieving diverse social images at mediaeval 2013: Objectives, dataset and evaluation. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, 2013.

[94] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînscă, and H. Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MediaEval 2014 Workshop*, Barcelona, 2014.

[95] A. Irtaza, M. Jaffar, and M. Muhammad. Content based image retrieval in a web 3.0 environment. *Multimedia Tools and Applications*, pages 1–18, 2013.

[96] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[97] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, Oct. 2002.

[98] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Proceedings of 30th European Conference on Advances in Information Retrieval*, pages 4–15, Berlin, Heidelberg, 2008. Springer-Verlag.

[99] X. Jin, J. French, and J. Michel. Toward consistent evaluation of relevance feedback approaches in multimedia retrieval. In *Proceedings of the Third international conference on Adaptive Multimedia Retrieval: user, context, and feedback*, pages 191–206, Berlin, Heidelberg, 2006. Springer-Verlag.

[100] X. Jin and J. C. French. Improving image retrieval effectiveness via multiple queries. In *Proceedings of the 1st ACM International Workshop on Multimedia Databases*, pages 86–93, New York, NY, USA, 2003. ACM.

[101] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1877–1890, Nov 2008.

[102] Y. Jing, M. Covell, D. Tsai, and J. Rehg. Learning query-specific distance functions for large-scale web image search. *IEEE Transactions on Multimedia*, 15(8):2022–2034, Dec 2013.

[103] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Tenth IEEE International Conference on Computer Vision*, volume 1, pages 604–610 Vol. 1, Oct 2005.

[104] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1053–1062, New York, NY, USA, 2011. ACM.

[105] E. Kanoulas, M. M. Hall, P. D. Clough, B. Carterette, and M. Sanderson. Overview of the TREC 2011 session track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18*, 2011.

[106] M. Karimzadehgan and C. Zhai. A learning approach to optimizing exploration-exploitation tradeoff in relevance feedback. *Information Retrieval*, 16(3):307–330, 2013.

[107] L. Kaufman and P. J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley, New York, 1990. A Wiley-Interscience publication.

[108] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224, Jan. 2009.

[109] D. Kelly, S. Dumais, and J. Pedersen. Evaluation challenges and directions for information-seeking support systems. *Computer*, 42(3):60–66, March 2009.

[110] D. Kelly and C. R. Sugimoto. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, 64(4):745–770, 2013.

[111] H. Keskustalo, K. Järvelin, and A. Pirkola. Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value. *Information Retrieval*, 11(3):209–228, June 2008.

[112] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, and M. Lykke. Test collection-based ir evaluation needs extension toward sessions — a case of extremely short queries. In *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, pages 63–74, Berlin, Heidelberg, 2009. Springer-Verlag.

[113] M. Kherfi, D. Brahmi, and D. Ziou. Combining visual features with semantics for a more effective image retrieval. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 961–964 Vol.2, Aug 2004.

[114] C. Kofler, M. Larson, and A. Hanjalic. Intent-aware video search result optimization. *IEEE Transactions on Multimedia*, 16(5):1421–1433, Aug 2014.

[115] A. Kotov, P. N. Bennett, R. W. White, S. T. Dumais, and J. Teevan. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 5–14, New York, NY, USA, 2011. ACM.

[116] V. Kovalev and S. Volmer. Color co-occurence descriptors for querying-by-example. In *Proceedings of the 1998 Conference on MultiMedia Modeling*, pages 32–38, 1998.

[117] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[118] M. K. Kundu, M. Chowdhury, and S. R. Bulò. A graph-based relevance feedback mechanism in content-based image retrieval. *Knowledge-Based Systems*, 73(0):254 – 264, 2015.

[119] A. Kustarev, Y. Ustinovskiy, A. Mazur, and P. Serdyukov. Session-based query performance prediction. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2563–2566, New York, NY, USA, 2012. ACM.

[120] L. J. Latecki and R. Lakämper. Shape similarity measure based on correspondence of visual parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1185–1190, Oct. 2000.

[121] E. Letouzé and J. Jütting. Official statistics, big data and human development. Technical report, Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute) and Paris2, 2015.

[122] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. Garner. Text similarity: an alternative way to search MEDLINE. *Bioinformatics*, 22(18):2298–2304, 2006.

[123] C. Li, Y. Wang, P. Resnick, and Q. Mei. Req-rec: High recall retrieval with query pooling and interactive classification. In *Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval*, pages 163–172, New York, NY, USA, 2014. ACM.

[124] J. Li and N. Allinson. Relevance feedback in content-based image retrieval: A survey. In M. Bianchini, M. Maggini, and L. C. Jain, editors, *Handbook on Neural Information Processing*, volume 49 of *Intelligent Systems Reference Library*, pages 433–469. Springer Berlin Heidelberg, 2013.

[125] J. Li, Q. Ma, Y. Asano, and M. Yoshikawa. Re-ranking by multi-modal relevance feedback for content-based social image retrieval. In *Proceedings of the 14th Asia-Pacific International Conference on Web Technologies and Applications*, pages 399–410, Berlin, Heidelberg, 2012. Springer-Verlag.

[126] L. T. Li, D. C. G. Pedronette, J. Almeida, O. A. B. Penatti, R. T. Calumby, and R. da Silva Torres. A rank aggregation framework for video multimodal geocoding. *Multimedia Tools and Applications*, 73(3):1323–1359, 2014.

[127] G. Lin, H. Peng, Q. Ma, J. Wei, and J. Qin. Improving diversity in web search results re-ranking using absorbing random walks. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pages 2116–2421, 2010.

[128] J. Lin and M. Efron. Overview of the trec-2013 microblog track. In *Proceedings of TREC*, volume 2013, 2013.

[129] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[130] T.-Y. Liu. *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.

[131] M. Lux and S. A. Chatzichristofis. LIRE: lucene image retrieval: an extensible java CBIR library. In *Proceedings of the 16th ACM International Conference on Multimedia*, pages 1085–1088, 2008.

[132] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696, New York, NY, USA, 2009. ACM.

[133] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys*, 46(3):38:1–38:38, Jan. 2014.

[134] J. Meng, J. Yuan, Y. Jiang, N. Narasimhan, V. Vasudevan, and Y. Wu. Interactive visual object search through mutual information maximization. In *Proceedings of the International Conference on Multimedia*, pages 1147–1150, New York, NY, USA, 2010. ACM.

[135] G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, Nov. 1995.

[136] I. Mironica, B. Ionescu, J. Uijlings, and N. Sebe. Fisher kernel based relevance feedback for multimodal video retrieval. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 65–72, New York, NY, USA, 2013. ACM.

[137] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel - evaluation in image retrieval. In M. Lew, N. Sebe, and J. Eakins, editors, *Image and Video Retrieval*, volume 2383 of *Lecture Notes in Computer Science*, pages 38–49. Springer Berlin Heidelberg, 2002.

[138] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-100). Technical report, Columbia University, 1996.

[139] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729, Dec 2008.

[140] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[141] P. Over. Trec-6 interactive track report. In *Proceedings of The Sixth Text REtrieval Conference, TREC 1997, Gaithersburg, Maryland, USA, November 19-21, 1997*, pages 73–82, 1998.

[142] P. Over. Trec-7 interactive track report. In *Proceedings of The Seventh Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998*, pages 33–39, 1998.

[143] P. Over, G. Awad, W. Kraaij, and A. F. Smeaton. Trecvid 2007 - overview. In *TRECVID 2007 workshop participants notebook papers, Gaithersburg, MD, USA, November 2007*.

[144] P. Over, G. M. Awad, J. Fiscus, M. Michel, A. F. Smeaton, and W. Kraaij. Trecvid 2009-goals, tasks, data, evaluation mechanisms and metrics. In *TRECVid Workshop 2009, 16-17 November 2009, Gaithersburg, MD, USA*, 2010.

[145] P. Over, G. M. Awad, T. Rose, J. Fiscus, W. Kraaij, and A. F. Smeaton. Trecvid 2008-goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2008 workshop participants notebook papers, Gaithersburg, MD, USA, November 2008*, 2008.

[146] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. Trecvid 2005 - an overview. In *TRECVid 2005 - Text REtrieval Conference TRECVID Workshop, 14-15 November 2005, Gaithersburg, Maryland*, 2005.

[147] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. Trecvid 2006 - an overview. In *TRECVid 2006 - Text REtrieval Conference TRECVid Workshop, 13-14 November 2006, Gaithersburg, Maryland*, 2006.

[148] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, 19(2):120–131, June 2009.

[149] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiaris. The 2012 social event detection dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 102–107, New York, NY, USA, 2013. ACM.

[150] S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, and I. Kompatsiaris. Social event detection at mediaeval 2011: Challenges, dataset and evaluation. In *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*, 2011.

[151] D. C. G. Pedronette, R. T. Calumby, and R. da Silva Torres. Semi-supervised learning for relevance feedback on image retrieval tasks. In *Proceedings of the 27th Conference on Graphics, Patterns and Images*, pages 243–250, Aug 2014.

[152] D. C. G. Pedronette and R. da S. Torres. Image re-ranking and rank aggregation based on similarity of ranked lists. In *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns - Volume Part I*, pages 369–376, 2011.

[153] D. C. G. Pedronette and R. da S. Torres. Exploiting pairwise recommendation and clustering strategies for image re-ranking. *Information Sciences*, 207(0):19 – 34, 2012.

[154] D. C. G. Pedronette, R. da S. Torres, and R. T. Calumby. Using contextual spaces for image re-ranking and rank aggregation. *Multimedia Tools and Applications*, 69(3):689–716, 2014.

[155] O. A. B. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. da S. Torres. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*, 47(2):705–720, 2014.

[156] S. M. Pinto-Caceres, J. Almeida, M. C. C. Baranauskas, and R. da Silva Torres. Fisir: A flexible framework for interactive search in image retrieval systems. In *Proceedings of the 21st International Conference MultiMedia Modeling, Sydney, NSW, Australia, January 5-7, 2015*, pages 335–347, 2015.

[157] S. M. Pinto-Caceres, J. Almeida, V. P. de Almeida Néris, M. C. C. Baranauskas, N. J. Leite, and R. da Silva Torres. Navigating through video stories using clustering sets. *International Journal of Multimedia Data Engineering and Management*, 2(3):1–20, 2011.

[158] A. Popescu and I. Kanellos. Creating visual summaries for geographic regions. In *European Conference on Information Retrieval, Workshop on Information Retrieval over Social Networks*, 2009.

[159] T. Qin and T. Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013.

[160] T. Qin, T.-Y. Liu, J. Xu, and H. Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, Aug. 2010.

[161] E. Rabinovich, O. Rom, and O. Kurland. Utilizing relevance feedback in fusion-based retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval*, pages 313–322, New York, NY, USA, 2014. ACM.

[162] K. Raman, P. Shivaswamy, and T. Joachims. Online learning to diversify from implicit feedback. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 705–713, 2012.

[163] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi. Information fusion between short term learning and long term learning in content based image retrieval systems. *Multimedia Tools and Applications*, 74(11):3799–3822, 2015.

[164] S. E. Robertson. Readings in information retrieval. chapter The Probability Ranking Principle in IR, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[165] S. E. Robertson and I. Soboroff. The trec 2002 filtering track report. In *TREC*, volume 2002, page 5, 2002.

[166] S. Rota Bulò, M. Rabbi, and M. Pelillo. Content-based image retrieval with relevance feedback using random walks. *Pattern Recognition*, 44(9):2109–2122, Sept. 2011.

[167] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2):95–145, June 2003.

[168] K. L. Santos, H. Almeida, R. da S. Torres, and M. A. Gonçalves. Recuperação de imagens da web utilizando múltiplas evidências textuais e programação genética. In *XXIV Simpósio Brasileiro de Bancos de Dados*, pages 91–105, 2009.

[169] R. L. T. Santos, P. Castells, I. S. Altingövde, and F. Can. Diversity and novelty in information retrieval. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, Dublin, Ireland - July 28 - August 01, 2013*, page 1130, 2013.

[170] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890, 2010.

[171] R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1179–1188, Toronto, Canada, 2010. ACM.

[172] R. L. T. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 595–604, 2011.

[173] R. L. T. Santos, C. Macdonald, and I. Ounis. Search result diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90, Mar. 2015.

[174] S. Schmiedeke, C. Kofler, and I. Ferrané. Overview of mediaeval 2012 genre tagging task. In *MediaEval 2012 Workshop, Pisa, Italy*, 2012.

[175] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems (NIPS)*, page 41, 2004.

[176] A. Shamsi, H. Nezamabadi-pour, and S. Saryazdi. A short-term learning approach based on similarity refinement in content-based image retrieval. *Multimedia Tools and Applications*, 72(2):2025–2039, 2014.

[177] J. A. Shaw, E. A. Fox, J. A. Shaw, and E. A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

[178] R. M. Silva, M. A. Gonçalves, and A. Veloso. Rule-based active sampling for learning to rank. In *European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, Greece, September 5-9, 2011, Proceedings, Part III*, pages 240–255, 2011.

[179] R. M. Silva, M. A. Gonçalves, and A. Veloso. A two-stage active learning method for learning to rank. *Journal of the American Society for Information Science and Technology*, 65(1):109–128, 2014.

[180] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pages 623–632, New York, NY, USA, 2007. ACM.

[181] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 399–402, New York, NY, USA, 2005. ACM.

[182] I. Soboroff, I. Ounis, J. Lin, and I. Soboroff. Overview of the trec-2012 microblog track. In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9*, 2012.

[183] R. Stehling, M. Nascimento, and A. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 102–109, 2002.

[184] N. Suditu and F. Fleuret. Iterative relevance feedback with adaptive exploration/exploitation trade-off. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1323–1331, New York, NY, USA, 2012. ACM.

[185] A. Sun and S. S. Bhowmick. Image tag clarity: In search of visual-representative tags for social images. In *Proceedings of the First SIGMM Workshop on Social Media*, pages 19–26, New York, NY, USA, 2009. ACM.

[186] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[187] B. Tao and B. Dickinson. Texture recognition and image retrieval using gradient indexing. *Journal of Visual Communication and Image Representation*, 11(3):327–342, 2000.

[188] D. Tao, X. Tang, and X. Li. Which components are important for interactive image searching? *IEEE Transactions on Circuits and Systems for Video Technology*, 18(1):3–11, Jan 2008.

[189] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1088–1099, July 2006.

[190] A. Thomas, C. Paul, M. Sanderson, and M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706, pages 500–511, 2009.

[191] B. Thomee, M. Huiskes, E. Bakker, and M. Lew. Using an artificial imagination for texture retrieval. In *19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008.

[192] B. Thomee, M. Huiskes, E. Bakker, and M. Lew. An exploration-based interface for interactive image retrieval. In *Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*, pages 188–193, Sept 2009.

[193] B. Thomee, M. J. Huiskes, E. Bakker, and M. S. Lew. Deep exploration for experiential image retrieval. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 673–676, New York, NY, USA, 2009. ACM.

[194] B. Thomee and M. S. Lew. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval*, 1(2):71–86, 2012.

[195] S. Tollari, P. Mulhem, M. Ferecatu, H. Glotin, M. Detyniecki, P. Gallinari, H. Sahbi, and Z.-Q. Zhao. A comparative study of diversity methods for hybrid text and image retrieval approaches. In *Cross Language Evaluation Forum*, pages 585–592, 2009.

[196] H. Tong, M. Li, H. Zhang, and C. Zhang. Blur detection for digital images using wavelet transform. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 17–20 Vol.1, 2004.

[197] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the Ninth ACM International Conference on Multimedia*, pages 107–118, New York, NY, USA, 2001. ACM.

[198] J. M. Torres, D. Hutchison, and L. P. Reis. Semantic image retrieval using region-based relevance feedback. In S. Marchand-Maillet, E. Bruno, A. Nürnberger, and M. Detyniecki, editors, *Adaptive Multimedia Retrieval: User, Context, and Feedback*, volume 4398 of *Lecture Notes in Computer Science*, pages 192–206. Springer Berlin Heidelberg, 2007.

[199] D. Tran, S. Pamidimukkala, and P. Nguyen. Relevance-feedback image retrieval based on multiple-instance learning. In *Seventh IEEE/ACIS International Conference on Computer and Information Science*, pages 597–602, May 2008.

[200] J. Urban and J. M. Jose. Adaptive image retrieval using a graph model for semantic feature integration. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 117–126, New York, NY, USA, 2006. ACM.

[201] J. Urban and J. M. Jose. Evaluating a workspace's usefulness for image retrieval. *Multimedia Systems*, 12(4-5):355–373, 2007.

[202] P. Vakkari. Exploratory searching as conceptual exploration. In *Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval*, pages 24–27, 2010.

[203] E. Valle and M. Cord. Advanced techniques in cbir: Local descriptors, visual dictionaries and bags of features. In *Tutorials of the XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI Tutorials)*, pages 72–78, 2009.

[204] R. van Zwol, V. Murdock, L. G. Pueyo, and G. Ramirez. Diversifying image search with user generated content. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 67–74, 2008.

[205] S. Vargas, P. Castells, and D. Vallet. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 75–84, 2012.

[206] J. A. Vargas Muñoz, R. da Silva Torres, and M. A. Gonçalves. A soft computing approach for learning to aggregate rankings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 83–92, New York, NY, USA, 2015. ACM.

[207] R. C. Veltkamp and M. Tanase. *Content-Based Image and Video Retrieval*, chapter A Survey of Content-Based Image Retrieval Systems, pages 47–101. Kluwer, 2002.

[208] M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina, and V. J. Tsotras. On query result diversification. In *Proceedings of the IEEE 27th International Conference on Data Engineering*, pages 1163–1174, 2011.

[209] R. Villa and M. Halvey. Is relevance hard work?: Evaluating the effort of making relevant assessments. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–768, New York, NY, USA, 2013. ACM.

[210] E. M. Voorhees. Overview of the trec 2003 robust retrieval track. In *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21*, pages 69–77, 2003.

[211] E. M. Voorhees and D. Harman. Overview of the seventh text retrieval conference trec-7. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7*, pages 1–24, 1998.

[212] E. M. Voorhees and D. Harman. Overview of the eighth text retrieval conference (trec-8). In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 1–24, 2000.

[213] E. M. Voorhees and D. Harman. Overview of the ninth text retrieval conference (trec-9). In *In Proceedings of the Ninth Text REtrieval Conference (TREC-9*, pages 1–14, 2000.

[214] E. M. Voorhees and D. Harman. Overview of the sixth text retrieval conference (trec-6). *Information Processing & Management*, 36(1):3–35, Jan. 2000.

[215] E. M. Voorhees and D. Harman. Overview of trec 2001. In *Proceedings of the Tenth Text REtrieval Conference*, 2001.

[216] M. Wang, L. Yang, and X.-S. Hua. MSRA-MM: Bridging research and industrial societies for multimedia information retrieval. Technical Report MSR-TR-2009-30, Microsoft, March 2009.

[217] R. Wang, S. J. McKenna, and J. Han. High-entropy layouts for content-based browsing and retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 16:1–16:8, New York, NY, USA, 2009. ACM.

[218] X. Wang and M. Kankanhalli. Multimedia fusion with mean-covariance analysis. *IEEE Transactions on Multimedia*, 15(1):120–128, Jan 2013.

[219] X.-Y. Wang, Y.-W. Li, H.-Y. Yang, and J.-W. Chen. An image retrieval scheme with relevance feedback using feature reconstruction and svm reclassification. *Neurocomputing*, 127:214–230, Mar. 2014.

[220] X.-Y. Wang, H.-Y. Yang, Y.-W. Li, W.-Y. Li, and J.-W. Chen. A new svm-based active feedback scheme for image retrieval. *Engineering Applications of Artificial Intelligence*, 37(0):43 – 53, 2015.

[221] X.-Y. Wei and Z.-Q. Yang. Coaching the exploration and exploitation in active learning for interactive video retrieval. *IEEE Transactions on Image Processing*, 22(3):955–968, 2013.

[222] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, June 2009.

[223] A. Williams and P. Yoon. Content-based image retrieval using joint correlograms. *Multimedia Tools and Applications*, 34(2):239–248, 2007.

[224] R. E. Williamson. Does relevance feedback improve document retrieval performance? In *Proceedings of the 1st annual international ACM SIGIR conference on Information storage and retrieval*, SIGIR '78, pages 151–170, New York, NY, USA, 1978. ACM.

[225] J. Wu, H. Shen, Y.-D. Li, Z.-B. Xiao, M.-Y. Lu, and C.-L. Wang. Learning a hybrid similarity measure for image retrieval. *Pattern Recognition*, 46(11):2927 – 2939, 2013.

[226] K. Wu, K.-H. Yap, and L.-P. Chau. Region-based image retrieval using radial basis function network. In *Proceeding of the IEEE International Conference on Multimedia and Expo*, pages 1777–1780, July 2006.

[227] P. Wu, B. S. Manjunanth, S. D. Newsam, and H. D. Shin. A texture descriptor for image retrieval and browsing. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, page 3, 1999.

[228] Z. Xiao and X. Qi. Complementary relevance feedback-based content-based image retrieval. *Multimedia Tools and Applications*, 73(3):2157–2177, 2014.

[229] H. Xie, V. Andreu, and A. Ortega. Quantization-based probabilistic feature modeling for kernel design in content-based image retrieval. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 23–32, New York, NY, USA, 2006. ACM.

[230] Q. Xing, Y. Zhang, and L. Zhang. On bias problem in relevance feedback. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1965–1968, New York, NY, USA, 2011. ACM.

[231] E. S. Xioufis, S. Papadopoulos, A. Gînsca, A. Popescu, Y. Kompatsiaris, and I. P. Vlahavas. Improving diversity in image search via supervised relevance scoring. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*, pages 323–330, 2015.

[232] F. Yan, K. Mikolajczyk, and J. Kittler. Multiple kernel learning via distance metric learning for interactive image retrieval. In C. Sansone, J. Kittler, and F. Roli, editors, *Multiple Classifier Systems*, volume 6713 of *Lecture Notes in Computer Science*, pages 147–156. Springer Berlin Heidelberg, 2011.

[233] J. Yang, Q. Li, and Y. Zhuang. Image retrieval and relevance feedback using peer indexing. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 409–412 vol.2, 2002.

[234] Y. Yang and S. Newsam. Geographic image retrieval using local invariant features. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(2):818–832, Feb 2013.

[235] E. Yilmaz and S. Robertson. On the choice of effectiveness measures for learning to rank. *Information Retrieval*, 13(3):271–290, June 2010.

[236] H. P. Young. An axiomatization of borda's rule. *Journal of Economic Theory*, 9(1):43–52, 1974.

[237] D. Zellhöfer. An extensible personal photograph collection for graded relevance assessments and user simulation. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pages 29:1–29:8, New York, NY, USA, 2012. ACM.

[238] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–17, 2003.

[239] C. Zhang and X. Chen. Region-based image clustering and retrieval using multiple instance learning. In W.-K. Leow, M. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. Bakker, editors, *Image and Video Retrieval*, volume 3568 of *Lecture Notes in Computer Science*, pages 194–204. Springer Berlin Heidelberg, 2005.

[240] J. Zhang and L. Ye. Content based image retrieval using unclean positive examples. *IEEE Transactions on Image Processing*, 18(10):2370–2375, Oct 2009.

[241] L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. In *Proceedings of the International Conference on Image Processing*, volume 2, pages 721–724 vol.2, Oct 2001.

[242] L. Zhang, S. Liu, Z. Wang, W. Cai, Y. Song, and D. D. Feng. Graph cuts based relevance feedback in image retrieval. In *IEEE International Conference on Image Processing, Melbourne, Australia, September 15-18, 2013*, pages 4358–4362, 2013.

[243] L. Zhang, L. Wang, and W. Lin. Semisupervised biased maximum margin analysis for interactive image retrieval. *IEEE Transactions on Image Processing*, 21(4):2294–2308, April 2012.

[244] Y. Zhang, W. Li, Z. Mo, T. Zhao, and J. Zhang. An adaptive-weight hybrid relevance feedback approach for content based image retrieval. In *20th IEEE International Conference on Image Processing*, pages 3977–3981, Sept 2013.

[245] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.

[246] Z.-H. Zhou and H.-B. Dai. Exploiting image contents in web search. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2928–2933, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Appendix A

IIR Bibliometrics

In this appendix, we provide a brief bibliometric view of the works covered in Chapter 2, ranging from 2011 to 2015. Considering such recent work, the corresponding conferences/journals covered in the period and the corresponding acronyms are presented in Tables A.1 and A.2.

We quantitatively analyzed the main target venues by showing their publication distribution in the period. Figures A.1 and A.2 present the number of articles in each of the analyzed conferences and journals, respectively. Similar to the findings in [110], the IIR works are concentrated in few venues but we can notice a slightly superior scattering on many conferences and journals. This suggests that researchers were able to introduce their work in several venues with different central subjects. This fact may be directly related to the multidisciplinary characteristic of the IIR field.

As depicted in Figure A.1, more than 60% of the papers from the last five years were published in three main conferences: SIGIR, ICIP, and CIKM. In turn, considering only journal papers (Figure A.2), roughly 58% of the papers were concentrated in four venues: MTAP, PRL, PR, and IEEE TIP.

Figure A.3 presents the number of papers per year and a visual representation of the contribution from each venue. We can observe that a similar amount of works were published in the last five years. The amount of papers for 2015 considers the works published until the date of the submission of this thesis.

Considering the described works which were published from 2011 to 2015, Figure A.4 presents a tag cloud for the twenty most frequent keywords whose sizes represent the corresponding number of occurrences. As a natural interactive retrieval method, "relevance feedback" was the most used keyword. One may also notice that many of the other most frequent keywords are related to image retrieval and machine learning.

Table A.1: Conference names and acronyms.

Acronym	**Conference Name**
APWEB	Asia-Pacific Web Conference
CIKM	ACM Conference on Information and Knowledge Management
CLEF	Conference and Labs of the Evaluation Forum
ECIR	European Conference on Information Retrieval
ICIP	IEEE International Conference on Image Processing
ICMR	ACM International Conference on Multimedia Retrieval
IGARSS	IEEE International Geoscience and Remote Sensing Symposium
MCS	International Workshop on Multiple Classifier Systems
SIBGRAPI	Conference on Graphics, Patterns and Images
SIGIR	ACM SIGIR Conference
SIGKDD	ACM Conference on Knowledge Discovery and Data Mining
WSDM	ACM International Conference on Web Search and Data Mining

Table A.2: Journal names and acronyms.

Acronym	**Journal Name**
ACS	ACM Computing Surveys
AMM	Advances in Multimedia Modeling (LNCS)
ASV	Applied Soft Computing
EAAI	Engineering Applications of Artificial Intelligence
IJMIR	International Journal of Multimedia Information Retrieval
IVC	Image and Vision Computing
IS	Information Sciences
JASSIST	Journal of the Association for Information Science and Technology
KBS	Knowledge-Based Systems
MTAP	Multimedia Tools and Applications
NC	Neurocomputing
PR	Pattern Recognition
PRL	Pattern Recognition Letters
TGRS	IEEE Transactions on Geoscience and Remote Sensing
TIP	IEEE Transactions on Image Processing

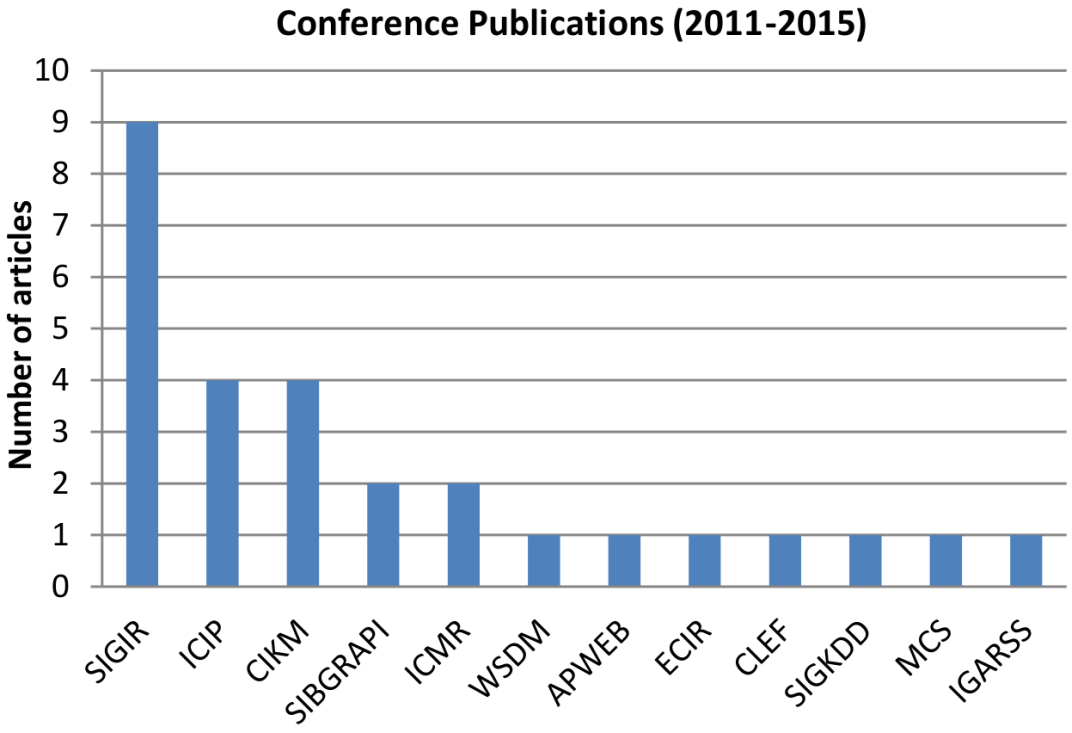

Figure A.1: Number of papers published per conference.

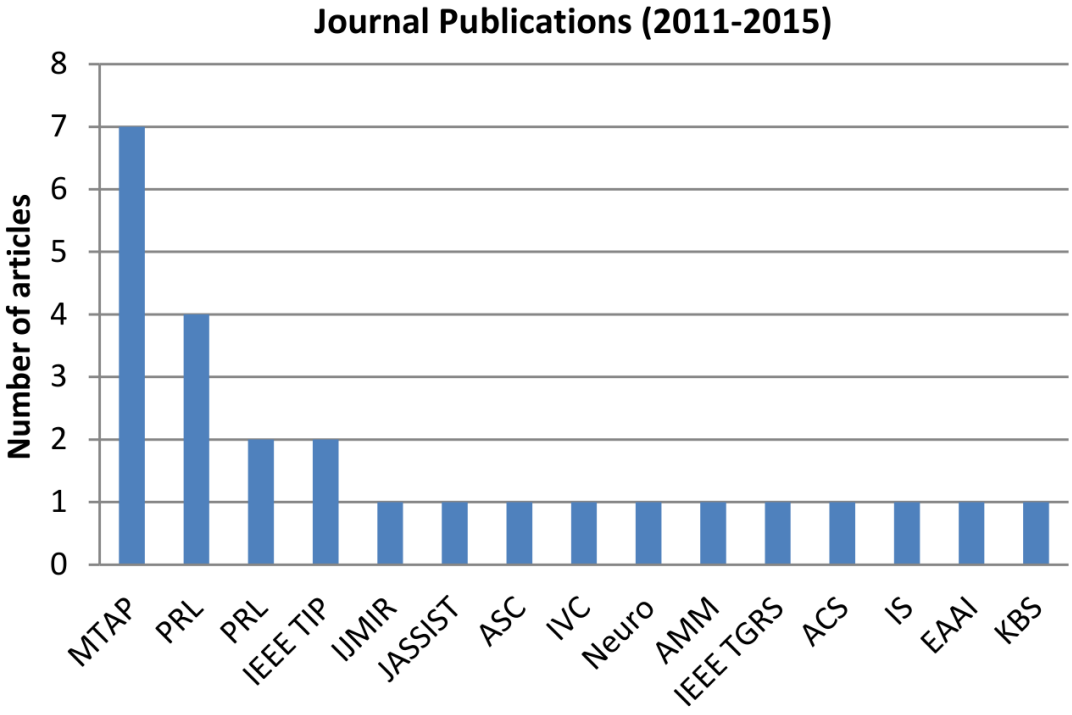

Figure A.2: Number of papers published per journal.

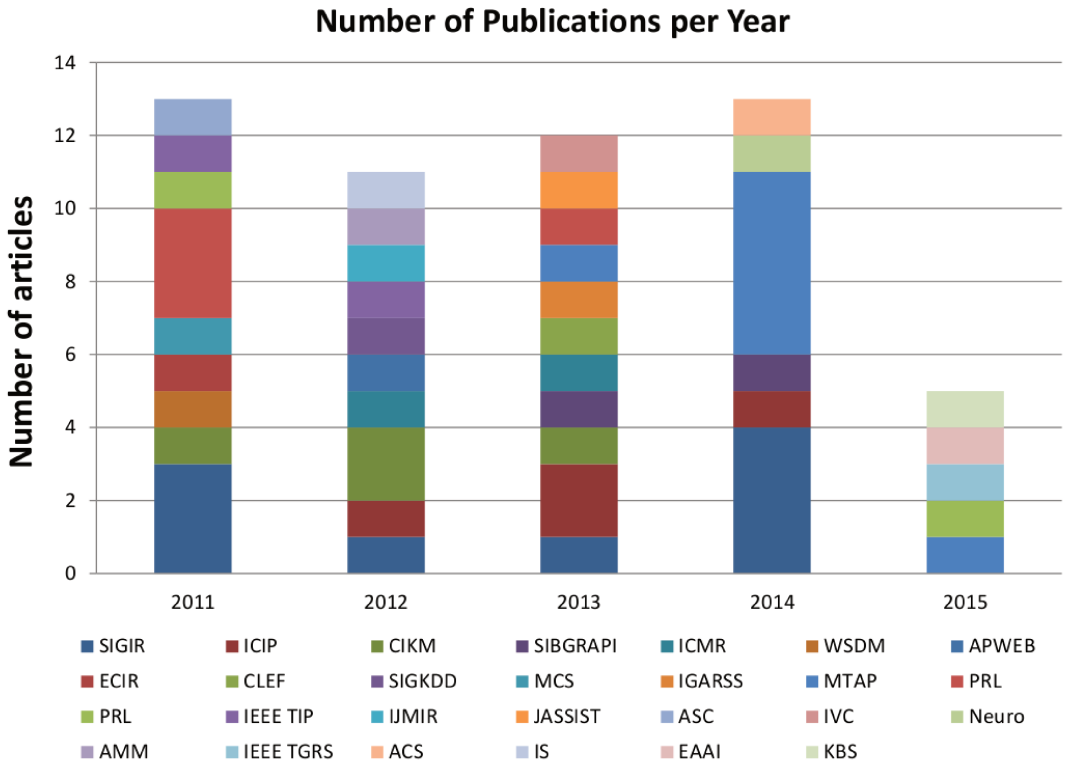

Figure A.3: Number of papers published per year in conferences and journals.

Figure A.4: Tag cloud for the 20 most frequent keywords in recent papers.

Appendix B

IIR Challenges and Trends

Considering the multidisciplinary characteristic of the interactive information retrieval field, the technological advances have integrated contributions from different research fields. Moreover, each of these fields presents specific challenges, which become even more complex with vast merging possibilities.

According to [194], the main challenges related to interactive search are:

1. Optimal user interface (query specification and results exhibition) design: in this aspect, in parallel to results accuracy, we have to target user's satisfaction and also her understanding of why such results were returned;

2. Tags and comments exploration: the huge amount of information produced on social networks can be explored as it provides knowledge for better estimating the relationship between images and their content;

3. Achieving good accuracy with a few training samples: such difficulty may be reduced by using new learning algorithms, for instance with manifold learning, improving multi-modal fusion methods, and making better use of implicit feedback. In our experimental analysis, we have seen how important is the usage of sufficient training information in order to properly answer the user needs and attenuate the drawbacks of the diversity promotion approaches (see Chapter 4); and

4. Overcoming evaluation issues: for better designing, evaluating, and tuning the interactive systems researchers have to pursue allowing high quality ground-truth construction, better benchmarks, proposing more suitable/effective evaluation measures, conducting real user experiments, and also more advanced user modeling. In our analysis, we strive to report as detailed as possible all processing steps and configurations in order to facilitate reproduction and allow the adequate reasoning over the outcomes. Additionally, we relied on an adequate and meaningful evaluation protocol and measures and strictly compared the results with statistical and stability analysis.

For simplicity we grouped some of the challenges, gathered and inferred from recent work, into the following:

1. Theory: researchers and industry possess some well-established theoretical foundation for IR, which is not yet the case for interactive methods. Therefore, proposing new formal foundations for interactive systems may allow the development of better solutions, better analysis, and superior user satisfaction. However, given the dynamic environment of interactive retrieval and the many interfering factors, integrating all such aspects into unified formal frameworks is a challenging endeavor. Even though it has not been a formal proposal for an standard procedure, the literature review, evaluation protocol, and measures reported in this thesis may serve as an important source of information for novel research;

2. Data: with the ever increasing availability of social, linked, log, and mainly unlabeled data (see Section 2.4.2), it becomes important to develop methods that are able to explore this wide sources of information, as well as integrating multiple sources of evidence particularly inherent to multimedia data. Our investigations were conducted with datasets, which provide multiple sources of information, e.g., visual, textual, and geographic, and we have proposed specific processing methods for each of them, as well as many integration alternatives;

3. Learning: the effectiveness of search systems in capturing real user interpretations and automatically adjusting internal models still needs to be further improved in order to attenuate, e.g., the cold-start problem (few training samples) or even the case of iterations with no feedback at all. Similarly, as described in Section 2.4.1, automatically adjusting the exploration-exploitation trade-off is still an open issue and may benefit from advanced learning models. In our interactive search analysis, we proposed avoiding promoting diversity in the first result of a session in order to allow the user to provide the system with more relevant information. This information was used in the following iterations to strengthen the learning process, find novel relevant items, and then allow proper diversification with attenuated drawbacks;

4. User: Regarding user interactions, the retrieval systems face important challenges considering user fallibility on providing correct feedback and also on drifting her information need within the same search session. Therefore, new studies are necessary on system's sensitivity to erroneous feedback and also the construction of benchmarks that properly assess these difficulties; and

5. Scale: In the age of the ever-increasing data generation rate, developing effective retrieval systems becomes even more crucial. Being capable of handling extremely large, dynamic, multimedia, and linked data is a must-have feature for modern search engines. Therefore, capturing, indexing, and searching over large amounts of data is a natural demand for future retrieval systems.

B.1 IIR Evaluation Challenges

Providing the adequate theoretical and practical tools for IIR research is an important factor for tackling several issues previously mentioned. As a special case, evaluation activities still suffer from the absence of integrated frameworks and standard approaches.

According to [69], "*the challenge is two-fold: developing a standard methodological protocol that may service multiple types of IIR evaluations and research, and developing a standard set of meaningful measures that are more than descriptive of the process... The main challenge lies in creating a framework that is sufficiently standardized to enable comparability of evaluation results, while at the same time being flexible enough to be applied to a wide range of experiments and variables in order to ensure its uptake.*"

Considering the recent works on IIR evaluation and the obstacles found, some of the main challenges are:

1. The development of effectiveness measures that are more informative and better suited for learning-to-rank methods;

2. The proposal of better interactivity cost functions to evaluate search strategies and user effort on retrieval sessions;

3. The development of better log analysis methods, click models, and user models considering reformulation understanding, stopping criteria, and erroneous feedback simulation; and

4. The performance of experiments with real-life settings. Conducting real user studies has always been a difficult task and often neglected. Nevertheless, contrasting lab-based analysis with real environment data is helpful not only for assessing system's performance, but also for validating modeling approaches.

Appendix C

IIR Promising Research Directions

According to [194], some promising directions on improving interactive search systems rely on exploring:

1. Q&A Paradigm focused on multi-modality and cross-modality;
2. Interaction by explanation: modern interactive search systems are expected to explain to the user why the results were chosen and also allow her to provide feedback based on the explanations;
3. Exploring external sources: interactive systems can explore additional image collections and knowledge sources for improving retrieval effectiveness; and
4. Social interaction for system's optimization through collaborative filtering.

As clearly observed in [110], and exposed in this text, there is a lack of standard evaluation methods and measures. As the availability of standards is considered a requirement for the maturation of a research field, there is still a great need for IIR standardization. As reported, the majority of evaluation datasets and benchmarks are constructed for system-centric research, which presents a promising direction on developing data infrastructure specifically designed for interactive retrieval.

Analyzing the recent proposals and trends, we can highlight some aspects of interactive learning-to-rank methods which deserve further investigation and development effort:

1. Exploring unlabeled data and semi-supervised methods, for reducing labeling effort, attenuating the cold-start problem, and consequently the effectiveness of classifiers;
2. Differentiating positive and negative samples treatment on the learning process for their different representativeness in relation to real data distribution;
3. Exploring learning boosting alternatives such as diversity promotion for handling ambiguous, multi-intent, overview, or underspecified queries;
4. Using reinforcement learning methods for combining multiple feature modalities or even multiple learning strategies such as active learning and exploration/exploitation;

5. Analyzing user behavior impacts on search tasks which will produce information for the development of better generalization models and more realistic user models;

6. Leveraging long-term learning and collaborative retrieval for effectiveness and efficiency improvement;

7. Using graded relevance assessments as a way to improve ground-truth quality and maximize feedback information. For conducting user-centric evaluations the work in [110] also suggests using nDCG [97] for effectiveness evaluation as it relies on graded relevance assessments and has been experimentally demonstrated effective for user-centered tasks. Moreover, nDCG is also capable of reflecting small changes or re-ordering of relevant documents; and

8. Reducing RF bias since the non-relevant samples are generally less representative than the relevant samples, w.r.t. the whole data collection, which leads to imbalanced training sets and consequently inaccurate classification boundaries.

Appendix D

Visual Examples

In this appendix, we present a set of visual examples of the results from Chapter 3, part 2. We present four examples (with top-20 images) from the test set queries for the Ranking 4 (which produced the best F1@20).

D.1 Venice carnival

Figure D.1 presents an example of result with high precision and low diversity. It corresponds to the query *Venice carnival* with 3 out of 10 clusters represented. The effectiveness values are:

- P@20 = 0.95
- CR@20 = 0.30
- F1@20 = 0.46

Figure D.1: Top-20 results for the query *Venice carnival*. Highlights: non-relevant (red), cluster 3 (green), cluster 4 (blue), and cluster 1 (all the others).

As we can observe from Figure D.1, although only one non-relevant image was retrieved, in the top-20 images only 30% of the available clusters were represented, which had great impact on the relevance-diversity balance.

D.2 Tropical rain

Figure D.2 presents an example of result with low precision and high diversity. It corresponds to the query *Tropical Rain* with 4 out of 5 clusters represented. The effectiveness values are:

- P@20 = 0.20
- CR@20 = 0.80
- F1@20 = 0.32

Figure D.2: Top-20 results for the query *Tropical rain*. Highlights: cluster 1 (green), cluster 2 (yellow), cluster 3 (blue), cluster 5 (magenta), and non-relevant (all the others).

As we can observe from Figure D.2, only 4 out of 20 images were relevant. Since all relevant images were grouped in five clusters, this result correspond to 80% of cluster coverage even with low precision. Therefore, the user received few relevant items but at least one relevant item for most of the clusters.

D.3 *Biennale de la danse de Lyon*

Figure D.3 presents an example of result with a good a precision-diversity balance. It corresponds to the query *Biennale de la danse de Lyon* with 7 out of 10 clusters represented. The effectiveness values are:

- P@20 = 0.80
- CR@20 = 0.70
- F1@20 = 0.75

Figure D.3: Top-20 results for the query *Biennale de la danse de Lyon*. The clusters of the images are represented by the numbers. Non-relevant images are highlighted in red.

As we can observe from Figure D.3, although some non-relevant images were retrieved, the other images covered 70% of the available clusters and consequently produced a good relevance-diversity balance.

D.4 Thanksgiving Day Parade New York

Figure D.4 presents an example of result for the query with the best precision-diversity balance (highest F1@20). It corresponds to the query *Macy's Thanksgiving Day Parade New York* with 9 out of 13 clusters represented. The effectiveness values are:

- P@20 = 1.0
- CR@20 = 0.69
- F1@20 = 0.82

Figure D.4: Top-20 results for the query *Thanksgiving Day Parade New York*. The clusters of the images are represented by the numbers.

As we can observe from Figure D.4, for this query, no non-relevant image was present. Therefore, with roughly 70% of cluster coverage, it allowed a high relevance-diversity balance.