Neurocomputing 208 (2016) 3–24

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

On interactive learning-to-rank for IR: Overview, recent advances, challenges, and directions

Rodrigo Tripodi Calumby [a,c,*], Marcos André Gonçalves [b], Ricardo da Silva Torres [c]

[a] Department of Exact Sciences, University of Feira de Santana, Avenida Transnordestina, s/n, Novo Horizonte, Feira de Santana, Bahia 44036-900, Brazil
[b] Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil
[c] RECOD Lab, Institute of Computing, University of Campinas, Campinas, Brazil

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form
16 March 2016
Accepted 22 March 2016
Available online 4 June 2016

Keywords:
Interactive retrieval
Learning-to-rank
Relevance feedback
Multimedia retrieval
Effectiveness evaluation
User behavior

ABSTRACT

With the amount and variety of information available on digital repositories, answering complex user needs and personalizing information access became a hard task. Putting the user in the retrieval loop has emerged as a reasonable alternative to enhance search effectiveness and consequently the user experience. Due to the great advances on machine learning techniques, optimizing search engines according to user preferences has attracted great attention from the research and industry communities. Interactively learning-to-rank has greatly evolved over the last decade but it still faces great theoretical and practical obstacles. This paper describes basic concepts and reviews state-of-the-art methods on the several research fields that complementarily support the creation of interactive information retrieval (IIR) systems. By revisiting ground concepts and gathering recent advances, this article also intends to foster new research activities on IIR by highlighting great challenges and promising directions. The aggregated knowledge provided here is intended to work as a comprehensive introduction to those interested in IIR development, while also providing important insights on the vast opportunities of novel research.

1. Introduction

In the last decades we have witnessed the production and sharing of huge amounts of data, boosted by a constantly growing data production rate. Human beings and electronic devices have never generated so much data in such a short time [1]. These factors were promoted with important advances on information technologies related to data capturing, storing, and sharing. Moreover, with the popularization of the Internet and mobile devices, a great portion of previously consumer-only users became prolific data generation sources. Therefore, with so much data around, the information technology industry is challenged to deliver more effective and efficient indexing and searching engines.

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 relied only on metadata as a source for relevance estimation, implying on important issues related to annotation costs and accuracy. Relying on textual annotations is subject to language problems related to synonym and polysemy. With the advances on data processing capabilities, content-based methods 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, sometimes are not able to properly represent data concepts and user preferences, causing the well-known *semantic-gap* problem [2].

Consequently, introducing user perception into retrieval methods became an important asset for effectiveness enhancement and result personalization. One common strategy relies on Relevance Feedback (RF) [3], in which the user interacts with the system by implicitly or explicitly providing relevance assessments for the retrieved items. This information is then explored by systems in order to refine and customize new retrieval results. Hence, by interactively exchanging information with the system, the user allows her preferences to be learned and optimized, improving the search experience.

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

* Corresponding author at: Department of Exact Sciences, University of Feira de Santana, Avenida Transnordestina, s/n, Novo Horizonte, Feira de Santana, Bahia 44036-900, Brazil.
E-mail addresses: rtcalumby@ecomp.uefs.br (R.T. Calumby), mgoncalv@dcc.ufmg.br (M.A. Gonçalves), rtorres@ic.unicamp.br (R.d.S. Torres).

http://dx.doi.org/10.1016/j.neucom.2016.03.084
0925-2312/

over time. By taking advantage of all available data and explicitly or implicitly collected user preferences, learning-to-rank models [4] leveraged online adaptiveness and consequently improved user search experience.

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 last decades [5,6], 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.

In the effort for jointly exploring several information related sciences (information retrieval, machine learning, human-computer interaction, computer vision, data mining, etc.) and boosted by the large-scale data production and sharing, interactive information retrieval (IIR) became a very active research field in the last decade. Moreover, for boosting the user–system knowledge transfer and personalization, recent work has gone beyond simple relevance feedback towards integrating more diverse information and techniques into the interactive search process (see Section 6).

This work 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 publication venues and periods consulted in this work were the following: (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 work that exploits 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 survey, in Fig. 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 1 presents a categorization and representative works on the concepts from Fig. 1.

The remainder of this text is organized as follows. In Section 3, the bibliometric information considering the main recent publications discussed throughout this paper is summarized. Section 2 summarizes the findings of previous overview works on the interactive retrieval field. Section 4 overviews traditional concepts on IIR and recent works. Next, Sections 6 and 7 describe common learning-to-rank strategies for IIR and recent boosting alternatives, respectively. With regard to experimental evaluation and user aspects, Sections 7 and 8 present common and new experimental and modeling theoretical and practical tools. Section 9 illustrates several interactive multimedia retrieval applications. Finally, Sections 10 and 11 describe the main open challenges and promising research directions in IIR and Section 12 presents our final considerations.

2. Previous work

Thomee and Lew [5] 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 2002–2011 (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

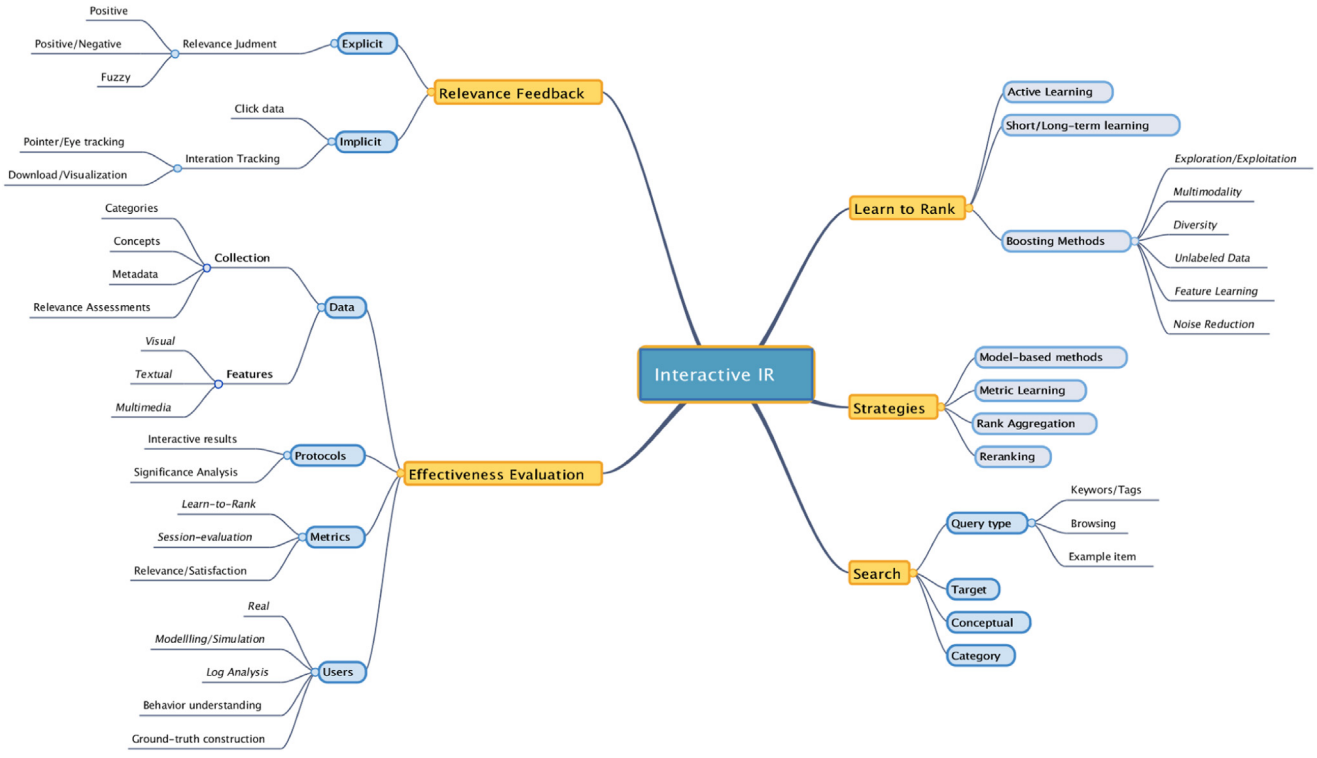


Fig. 1. Conceptual map of the interactive information retrieval field.

Table 1
IIR Concepts and Representative Works.

IIR Concepts	Representative works
Interactivity	
Relevance Feedback	Distance-based learning [7,8], Random walks [9,10], Graph Cuts and Manifold learning [11], Evolutionary methods [7,12–14], Query-point movement [15], Query expansion [16], Query Reformulation [17], Implicit vs. Explicit RF [18].
Active Learning	Most positive/informative samples [19], Uncertainty/Diversity/Density [20], Positive/Negative samples unbalance [16,21].
Short/Long-term learning	Short-term [22,23], Long-term [24], Short and long-term fusion: [16,24,25]
Learning strategies	
Classification-based methods	SVM [11,12,20,21,26], Evolutionary algorithms [12,14,23,27], Logistic Regression [28], Optimum-path forest [19]
Metric learning	Kernel Combination [29], Similarity function optimization [13,14], Features and components weighs adjustment [30]
Rank Aggregation	Ranked-lists fusion [31]
Reranking	Multi-instance learning [32], Reinforcement learning [33]
Learning clues	
Exploration × Exploitation	Interleaving [34,23], Redundancy minimization [35], Exploration–exploitation nesting [36]
Diversity	Dynamic ranked retrieval [28], learn from diversity [27]
Unlabeled Data	Heuristic selection [37], Subspace learning [38], Contextual information [22]
Noise Reduction	Feedback samples similarity [39]
Feature Learning	Dynamic visual dictionaries [40,41,26], Adaptive feature space [42]
Multimodality	Multimodal feature space [43], Multi-form image representation [15], Multimodal ranking functions [27]
Experimental evaluation	
Protocols	Rank-shit [36], residual collection [31], freezing [27,44]
Datasets	PICv1 [45]
Measures	Learning-to-rank [46], Session-based [17,47]
User Aspects	User modeling [48,49], Ground-truth generation [50], judgment effort analysis [50,51].

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 [5], 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 [52], example images [53], random selection of images from the database [54], selected segmented regions [55], and outlines [56]. 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 [57]. The results are usually presented as a ranked list of items that may include the best matching images and/or the most informative ones [58]. The interaction with the user continues with feedback that may be provided using different possibilities of relevance levels: positive only [59], positive/negative [60], positive/neutral/negative [61], or multiple/fuzzy relevance levels [62]. The user feedback may also be collected using region selection on images [63] or implicitly, according to user's actions [64]. The input, results, and feedback may include items from different modalities [65] (text, audio, images, etc.) In turn, the development of new interactive interfaces have focused on better collection browsing [66] and results presentation [67], as well as handling multiple query and feedback modalities [68] (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 [69]. 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 work has explored several directions, such as Feature selection and weighting – principal component analysis (PCA) [70], discriminant component analysis [71] or linear discriminant analysis [72]; Manifold learning [73]; Synthetic and pseudo-imagery [74]; Learning Methods – Artificial Neural Networks [75] and SVM [76]; Kernels [77]; Learners combination [78]; and Probabilistic classifiers [79]. Similarly, long-term learning approaches (see Section 4.3) 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 [5], 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 and Allinson [80] presented an overview of relevance feedback-based methods for Content-based Image Retrieval (CBIR).

Different from [5], in [80], the RF were grouped according to two learning models: short-term learning and long-term learning. 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 and Sugimoto [6] overviews 40 years of Interactive Information Retrieval (IIR) evaluation works (1967–2006). 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.

The authors of [6] defined different scopes for the IIR field. 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 [81], the works that fit both descriptions were defined as the classic core of IIR. Such works include experiments conducted for evaluating the engines and also the retrieval interfaces. Although the IIR research evolves based on 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 [6] 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 this survey (see Section 3).

Complementary to [5,6,80], this article reviews IIR concepts, considering the whole information retrieval field, including traditional text-related methods and modern multimedia-oriented proposals (and not only image retrieval, as [5]). We broadly cover recent proposals and current open issues and challenges. We focus on recent work, mostly related to machine learning strategies (e.g., learning-to-rank [82]), considering multiple modes of information. We specially concentrate on the relationship between interactive systems and learning-to-rank methods.

Our work also comprehensively covers the field in terms of theoretical and practical resources available for the interactive retrieval research and development, including datasets and evaluation protocols. Considering the user aspects, this work presents a much deeper discussion on the advances and challenges of making systems adaptive to user preferences and the difficulties of modeling and learning from user interactions. Besides that, we also cover modern proposals on boosting the interaction effectiveness, considering, for instance, multimodal and diversification techniques. In this sense our work updates and supplements previous efforts in summarizing and understanding such a rich research area, which has evolved a lot in the last few years, as we shall see.

3. Bibliometric analysis

Before diving into the main concepts discussed in this article, we provide a brief bibliometric view of the work covered in this survey, 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 2 and 3. We quantitatively analyzed the main target venues by showing their publication distribution in the period.

Figs. 2 and 3 present the number of articles in each of the analyzed conferences and journals, respectively. Similar to the findings in [6], 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 Fig. 2, 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 (Fig. 3), roughly 58% of the papers were concentrated in four venues: MTAP, PRL, PR, and IEEE TIP.

Fig. 4 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 article.

Considering the described works, which were published from 2011 to 2015, Fig. 5 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.

4. Interactive retrieval

According to Kelly and Sugimoto [6], "*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*

Table 2
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 3
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
JASIST	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

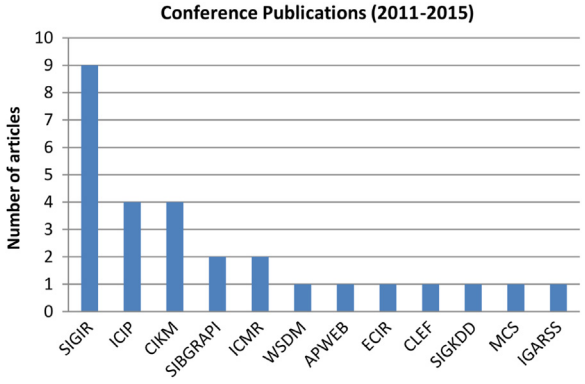


Fig. 2. Number of papers published per conference.

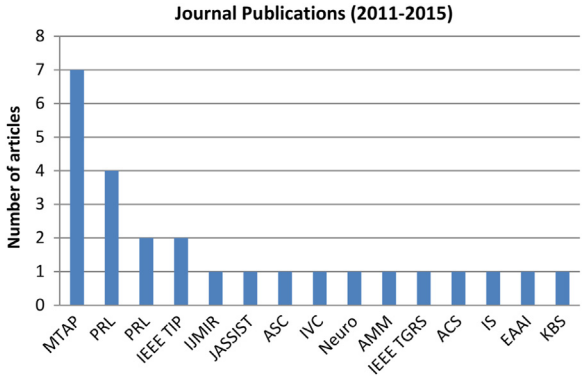


Fig. 3. Number of papers published per journal.

interactions with information systems." In the image retrieval context, Thomee et al. [5] 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 (Section 4.1) and its Implicit and Explicit variations (Section 4.1.1), Active Learning strategies (Section 4.2), and Short-term and Long-term Learning (Section 4.3).

4.1. 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 uses them as positive samples for improving the search results [32].

Recently proposed relevance feedback approaches have relied on several methods such as Random walks [9,10], Genetic Algorithms [7,12], Graph Cuts [11], Manifold learning [11], Distance-based methods (e.g., kNN) [7,8], Genetic Programming (GP) [13,14], Query-point Movement [15], Query Expansion [16], and Query Reformulation [17].

In [9,10], positive and negative feedback samples were used as starting 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 [7], 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 [12], the authors boosted an SVM-based RF approach by optimizing feedback samples' features using genetic algorithms.

In [11], 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 [8], 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 [8] 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.

Other RF-based learning-to-rank proposals and strategies are described throughout this paper. For the interested reader, we refer to Sections 5 and 6 for more details.

4.1.1. 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.

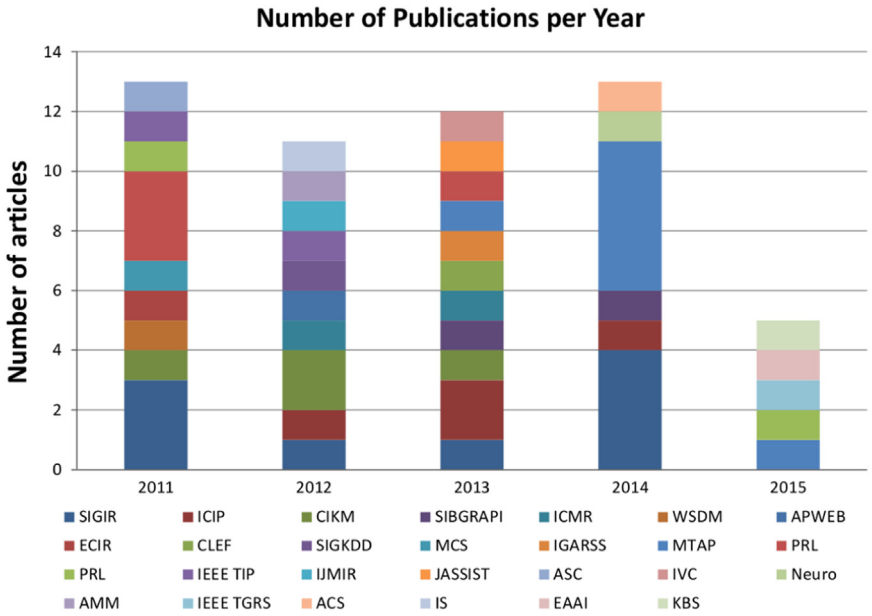


Fig. 4. Number of papers published per year in conferences and journals.

Fig. 5. Tag cloud for the 20 most frequent keywords in recent papers.

Though explicit and implicit RF present different practical challenges and information gain potentials, some work suggest that their combined usage may be beneficial to the overall system effectiveness and user satisfaction. For instance, Zhang et al. [18] 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.

4.2. 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 items. After a few iterations and the labeling of a “proper” amount of informative items, 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 Sections 6.1 and 6.2, respectively.

Different from traditional active learning techniques, which explore user feedback for the most ambiguous (relevant and irrelevant) samples, in [19], an active learning model is proposed for feedback over the most informative samples selected only from the set of relevant images. The method in [19], based on the optimum-path forest classifier [83], 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 [84], with significantly lower processing time.

In the context of remote sensing image retrieval, the work described in [20], which extends [85], 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 criterion based

on the distance between the most informative samples [86]. 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 [21] presented a comprehensive overview of SVM-based relevance feedback and active learning methods and highlighted related open issues. Relevance-based ranking using SVM classifiers, especially 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 [21] 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 [87], 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 [88]). 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 8.2 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 AL "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 bigger than a few thousands of documents.

4.3. 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 are 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 [80], these methods can be categorized regarding how the labeled samples are treated, such as:

(a) *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).
(b) *Two-class* (*one class for positive samples and the other 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.
(c) *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:

(a) *Latent semantic indexing-based techniques*: Among such methods the most commonly used is the Singular Value Decomposition. Chen et al. [89], for instance, explored semantic regions segmented from images and user feedback for constructing the long-term knowledge base.
(b) *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 and Jose [90], for example, proposed an image-context graph for representing the correlation between images, terms, and low-level features.
(c) *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. [91] proposed semantic-correlated clusters constructed based on co-positive-feedback frequency and the co-feedback frequency between the images.
(d) *Feature representation-based methods*: These methods try to improve retrieval effectiveness by properly adjusting relative feature weighs using accumulated feedback information [92].
(e) *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 [93].

LTL methods usually rely on storing pairwise relevance correlation, usually aggregated on an affinity matrix between images or between images and semantic concepts [24]. 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 [25].

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. [16] 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 similarities between unlabeled images to the positively and negatively labeled sets from previous iterations were 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. [25] 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. [24] 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.

5. 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 5.1), Metric learning (Section 5.2), Rank aggregation (Section 5.3), and Reranking methods (Section 5.4).

5.1. Model-based methods

Beyond feature weight adjustment and (multi) query-point movement [43], 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 [11,12,26]. 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 [20,21], 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 4.1, besides the SVM technique, several other machine learning methods have been explored for capturing user preferences such as Genetic Algorithms [12,23] and Programming [14,27], Logistic Regression [28], Optimum-path forest [19], 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 works in [94].

5.2. Metric learning

Analogous to feature components weight learning, when retrieval systems consider multiple features, with early or late fusion approaches [95], users' preferences may be explored for adjusting inter-feature importance and has 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 6.6. For instance, the authors in [29] 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 [96] and LMNN [97]).

For the automatic and adaptive combination of similarity functions from different visual features, the work in [13] 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 [14] 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 ranking functions that automatically adjust the importance of the different modalities and the different similarity measures from each modality according to user preferences expressed through RF.

Alternatively, with a hybrid approach, Shamsi et al. [30] 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.

5.3. 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 is limited research when it comes to IIR approaches.

For exploring relevance feedback over fusion-based improved ranked lists, the work in [31] 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 intermediate lists relative effectiveness 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.

5.4. 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 [98], 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 [32] 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 [99] 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 [32], 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. Different from traditional bag-based methods, in [32], 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 [100,101]. Additionally, the user labeling simulation allowed further effectiveness improvements over the pseudo-feedback method.

Also exploiting image social tags, the work in [33] proposed a multimodal relevance feedback method for image re-ranking 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 [102], VisualRank [103], and clustering-based reranking [100].

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 [98].

6. 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 6.1), diversity promotion (Section 6.2), semi-supervised learning (Section 6.3), noisy feedback reduction (Section 6.4), and other alternatives such as feature learning (Section 6.5), and multimodal feature combination (Section 6.6).

6.1. Exploration and exploitation

Hofmann et al. [34] regard exploitation as using 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 and Fleuret [35], 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. [23] 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 activity. 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 [34].

In this context, the work in [34] 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 [104] 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 for 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 [35] proposed an extension of [105,106], 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 [105,106] 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 the dataset over the iterations even when an image from a relevant region is found. For eliminating such limitations, the work in [35] 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 intent. 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 [23] 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 task, combining exploration–exploitation optimization and diversity promotion, the work in [36] 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 classifiers 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 fifty to ninety judged documents. Moreover, the benefits of the diversity-based method emerged only when around 150 judgments were collected. For clarity purposes, interactive diversity-oriented works are discussed in the next section.

6.2. Diversity

Promoting diversity in retrieval results has emerged as an effective way of maximizing the satisfaction rate in several different scenarios [107,108]. For instance, it has been applied for tackling ambiguous or underspecified queries for which there is no specific answer item or search intent [109]. By covering as many query interpretations as possible, a retrieval system may not provide several relevant answer items for a given interpretation but at least some relevant samples for each possible user intent.

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 [36]. Nevertheless, while it is a very active research field only a few works have investigated the relationship between diversity and user preferences.

Brandt et al. [28] 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. Different from greedy static ranking 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. [110] proposed a set of interactive approaches for text retrieval with diversity promotion using implicit feedback. An MMR-like [111] diversification is applied using the relevance model learned from feedback. Similarly, Calumby et al. [27] introduced a new genetic programming framework for improving relevance feedback session effectiveness on multimodal image retrieval scenarios. For improving the learning models, the relevance feedback was taken over explicitly 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. The authors 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 explicitly diversified results allowed retrieving more relevant items and also in earlier iterations.

6.3. 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 was 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. [37] 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 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 [38] 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 not related concepts. Another discussed issue related not only to SVM-based RF methods, but also generally present on image-based RF schemes is the small group of samples. In order to reduce such problems, the work in [38] 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 methods and traditional SVM approach.

Pedronette et al. [22] 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 [112] 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 combining visual and textual information. The proposed method outperformed a similarity combination function optimization baseline.

6.4. 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 [39] 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 [38], the authors of [39] 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 4.3, Rashedi et al. [24] also achieved noisy feedback reduction by jointly fusing short-term and long-term learning models.

6.5. Feature learning

An effective approach for CBIR is the construction of dictionaries of visual features [69,113]. 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 [40]. In a further step, Gosselin et al. [41] 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 [26] 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 [41], Wang et al. [42] 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 [114], 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.

6.6. 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, the 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. [43] enhanced a multimodal object retrieval system by incorporating relevance feedback. For fast and effective retrieval, the information from all objects' modalities was 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 [15], Guldogan et al. 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 weights of the different forms, and consequently their contribution to the results 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. [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.

7. 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 [48]. Different from the document relevance theories, e.g., Probability Ranking Principle (PRP) [115], user–system interaction theoretic models for describing, predicting, and explaining search behavior are still an open issue [49]. Nevertheless, an interesting extension of the PRP theory for interactive retrieval was presented by Fuhr [116]. Moreover, some works [34,39,47,117] 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 [49].

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 [50]. As described in [48,49], 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. [48] 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 [49].

In [49] (extended from [118]), Azzopardi proposed a more realistic theoretical modeling for search behavior understanding and prediction based on the search economic theory [119]. As proposed in [119] and evaluated in [118], 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 [49], 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 considering 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 [51], on the relevance assessment effort evaluation for text retrieval, Halvey and Villa [50] 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 [50,51] could have positive impact on user behavior modeling, such as in [49,118], by simulating and assessing different user patterns considering different topic difficulties and semantics, which should also be incorporated into effectiveness evaluation measures.

8. 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 [120].

Kelly et al. [121] 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 variations; (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 (1967–2006) of [6], 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 [6] and also in the recent works discussed here, specially considering evaluation measures and statistical analysis methods.

Considering IIR research, Thomee and Lew [5] 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 on 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 [6,119,120] and practical [17,47,120] 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 [120].

8.1. 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 user approach (real or modeled) – Sections 7 and 8.1.1, search type (target, category, achievement-based, etc.) – Section 8.1.2, result evaluation protocol (Section 8.1.3), training and testing datasets (Section 8.2), effectiveness measures (Sections 8.3–8.5), number of queries/topics, number of items retrieved per iteration, number of feedback samples, relevance assessment grades, among others. In active learning studies, it is also important to establish the number of learning iterations before the user has access to final results and how the samples are selected for user assessment.

8.1.1. User modeling

Considering the user complexity presented in Section 7, we observed that IIR works are still conducted with non-standard modeling but some groups of approaches can be highlighted: (i) Perfect user simulation (with classes/categories information or relevance assessments [21,23,27]; (ii) Probabilistic modeling [18,39,48]; (iii) Click model [34,104]; (iv) Log analysis [16,24,49]; and (v) Real users [33,35,47,50,117];

8.1.2. 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 [49], target search [122], conceptual search [32,38], category search [15,42,123], 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 [35].

8.1.3. 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 [124] 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" [125]. Alternatively, with the collection reranking procedure, all items in the target dataset are reranked in future iterations. In turn, with a residual collection strategy [126], only the items not previously seen are presented in further iterations, no matter if they were judged relevant or not. Differently, the freezing approach [44] keeps the relevant items in the same rank positions they were firstly retrieved. As a variation, the full freezing protocol [127] 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 [127] 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 [126]. 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 [9,36], collection reranking [18,25], residual collection [7,31,38], and full freezing [27].

Table 4
Datasets explored in recent IIR works.

Type[a]	Dataset	Size[b]
Text	Letor 3.0 [136] and 4.0 [137]	Multiple datasets
Text	TREC 1 [138] and 2 [139] Ad hoc tracks	742,611 docs
Text	TREC 3 Ad hoc track [140]	741,856 docs
Text	TREC 6 Ad hoc track [141]	556,077 docs
Text	TREC 7 [142] and 8 [143] Ad hoc tracks	528,155 docs
Text	TREC 9 [144] and 10 [145] Ad hoc tracks	Multiple datasets
Text	TREC 9 Query Track [146]	510,000 docs
Text	TREC Filtering Track 2002 [147]	800,000 docs
Text	TREC HARD Track 2005 [148]	1,033,461 docs
Text	TREC 6 [149], 7 [150], and 8 [151] Interactive Tracks	210,158 articles
Text	TREC Microblog Track 2012 [152]	16mi tweets
Text	TREC Microblog Track 2013 [153]	243mi tweets
Text	TREC Robust Topics 2005 [154]	1,033,461 docs
Text	ClueWeb09[c]	1.04bi web pages
Image	Aerial orthoimagery [155]	600 (6)
Image	Brodatz [156]	1776 (111)
Image	Caltech-101 [128]	8677 (101)
Image	Caltech-256 [129]	30,607 (256)
Image	Coil-100 [157]	7200 (100)
Image	Corel [158]	circa 80,000 (800)
Image	ImageCLEF Photographic Retrieval Task 2007 [159] and 2008 [160]	20,000
Image	IRMA (Medical Collection)[d]	Multiple datasets
Image	MIRFlickr [161]	25,000 (1386)
Image	MPEG-7 Part B [162]	1400 (70)
Image	MSRCORID[e]	4320 (20)
Image	NUS-WIDE [163]	269,648 (81)
Image	Oxford Flower17 [164]	8189 (103)
Image	PASCAL VOC 2006 [165]	2618 (10)
Image	PASCAL VOC 2007 [166]	9963 (20)
Image	PASCAL VOC 2012 [167]	11,530 (20)
Image	University of Washington[f]	1109 (20)
Video	MediaEval Video Genre Tagging Task 2012 [168]	15,000 (26)
Video	TRECVID 2005 [169]	169 h of video
Video	TRECVID 2006 [170]	328 h of video
Video	TRECVID 2007 [171]	200 h of video
Video	TRECVID 2008 [172]	253 h of video
Video	TRECVID 2009 [173]	410 h of video

[a] (Main) data type.
[b] Number of classes/concepts/tags.
[c] ⟨http://lemurproject.org/clueweb09/⟩ (as of September 18, 2015).
[d] ⟨http://http://www.irma-project.org/⟩ (as of September 24, 2015).
[e] ⟨http://research.microsoft.com/en-us/projects/objectclassrecognition/⟩ (as of September 24, 2015).
[f] ⟨http://imagedatabase.cs.washington.edu/⟩ (as of September 24, 2015).

8.2. Datasets

A summary and brief description of the datasets used in recent interactive retrieval works are presented in Table 4. 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 4, 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 [45], 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 [45] 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 [128] and 256 [129], MIR Flickr [130], MSRA-MM [131], and Social Event Detection Task [132] (extended in [133]). None of these collections have 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. 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:

(a) 5555 personal photos from 19 photographers;
(b) demographic metadata of the photographers and assessors, which allows persona creations for user simulation;
(c) EXIF data, GPS coordinates (automatically or manually included), and city or country names;
(d) tags: indoor/outdoor, day/night, altered, blurred etc.;
(e) number of people in the photo;
(f) event information/ground-truth using WordNet [134];
(g) 130 fully assessed topics from different domains;
(h) 32 topics with graded relevance assessments (0 – irrelevant to 3 – fully relevant);
(j) ideal DCG curves [135];
(k) 18 low-level visual features.

8.3. Effectiveness measures

The historical analysis in [6] 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 [174] and TREC-like[1] 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 5.

Table 5
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

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 $\times$ Recall, and NDCG. Several works have computed these measures in a per-iteration basis, e.g., Recall@N $\times$ Iteration. Alternatively, several studies applied not so common measures such as R-Precision (in [36]), BPREF, and GMAP (in [27]), and the number of relevant items per iteration (in [13]). Some measures were also reported for evaluating results' diversity, such as Intent-aware measures (in [28]) 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 [35] and session time in [50]. Some measures related to learn-to-rank and session-based retrieval are discussed in Sections 8.4 and 8.5, respectively.

8.4. Learn-to-rank evaluation and measures

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) [46].

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 [46] has experimentally shown that, under certain circumstances, it is not the case. The authors in [46] 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 [46] 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 [175].

[1] http://trec.nist.gov (as of February 21, 2016).

8.5. Session-based effectiveness

As stated in [122], 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 [176], 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 [177,178]. Therefore, as pointed out in [179], 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 that 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 [47] 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:

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

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 [135], *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 [47], this new measure is considered more suitable for session-based evaluation for:

(a) considering items in equivalent rank position more relevant when returned for an earlier query;
(b) using smooth discount for document-based gains and query sequence effectiveness importance; and
(c) is 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 [17] argued that using an interactive evaluation paradigm can better assess the real user experience but previously proposed measures, e.g., instance recall [149] and *nsDCG* [47], 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 behaviors 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 [17]. 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 the following equation:

$$sP = \frac{rR@j,k}{k} \tag{2}$$

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 Eq. (2) 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 [17] also formulated the session-based measures according to the expected retrieval effectiveness (Eq. (3)) and not the maximum values, as used for interpolated measures:

$$esM = \sum_{w \in W} P(w) M_w \tag{3}$$

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 [17].

For effectiveness prediction, by describing session-based features for queries, the authors in [180] 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 paper and for more information the reader is directed to the works in [121,177,181].

8.5.1. Significance analysis

For strict result analysis and the construction of an adequate comparison between different retrieval systems of 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 [34,38,39]. 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 [182].

9. Multimedia retrieval and applications

In the works described in this paper, 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 [129], Corel [158], and Pascal VOC [167] datasets. Nevertheless, some interesting works on interactive retrieval have been conducted for medical images [39], remote sensing images [20], soccer teams [22], fish images [13], and flowers [29].

In [5], 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 [20,85], interactive strategies with active learning were proposed for remote sensing images retrieval on earth observation data archives.

Wei and Yang [117] highlighted some important and interdependent factors related to interactive video retrieval: (i) the exploration–exploitation dilemma (see Section 6.1); (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, Wei and Yang [117] proposed an integrated framework for video retrieval with relevance feedback based on an active learning model (see Section 4.2) 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 [117] integrates several research alternatives described in the previous sections for enhancing retrieval effectiveness. Another interesting alternative intrinsically related to video retrieval and analysis is the combination of multiple features [183,184], multi-view information [185], and also multiple information modalities (see Section 6.6). For instance, Mironica et al. [123] proposed a RF method with the combination of several visual, audio, and textual features from videos.

10. 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 [5], the main challenges related to interactive search are:

(a) *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.
(b) *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.
(c) *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;
(d) *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.

For simplicity we grouped some of the challenges into the following:

(a) *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.
(b) *Data*: With the ever increasing availability of social, linked, log, and mainly unlabeled data (see Section 6.3), 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.
(c) *Learning*: The effectiveness of search systems in capturing real user intents 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 6.1, automatically adjusting the exploration–exploitation trade-off is still an open issue and may benefit from advanced learning models.
(d) *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.
(e) *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.

10.1. 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 [120], "*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 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:

(a) The development of effectiveness measures that are more informative and better suited for learn-to-rank methods.
(b) The proposal of better interactivity cost functions to evaluate search strategies and user effort on retrieval sessions.
(c) The development of better log analysis methods, click models, and user models considering reformulation understanding, stopping criteria, and erroneous feedback simulation.
(d) The performance of experiments with real-life settings. Conducting real-user study 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.

11. Promising directions

According to [5], some promising directions on improving interactive search systems rely on exploring:

(a) Question Answering Paradigm focused on multi-modality and cross-modality. For instance, these aspects may trigger effectiveness improvements for several applications like Multimedia Answering [186].
(b) 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.
(c) Exploring external sources: Interactive systems can explore additional image collections and knowledge sources for improving retrieval effectiveness.
(d) Social interaction for system's optimization through collaborative filtering.

As clearly observed in [6], and exposed in this paper, 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 learn-to-rank methods which deserve further investigation and development effort:

(a) Exploring unlabeled data and semi-supervised methods, for reducing labeling effort, attenuating the cold start problem, and consequently classifier effectiveness.
(b) Differentiating positive and negative samples treatment on the learning process for their different representativeness in relation to real data distribution.
(c) Integrating advanced procedures for handling complex queries [187], which embed multiple search concepts, specially considering high-level semantic requirements and the relationship among such concepts.
(d) Exploring learning boosting alternatives such as diversity promotion for handling ambiguous, multi-intent, overview, or underspecified queries.
(e) Using reinforcement learning methods for combining multiple feature modalities or even multiple learning strategies such as active learning and exploration/exploitation.
(f) Analyzing user behavior impacts on search tasks, which will produce information for the development of better generalization models and more realistic user models.
(g) Leveraging long-term learning and collaborative retrieval for effectiveness and efficiency improvement.
(h) Using graded relevance assessments as a way to improve ground-truth quality and maximize feedback information. For conducting user-centric evaluations the work in [6] also suggests using nDCG [135] 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
(i) 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.

12. Final considerations

In this article, we reviewed many aspects related to interactive learn-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 pose 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 the latest works in the field have demonstrated, IIR research has been directed towards integrating as much information as possible, fusing multiple data sources and analytical methods, which allowed targeting customized user experiences. 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 user standing is one of the most important factors of the interaction loop, it 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 research challenge, it opens opportunities due to 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 paper 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.

Acknowledgments

The authors thank CAPES (Grant no. P-4388/2010), CNPq (Grant no. 140977/2012-0), FAPESP, and FAPEMIG for supporting this work.

References

[1] E. Letouzé, J. Jütting, Official Statistics, Big Data and Human Development, Technical Report, Data-Pop Alliance, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute, Paris, 2015.
[2] R.C. Veltkamp, M. Tanase, A survey of content-based image retrieval systems, in: Content-Based Image and Video Retrieval, Kluwer, Norwell, Massachusetts, USA, 2002, pp. 47–101.
[3] X.S. Zhou, T.S. Huang, Relevance feedback in image retrieval: a comprehensive review, Multimed. Syst. 8 (6) (2003) 536–544.
[4] T.-Y. Liu, Learning to rank for information retrieval, Found. Trends Inf. Retr. 3 (3) (2009) 225–331.
[5] B. Thomee, M. Lew, Interactive search in image retrieval: a survey, Int. J. Multimed. Inf. Retr. 1 (2) (2012) 71–86.
[6] D. Kelly, C.R. Sugimoto, A systematic review of interactive information retrieval evaluation studies, 1967–2006, J. Am. Soc. Inf. Sci. Technol. 64 (4) (2013) 745–770.
[7] M. Arevalillo-Herráez, F.J. Ferri, S. Moreno-Picot, Distance-based relevance feedback using a hybrid interactive genetic algorithm for image retrieval, Appl. Soft Comput. 11 (2) (2011) 1782–1791.
[8] M. Arevalillo-Herráez, F.J. Ferri, An improved distance-based relevance feedback strategy for image retrieval, Image Vis. Comput. 31 (10) (2013) 704–713.
[9] S. Rota Bulò, M. Rabbi, M. Pelillo, Content-based image retrieval with relevance feedback using random walks, Pattern Recognit. 44 (9) (2011) 2109–2122.
[10] M.K. Kundu, M. Chowdhury, S.R. Bulò, A graph-based relevance feedback mechanism in content-based image retrieval, Knowl.-Based Syst. 73 (2015) 254–264.
[11] L. Zhang, S. Liu, Z. Wang, W. Cai, Y. Song, D. D. Feng, Graph cuts based relevance feedback in image retrieval, in: IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15–18, 2013, pp. 4358–4362.
[12] A. Irtaza, M. Jaffar, M. Muhammad, Content based image retrieval in a web 3.0 environment, Multimed. Tools Appl. (2013) 1–18.
[13] C. Ferreira, J. Santos, R. da S. Torres, M. Gonalves, R. Rezende, W. Fan, Relevance feedback based on genetic programming for image retrieval, Pattern Recognit. Lett. 32 (1) (2011) 27–37 (Image Processing, Computer Vision and Pattern Recognition in Latin America).
[14] R.T. Calumby, R.d.S. Torres, M.A. Gonçalves, Multimodal retrieval with relevance feedback based on genetic programming, Multimed. Tools Appl. 69 (3) (2014) 991–1019.
[15] E. Guldogan, T. Olsson, E. Lagerstam, M. Gabbouj, Instance based personalized multi-form image browsing and retrieval, Multimed. Tools Appl. 71 (3) (2014) 1087–1104.
[16] J. Wu, H. Shen, Y.-D. Li, Z.-B. Xiao, M.-Y. Lu, C.-L. Wang, Learning a hybrid similarity measure for image retrieval, Pattern Recognit. 46 (11) (2013) 2927–2939.
[17] E. Kanoulas, B. Carterette, P.D. Clough, M. Sanderson, Evaluating multi-query sessions, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2011, pp. 1053–1062.
[18] Y. Zhang, W. Li, Z. Mo, T. Zhao, J. Zhang, An adaptive-weight hybrid relevance feedback approach for content based image retrieval, in: 2013 20th IEEE International Conference on Image Processing (ICIP), 2013, pp. 3977–3981.
[19] A.T. da Silva, A.X. Falcão, L.P. Magalhães, Active learning paradigms for CBIR systems based on optimum-path forest classification, Pattern Recognit. 44 (12) (2011) 2971–2978.
[20] B. Demir, L. Bruzzone, A novel active learning method in relevance feedback for content-based remote sensing image retrieval, IEEE Trans. Geosci. Remote Sens. 53 (5) (2015) 2323–2334.
[21] X.-Y. Wang, H.-Y. Yang, Y.-W. Li, W.-Y. Li, J.-W. Chen, A new SVM-based active feedback scheme for image retrieval, Eng. Appl. Artif. Intell. 37 (2015) 43–53.
[22] D. Guimaraes Pedronette, R. Calumby, R. da S. Torres, Semi-supervised learning for relevance feedback on image retrieval tasks, in: 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2014, pp. 243–250.
[23] M. Arevalillo-Herrez, F.J. Ferri, S. Moreno-Picot, Improving distance based image retrieval using non-dominated sorting genetic algorithm, Pattern Recognit. Lett. 53 (2015) 109–117.
[24] E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, Information fusion between short term learning and long term learning in content based image retrieval systems, Multimed. Tools Appl. 74 (11) (2015) 3799–3822.
[25] Z. Xiao, X. Qi, Complementary relevance feedback-based content-based image retrieval, Multimed. Tools Appl. 73 (3) (2014) 2157–2177.
[26] P.-H. Gosselin, Online kernel learning for interactive retrieval in dynamic image databases, in: 2012 19th IEEE International Conference on Image Processing (ICIP), 2012, pp. 1921–1924.
[27] R.T. Calumby, R.d.S. Torres, M.A. Gonçalves, Diversity-Driven Learning for Multimodal Image Retrieval with Relevance Feedback, 2014, pp. 2197–2201.
[28] C. Brandt, T. Joachims, Y. Yue, J. Bank, Dynamic ranked retrieval, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2011, pp. 247–256.
[29] F. Yan, K. Mikolajczyk, J. Kittler, Multiple kernel learning via distance metric learning for interactive image retrieval, in: C. Sansone, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 6713, Springer, Berlin, Heidelberg, 2011, pp. 147–156.
[30] A. Shamsi, H. Nezamabadi-pour, S. Saryazdi, A short-term learning approach based on similarity refinement in content-based image retrieval, Multimed. Tools Appl. 72 (2) (2014) 2025–2039.
[31] E. Rabinovich, O. Rom, O. Kurland, Utilizing relevance feedback in fusion-based retrieval, in: Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval, ACM, New York, NY, USA, 2014, pp. 313–322.
[32] L. Duan, W. Li, I. Tsang, D. Xu, Improving web image search by bag-based reranking, IEEE Trans. Image Process. 20 (11) (2011) 3280–3290.
[33] J. Li, Q. Ma, Y. Asano, 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, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 399–410.
[34] K. Hofmann, S. Whiteson, M. de Rijke, Balancing exploration and exploitation in learning to rank online, in: Proceedings of the 33rd European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 251–263.
[35] N. Suditu, F. Fleuret, Iterative relevance feedback with adaptive exploration/exploitation trade-off, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2012, pp. 1323–1331.
[36] C. Li, Y. Wang, P. Resnick, 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, ACM, New York, NY, USA, 2014, pp. 163–172.
[37] Q. Xing, Y. Zhang, L. Zhang, On bias problem in relevance feedback, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2011, pp. 1965–1968.
[38] L. Zhang, L. Wang, W. Lin, Semisupervised biased maximum margin analysis for interactive image retrieval, IEEE Trans. Image Process. 21 (4) (2012) 2294–2308.
[39] Y. Huang, H. Huang, J. Zhang, A noisy-smoothing relevance feedback method for content-based medical image retrieval, Multimed. Tools Appl. 73 (3) (2014) 1963–1981.
[40] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA, 2009, pp. 689–696.
[41] P.H. Gosselin, F. Precioso, S. Philipp-Foliguet, Incremental kernel learning for active image retrieval without global dictionaries, Pattern Recognit. 44 (10) (2011) 2244–2254.
[42] X.-Y. Wang, Y.-W. Li, H.-Y. Yang, J.-W. Chen, An image retrieval scheme with relevance feedback using feature reconstruction and svm reclassification, Neurocomputing 127 (2014) 214–230.
[43] A. Axenopoulos, S. Manolopoulou, P. Daras, Optimizing multimedia retrieval using multimodal fusion and relevance feedback techniques, in: K. Schoeffmann, B. Merialdo, A. Hauptmann, C.-W. Ngo, Y. Andreopoulos, C. Breiteneder (Eds.), Advances in Multimedia Modeling, Lecture Notes in Computer Science, vol. 7131, Springer, Berlin, Heidelberg, 2012, pp. 716–727.
[44] H. Keskustalo, K. Järvelin, A. Pirkola, Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value, Inf. Retr. 11 (3) (2008) 209–228.
[45] 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, ACM, New York, NY, USA, 2012, pp. 29:1–29:8.
[46] E. Yilmaz, S. Robertson, On the choice of effectiveness measures for learning to rank, Inf. Retr. 13 (3) (2010) 271–290.
[47] K. Järvelin, S.L. Price, L.M.L. Delcambre, M.L. Nielsen, Discounted cumulated gain based evaluation of multiple-query ir sessions, in: Proceedings of 30th European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 4–15.

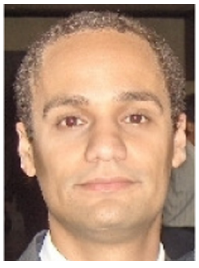
[48] F. Baskaya, H. Keskustalo, K. Järvelin, Modeling behavioral factors ininteractive information retrieval, in: Proceedings of the 22nd ACM International Conference on Conference on Information Knowledge Management, ACM, New York, NY, USA, 2013, pp. 2297–2302.
[49] L. Azzopardi, Modelling interaction with economic models of search, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, New York, NY, USA, 2014, pp. 3–12.
[50] M. Halvey, 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, ACM, New York, NY, USA, 2014, pp. 887–890.
[51] R. Villa, 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, ACM, New York, NY, USA, 2013, pp. 765–768.
[52] M. Kherfi, D. Brahmi, D. Ziou, Combining visual features with semantics for a more effective image retrieval, in: Proceedings of the 17th International Conference on Pattern Recognition, vol. 2, 2004, pp. 961–964.
[53] G. Aggarwal, T. Ashwin, S. Ghosal, An image retrieval system with automatic query modification, IEEE Trans. Multimed. 4 (2) (2002) 201–214.
[54] B. Thomee, M.J. Huiskes, E. Bakker, M.S. Lew, Deep exploration for experiential image retrieval, in: Proceedings of the 17th ACM International Conference on Multimedia, ACM, New York, NY, USA, 2009, pp. 673–676.
[55] C.-C. Chiang, M.-H. Hsieh, Y.-P. Hung, 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, E. Bakker (Eds.), Image and Video Retrieval, Lecture Notes in Computer Science, vol. 3568, Springer, Berlin, Heidelberg, 2005, pp. 487–496.
[56] J. Amores, N. Sebe, P. Radeva, T. Gevers, A. Smeulders, Boosting contextual information in content-based image retrieval, in: Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, ACM, New York, NY, USA, 2004, pp. 31–38.
[57] J.M. Torres, D. Hutchison, L.P. Reis, Semantic image retrieval using region-based relevance feedback, in: S. Marchand-Maillet, E. Bruno, A. Nürnberger, M. Detyniecki (Eds.), Adaptive Multimedia Retrieval: User, Context, and Feedback, Lecture Notes in Computer Science, vol. 4398, Springer, Berlin, Heidelberg, 2007, pp. 192–206.
[58] M. Huiskes, Image searching and browsing by active aspect-based relevance learning, in: H. Sundaram, M. Naphade, J. Smith, Y. Rui (Eds.), Image and Video Retrieval, Lecture Notes in Computer Science, vol. 4071, Springer, Berlin, Heidelberg, 2006, pp. 211–220.
[59] X. Jin, J.C. French, Improving image retrieval effectiveness via multiple queries, in: Proceedings of the 1st ACM International Workshop on Multimedia Databases, ACM, New York, NY, USA, 2003, pp. 86–93.
[60] C. Zhang, 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, E. Bakker (Eds.), Image and Video Retrieval, Lecture Notes in Computer Science, vol. 3568, Springer, Berlin, Heidelberg, 2005, pp. 194–204.
[61] J. Yang, Q. Li, Y. Zhuang, Image retrieval and relevance feedback using peer indexing, in: Proceedings of 2002 IEEE International Conference on Multimedia and Expo, 2002. ICME '02, vol. 2, 2002, pp. 409–412.
[62] X. Hunag, S.-C. Chen, M.-L. Shyu, Incorporating real-valued multiple instance learning into relevance feedback for image retrieval, in: Proceedings of 2003 International Conference on Multimedia and Expo, 2003. ICME '03, vol. 1, 2003, pp. I-321–4.
[63] D. Tran, S. Pamidimukkala, P. Nguyen, Relevance-feedback image retrieval based on multiple-instance learning, in: Seventh IEEE/ACIS International Conference on Computer and Information Science, 2008, ICIS 08, 2008, pp. 597–602.
[64] E. Cheng, F. Jing, L. Zhang, A unified relevance feedback framework for web image retrieval, IEEE Trans. Image Process. 18 (6) (2009) 1350–1357.
[65] J. Meng, J. Yuan, Y. Jiang, N. Narasimhan, V. Vasudevan, Y. Wu, Interactive visual object search through mutual information maximization, in: Proceedings of the International Conference on Multimedia, ACM, New York, NY, USA, 2010, pp. 1147–1150.
[66] B. Thomee, M. Huiskes, E. Bakker, M. Lew, An exploration-based interface for interactive image retrieval, in: Proceedings of 6th International Symposium on Image and Signal Processing and Analysis, 2009, ISPA 2009, 2009, pp. 188–193.
[67] R. Wang, S.J. McKenna, J. Han, High-entropy layouts for content-based browsing and retrieval, in: Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, New York, NY, USA, 2009, pp. 16:1–16:8.
[68] J. Urban, J.M. Jose, Evaluating a workspace's usefulness for image retrieval, Multimed. Syst. 12 (4–5) (2007) 355–373.
[69] F. Jurie, B. Triggs, Creating efficient codebooks for visual recognition, in: Tenth IEEE International Conference on Computer Vision, 2005, ICCV 2005, vol. 1, 2005, pp. 604–610.
[70] A. Franco, A. Lumini, D. Maio, A new approach for relevance feedback through positive and negative samples, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, vol. 4, 2004, pp. 905–908.
[71] S. Hoi, W. Liu, M. Lyu, W.-Y. Ma, Learning distance metrics with contextual constraints for image retrieval, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2072–2078.
[72] R. Huang, Q. Liu, H. Lu, S. Ma, Solving the small sample size problem of lda, in: Proceedings of 16th International Conference on Pattern Recognition, vol. 3, 2002, pp. 29–32.
[73] W. Bian, D. Tao, Biased discriminant euclidean embedding for content-based image retrieval, IEEE Trans. Image Process. 19 (2) (2010) 545–554.
[74] B. Thomee, M. Huiskes, E. Bakker, M. Lew, Using an artificial imagination for texture retrieval, in: 19th International Conference on Pattern Recognition, 2008, pp. 1–4.
[75] K. Wu, K.-H. Yap, L.-P. Chau, Region-based image retrieval using radial basis function network, in: 2006 IEEE International Conference on Multimedia and Expo, 2006, pp. 1777–1780.
[76] J. Zhang, L. Ye, Content based image retrieval using unclean positive examples, IEEE Trans. Image Process. 18 (10) (2009) 2370–2375.
[77] H. Xie, V. Andreu, 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, ACM, New York, NY, USA, 2006, pp. 23–32.
[78] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1088–1099.
[79] T. Amin, M. Zeytinoglu, L. Guan, Application of Laplacian mixture model to image and video retrieval, IEEE Trans. Multimed. 9 (7) (2007) 1416–1429.
[80] J. Li, N. Allinson, Relevance feedback in content-based image retrieval: a survey, in: M. Bianchini, M. Maggini, L.C. Jain (Eds.), Handbook on Neural Information Processing, Intelligent Systems Reference Library, vol. 49, Springer, Berlin, Heidelberg, 2013, pp. 433–469.
[81] D. Kelly, Methods for evaluating interactive information retrieval systems with users, Found. Trends Inf. Retr. 3 (1–2) (2009) 1–224.
[82] F.F. Faria, A. Veloso, H.M. Almeida, E. Valle, R.d.S. Torres, M.A. Gonçalves, W. Meira, Jr., Learning to rank for content-based image retrieval, in: Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10, ACM, New York, NY, USA, 2010, pp. 285–294.
[83] J.P. Papa, A.X. Falcão, C.T.N. Suzuki, Supervised pattern classification based on optimum-path forest, Int. J. Imaging Syst. Technol. 19 (2) (2009) 120–131.
[84] S. Tong, E. Chang, Support vector machine active learning for image retrieval, in: Proceedings of the Ninth ACM International Conference on Multimedia, ACM, New York, NY, USA, 2001, pp. 107–118.
[85] B. Demir, L. Bruzzone, An effective active learning method for interactive content-based retrieval in remote sensing images, in: 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2013, pp. 4356–4359.
[86] M. Ferecatu, N. Boujemaa, Interactive remote-sensing image retrieval using active relevance feedback, IEEE Trans. Geosci. Remote Sens. 45 (4) (2007) 818–826.
[87] R.M. Silva, M.A. Gonçalves, A. Veloso, A two-stage active learning method for learning to rank, J. Assoc. Inf. Sci. Technol. 65 (1) (2014) 109–128.
[88] R.M. Silva, M.A. Gonçalves, A. Veloso, Rule-based active sampling for learning to rank, in: Proceedings of Machine Learning and Knowledge Discovery in Databases – European Conference, ECML PKDD 2011, Athens, Greece, September 5–9, Part III, 2011, pp. 240–255.
[89] X. Chen, C. Zhang, S.-C. Chen, 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, 2005, 37–45.
[90] J. Urban, 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, ACM, New York, NY, USA, 2006, pp. 117–126.
[91] J. Han, K. Ngan, M. Li, H.-J. Zhang, A memory learning framework for effective image retrieval, IEEE Trans. Image Process. 14 (4) (2005) 511–524.
[92] M. Cord, P. Gosselin, Image retrieval using long-term semantic learning, in: IEEE International Conference on Image Processing, 2006, pp. 2909–2912.
[93] S. Hoi, M. Lyu, R. Jin, A unified log-based relevance feedback scheme for image retrieval, IEEE Trans. Knowl. Data Eng. 18 (4) (2006) 509–524.
[94] T.-Y. Liu, Learning to Rank for Information Retrieval, Springer Science & Business Media, Beijing, People's Republic of China, 2011.
[95] C.G.M. Snoek, M. Worring, A.W.M. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, ACM, New York, NY, USA, 2005, pp. 399–402.
[96] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, in: Advances in Neural Information Processing Systems (NIPS), 2004, p. 41.
[97] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[98] T. Mei, Y. Rui, S. Li, Q. Tian, Multimedia search reranking: a literature survey, ACM Comput. Surv. 46 (3) (2014) 38:1–38:38.
[99] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems, 2002, pp. 561–568.
[100] W.H. Hsu, L.S. Kennedy, S.-F. Chang, Video search reranking via information bottleneck principle, in: Proceedings of the 14th Annual ACM International Conference on Multimedia, ACM, New York, NY, USA, 2006, pp. 35–44.
[101] Z.-H. Zhou, H.-B. Dai, Exploiting image contents in web search, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence,

Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007, pp. 2928–2933.
[102] L. Zhang, F. Lin, B. Zhang, Support vector machine learning for image retrieval, in: 2001 International Conference on Proceedings of Image Processing, vol. 2, 2001, pp. 721–724.
[103] Y. Jing, S. Baluja, Visualrank: applying pagerank to large-scale image search, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 1877–1890.
[104] F. Guo, C. Liu, Y.M. Wang, Efficient multiple-click models in web search, in: Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA, 2009, pp. 124–131.
[105] M. Ferecatu, D. Geman, Interactive search for image categories by mental matching, in: IEEE 11th International Conference on Computer Vision, 2007, ICCV 2007, 2007, pp. 1–8.
[106] M. Ferecatu, D. Geman, A statistical framework for image category search from a mental picture, IEEE Trans. Pattern Anal. Mach. Intell 31 (6) (2009) 1087–1101.
[107] S. Vargas, P. Castells, D. Vallet, Explicit relevance models in intent-oriented information retrieval diversification, in: ACM SIGIR, 2012, pp. 75–84.
[108] C. Kofler, M. Larson, A. Hanjalic, Intent-aware video search result optimization, IEEE Trans. Multimed. 16 (5) (2014) 1421–1433.
[109] M.R. Vieira, H.L. Razente, M.C.N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina, V.J. Tsotras, On query result diversification, in: IEEE ICDE, 2011, pp. 1163–1174.
[110] K. Raman, P. Shivaswamy, T. Joachims, Online learning to diversify from implicit feedback, in: ACM SIGKDD, 2012, pp. 705–713.
[111] J. Carbonell, J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, in: ACM SIGIR, 1998, pp. 335–336.
[112] D.C.G. Pedronette, R. da, S. Torres, Exploiting pairwise recommendation and clustering strategies for image re-ranking, Inf. Sci. 207 (2012) 19–34.
[113] Y. Yan, Y. Yang, D. Meng, G. Liu, W. Tong, A. Hauptmann, N. Sebe, Event oriented dictionary learning for complex event detection, IEEE Trans. Image Process. 24 (6) (2015) 1867–1878.
[114] D. Tao, X. Tang, X. Li, Which components are important for interactive image searching? IEEE Trans. Circuits Syst. Video Technol. 18 (1) (2008) 3–11.
[115] S.E. Robertson, The Probability Ranking Principle in IR, in: Readings in Information Retrieval, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 281–286.
[116] N. Fuhr, A probability ranking principle for interactive information retrieval, Inf. Retr. 11 (3) (2008) 251–265.
[117] X.-Y. Wei, Z.-Q. Yang, Coaching the exploration and exploitation in active learning for interactive video retrieval, IEEE Trans. Image Process. 22 (3) (2013) 955–968.
[118] L. Azzopardi, D. Kelly, K. Brennan, How query cost affects search behavior, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2013, pp. 23–32.
[119] L. Azzopardi, The economics in interactive information retrieval, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2011, pp. 15–24.
[120] M.M. Hall, E.G. Toms, Building a common framework for IIR evaluation, in: Proceedings of Information Access Evaluation. Multilinguality, Multimodality, and Visualization – 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23–26, 2013, pp. 17–28.
[121] D. Kelly, S. Dumais, J. Pedersen, Evaluation challenges and directions for information-seeking support systems, Computer 42 (3) (2009) 60–66.
[122] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, 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, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 63–74.
[123] I. Mironica, B. Ionescu, J. Uijlings, N. Sebe, Fisher kernel based relevance feedback for multimodal video retrieval, in: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, ACM, New York, NY, USA, 2013, pp. 65–72.
[124] X. Jin, J. French, 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, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 191–206.
[125] I. Ruthven, M. Lalmas, A survey on the use of relevance feedback for information access systems, Knowl. Eng. Rev. 18 (2) (2003) 95–145.
[126] D. Harman, Relevance Feedback and other query reformulation techniques, in: Information Retrieval: Data Structures & Algorithms, Prentice-Hall, Upper Saddle River, NJ, USA, 1992.
[127] 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, ACM, New York, NY, USA, 1978, pp. 151–170.
[128] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, in: Conference on Computer Vision and Pattern Recognition Workshop, 2004, CVPRW '04, 2004, pp. 178–178.
[129] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset, Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.
[130] M.J. Huiskes, B. Thomee, M.S. Lew, New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative, in: MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval, ACM, New York, NY, USA, 2010, pp. 527–536.
[131] M. Wang, L. Yang, X.-S. Hua, MSRA-MM: Bridging Research and Industrial Societies for Multimedia Information Retrieval, Technical Report MSR-TR-2009-30 (March 2009).
[132] S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, 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.
[133] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, I. Kompatsiaris, The 2012 social event detection dataset, in: Proceedings of the 4th ACM Multimedia Systems Conference, ACM, New York, NY, USA, 2013, pp. 102–107.
[134] G.A. Miller, Wordnet: a lexical database for English, ACM Commun. 38 (11) (1995) 39–41.
[135] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (4) (2002) 422–446.
[136] T. Qin, T.-Y. Liu, J. Xu, H. Li, LETOR: a benchmark collection for research on learning to rank for information retrieval, Inf. Retr. 13 (4) (2010) 346–374.
[137] T. Qin, T. Liu, Introducing LETOR 4.0 datasets, CoRR abs/1306.2597.
[138] 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, ACM, New York, NY, USA, 1993, pp. 36–47.
[139] D. Harman, Overview of the second text retrieval conference (TREC-2), Inf. Process. Manag. 31 (3) (1995) 271–289.
[140] D. Harman, Overview of the Third Text Retrieval Conference (TREC-3), in: Proceedings of The Third Text REtrieval Conference, (TREC), Gaithersburg, Maryland, USA, 1994, pp. 1–20.
[141] E.M. Voorhees, D. Harman, Overview of the sixth text retrieval conference (TREC-6), Inf. Process. Manag. 36 (1) (2000) 3–35.
[142] E.M. Voorhees, D. Harman, Overview of the seventh text retrieval conference TREC-7, in: Proceedings of the Seventh Text REtrieval Conference (TREC-7), 1998, pp. 1–24.
[143] E.M. Voorhees, D. Harman, Overview of the eighth text retrieval conference (TREC-8), in: Proceedings of the Eighth Text REtrieval Conference (TREC-8), 2000, pp. 1–24.
[144] E.M. Voorhees, D. Harman, Overview of the ninth text retrieval conference (TREC-9), in: Proceedings of the Ninth Text REtrieval Conference (TREC-9), 2000, pp. 1–14.
[145] E.M. Voorhees, D. Harman, Overview of TREC 2001, in: Proceedings of the Tenth Text REtrieval Conference, 2001.
[146] C. Buckley, The TREC-9 query track., in: TREC, 2000.
[147] S.E. Robertson, I. Soboroff, The TREC 2002 filtering track report., in: TREC, vol. 2002, 2002, p. 5.
[148] J. Allan, Hard Track Overview in TREC 2003 High Accuracy Retrieval From Documents, Technical Report, DTIC Document, 2005.
[149] P. Over, Trec-6 Interactive Track Report, NIST Special Publication SP, 1998, pp. 73–82.
[150] P. Over, Trec-7 Interactive Track Report.
[151] W. Hersh, P. Over, Trec-8 interactive track report, NIST Special Publication SP 246, 2000, pp. 57–64.
[152] I. Soboroff, I. Ounis, J. Lin, I. Soboroff, Overview of the TREC-2012 microblog track, in: Proceedings of TREC, vol. 2012, 2012.
[153] J. Lin, M. Efron, Overview of the TREC-2013 microblog track, in: Proceedings of TREC, vol. 2013, 2013.
[154] E.M. Voorhees, Overview of the TREC 2003 robust retrieval track, in: TREC, 2003, pp. 69–77.
[155] Y. Yang, S. Newsam, Geographic image retrieval using local invariant features, IEEE Trans. Geosci. Remote Sens. 51 (2) (2013) 818–832.
[156] P. Brodatz, Textures: A Photographic Album for Artists and Designers, Dover Publications, New York, 1966.
[157] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (COIL-100), Technical Report, Columbia University. ⟨http://www.cs.columbia.edu/CAVE/databases/papers/nene/nene-nayar-murase-coil-100.ps⟩, 1996 [cited September 24, 2015].
[158] H. Müller, S. Marchand-Maillet, T. Pun, The truth about corel—evaluation in image retrieval, in: M. Lew, N. Sebe, J. Eakins (Eds.), Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, Berlin, Heidelberg, 2002, pp. 38–49.
[159] M. Grubinger, P. Clough, A. Hanbury, H. Müller, Overview of the ImageCLEFphoto 2007 photographic retrieval task, in: Advances in Multilingual and Multimodal Information Retrieval, Springer, Berlin, Heidelberg, 2008, pp. 433–444.
[160] T. Arni, P. Clough, M. Sanderson, 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, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 500–511.
[161] M.J. Huiskes, M.S. Lew, The MIR Flickr retrieval evaluation, in: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval, ACM, New York, NY, USA, 2008.
[162] L.J. Latecki, R. Lakämper, Shape similarity measure based on correspondence of visual parts, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1185–1190.

[163] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, 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, ACM, New York, NY, USA, 2009, pp. 48:1–48:9.
[164] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Sixth Indian Conference on Computer Vision, Graphics Image Processing, 2008. ICVGIP '08, 2008, pp. 722–729.
[165] M. Everingham, A. Zisserman, C.K.I. Williams, L. Van Gool, The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results, ⟨http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf⟩.
[166] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, ⟨http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html⟩.
[167] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, ⟨http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html⟩.
[168] S. Schmiedeke, C. Kofler, I. Ferrané, Overview of MediaEval 2012 genre tagging task, in: MediaEval 2012 Workshop, Pisa, Italy, 2012.
[169] P. Over, T. Ianeva, W. Kraaij, A.F. Smeaton, TRECVID 2005—An Overview, 2006.
[170] P. Over, T. Ianeva, W. Kraaij, A.F. Smeaton, TRECVID 2006—An Overview, 2007.
[171] P. Over, G. Awad, W. Kraaij, A.F. Smeaton, TRECVID 2007—Overview, 2014.
[172] P. Over, G.M. Awad, T. Rose, J. Fiscus, W. Kraaij, A.F. Smeaton, TRECVID 2008—Goals, Tasks, Data, Evaluation Mechanisms and Metrics, 2009.
[173] P. Over, G.M. Awad, J. Fiscus, M. Michel, A.F. Smeaton, W. Kraaij, TRECVID 2009—Goals, Tasks, Data, Evaluation Mechanisms and Metrics, 2010.
[174] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval—The Concepts and Technology Behind Search, 2nd edition, Pearson Education Ltd., Harlow, England, 2011.
[175] K. Hofmann, S. Whiteson, M. de Rijke, A probabilistic method for inferring preferences from clicks, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2011, pp. 249–258.
[176] E. Kanoulas, M.M. Hall, P.D. Clough, B. Carterette, 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.
[177] P. Vakkari, Exploratory searching as conceptual exploration, in: Proceedings of the Fourth Workshop on Human–Computer Interaction and Information Retrieval, 2010, pp. 24–27.
[178] S. Dumais, Whole-session Evaluation of Interactive Information Retrieval Systems: Compilation of Homework (NII Shonan Workshop, October 2012). ⟨http://research.microsoft.com/en-us/um/people/sdumais/niishonanworkshop-web/NII-Shonan-CompiledHomework_Final.pdf⟩ [cited May 20].
[179] F. Baskaya, H. Keskustalo, 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, ACM, New York, NY, USA, 2012, pp. 105–114.
[180] A. Kustarev, Y. Ustinovskiy, A. Mazur, P. Serdyukov, Session-based query performance prediction, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2012, pp. 2563–2566.
[181] A. Kotov, P.N. Bennett, R.W. White, S.T. Dumais, 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, ACM, New York, NY, USA, 2011, pp. 5–14.
[182] M.D. Smucker, J. Allan, 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, ACM, New York, NY, USA, 2007, pp. 623–632.
[183] Y. Yan, H. Shen, G. Liu, Z. Ma, C. Gao, N. Sebe, {GLocal} tells you more: coupling {GLocal} structural for feature selection with sparsity for image and video classification, Comput. Vis. Image Underst. 124 (2014) 99–109 (Large Scale Multimedia Semantic Indexing).
[184] Y. Yan, E. Ricci, G. Liu, N. Sebe, Egocentric daily activity recognition via multitask clustering, IEEE Trans. Image Process. 24 (10) (2015) 2984–2995.
[185] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, N. Sebe, A multi-task learning framework for head pose estimation under target motion, IEEE Trans. Pattern Anal. Mach. Intell. 38 (6) (2016) 1070–1083.
[186] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, T.-S. Chua, Beyond text QA: multimedia answer generation by harvesting web information, IEEE Trans. Multimed. 15 (2) (2013) 426–441.
[187] L. Nie, S. Yan, M. Wang, R. Hong, T.-S. Chua, Harvesting visual concepts for image search with complex queries, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, New York, NY, USA, 2012, pp. 59–68.

Rodrigo Tripodi Calumby is an Assistant Professor at the University of Feira de Santana, Brazil. He has a B.Sc. in Computer Science from the University of Santa Cruz, Brazil (2007). He has a M.Sc. (2010) and Ph.D. (2015) in Computer Science from the University of Campinas, Brazil. His main research interests include content-based information retrieval and classification, interactive information retrieval, effectiveness evaluation, diversity promotion, multimodal fusion, and machine learning applications.

Marcos André Gonçalves is an Associate Professor of Computer Science at the Universidade Federal de Minas Gerais, Brazil. His research interests include information retrieval, digital libraries, text classification, and text mining. Gonçalves has a Ph.D. in Computer Science from Virginia Tech. He is an Affiliated Member of the Brazilian Academy of Sciences.



Ricardo da Silva Torres received a B.Sc. in Computer Engineering from the University of Campinas, Brazil, in 2000. He got his Doctorate in Computer Science at the same university in 2004. He is an Full Professor at the Institute of Computing, University of Campinas. His research interests include image analysis, content-based image retrieval, databases, digital libraries, and geographic information systems.