Int J Multimed Info Retr (2012) 1:139–152
DOI 10.1007/s13735-012-0019-z

TRENDS AND SURVEYS

# New grand challenge for multimedia information retrieval: bridging the utility gap

**Alan Hanjalic**

Received: 19 June 2012 / Accepted: 12 July 2012 / Published online: 7 September 2012


**Abstract** The needs and expectations regarding multimedia content access have grown rapidly with the fast development of multimedia technology and the explosion of multimedia content around us. This imposed high demands on the level of sophistication of multimedia information retrieval (MIR) solutions. Although the potential to develop the MIR technology that meets such high demands has also rapidly grown over the years, we are not there yet with adequate solutions. This paper states that a significant step forward could become possible if the MIR field moves towards a *utility-centered* research focus. There, the criteria related to utility should be deployed to help us bridge the critical remaining gap that is in front of us—the *utility gap*, the gap between the expected and de facto usefuleness of MIR systems. Utility criteria reach beyond the objective relevance of MIR results to also consider their informativeness and how helpful they are for user's further actions. Bridging the utility gap can therefore be seen as the next grand challenge in the MIR research field. To pursue this challenge, we propose a *utility-by-design* approach, by which utility is targeted explicitly and embedded deep in the foundations of MIR solutions. The paper will first motivate this new MIR grand challenge and position it with respect to the current efforts in the field. Then, some possibilities for realizing the utility-by-design approach will be highlighted and translated into a number of recommended research directions.

This publication was supported by the Dutch national research program COMMIT.

A. Hanjalic (✉)
Multimedia Information Retrieval Lab, Delft University of Technology, Delft, The Netherlands
e-mail: a.hanjalic@tudelft.nl

**Keywords** Multimedia information retrieval · Multimedia search · Multimedia indexing · Utility-by-design

## 1 Introduction

Multimedia that is lost in a huge collection or in a back alley of the Internet is essentially useless. Research in *multimedia information retrieval* (*MIR*) is directed at preventing this. It aims at matching multimedia content and user needs and so at bringing image, audio and video items, further in the paper referred to as *multimedia items*, together with users. This is pursued by developing theories and algorithms that automatically assign, process and verify the descriptors (*metadata*) pertinent to the content of images, videos and music and then deploy them to retrieve the multimedia items required by the user.

The importance of MIR has grown rapidly with the fast development of multimedia technology and the explosion of multimedia content around us. Growing needs and expectations of users regarding multimedia content access in terms of semantically rich, personalized and context-aware relevance criteria have imposed high demands on the level of sophistication of MIR solutions. The potential to develop MIR technology that meets such high demands has increased over the past 20 years by building on intensive international research efforts [23]. This development accelerated, however, with the increasing contextualization of images, video and music in online networked communities of users formed around social network or content sharing sites such as Facebook, MySpace, Google+, YouTube, Flickr or Twitter. Such sites, frequently referred to as *social media*, link together multimedia content, diverse metadata and users of various profiles and interests and enable the users to interact with the content and with other people via and about the content.

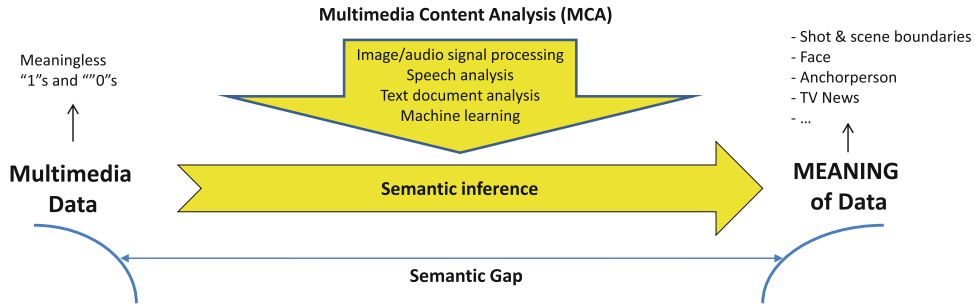


**Fig. 1** Illustration of the *semantic gap*, the gap between the representations of a multimedia content item at the data level and its interpretation at the semantic level [64]. Bridging this gap using the theory and algorithms of *multimedia content analysis* (MCA) enables automatic inference of semantics from data (*semantic inference*)

Richness of information that can be drawn from social media has brought vast new opportunities for improving the quality of MIR solutions [2]. Revisiting MIR in view of these opportunities, using the approaches that can jointly be referred to as *social media retrieval* [25], has helped the field to resolve some critical problems that impeded its development in the past. However, it could also help the field to address the new emerging demands.

The mission of this paper is to show how we can depart from the state of the art in social media retrieval to enable a significant step forward in the development of MIR theory and algorithms. We believe that this step could be realized through a focus shift in the MIR research, from a *technology-inspired* towards a *utility-centered* approach. Through this shift, the emphasis in the design and evaluation of MIR solutions should be put on optimizing the overal usefulness of MIR results. Here, the usefulness can be seen as a complex notion reaching beyond the objective relevance only and also encompassing the informativeness and helpfulness regarding user's further actions.

We start the technical part of the paper by reflecting in Sect. 2 upon current trends in the MIR field and discussing the rationale behind the theoretical and algorithmic concepts of social media retrieval. Then, in Sect. 3, we discuss the deficiencies of the existing social media retrieval solutions and justify the need for the above-mentioned focus shift in the MIR research approach. After proposing a strategy in Sect. 4 on how to change the MIR research focus, we revisit in Sect. 5 the foundations of social media retrieval and analyze how these foundations could be strengthened to help realize the proposed strategy. We conclude the paper with Sect. 6, in which we recommend a number of research directions that build on the new foundations and that should enable us to come a substantial step closer to truly useful MIR solutions.

## 2 Social media retrieval

### 2.1 Content access via semantic inference

For many years, the research in the MIR field has been dominated by the development of fully automated methods for *indexing* and *relevance ranking* of multimedia content. These methods are jointly referred to in this paper as *semantic inference*. Semantic inference has the objective to infer the 'meaning' (semantics) of the 'meaningless' 1s and 0s of a multimedia data stream and in this way bridge the *semantic gap* [64] between the representations of a multimedia content item at the data level and its interpretation at the semantic level.

As illustrated in Fig. 1, semantic inference is realized by means of specialized techniques for processing and analysis of data that are jointly referred to as *multimedia content analysis (MCA)*. The rationale behind the research on MCA was to provide a solid technological base for easily accessing multimedia content in a non-linear fashion and without the need for (extensive) manual annotation.

The fascination and challenge of the underlying scientific problem have mobilized a significant part of the research community to address MCA-related issues over the years, which has made MIR one of the most rapidly growing research fields within computer science. Impressive results achieved in the domain of automated video content analysis and representation [13,20], image and video indexing in terms of *semantic concepts* [27,65] and music content analysis and classification [5,16,46], as well as the immense popularity and growth of related international evaluation benchmarks, such as TRECVID [50] or Mirex [15], are the best indicators of the rise and achievements of what we refer to in this paper as the *content-analysis movement* in MIR.

Despite the growth and successes mentioned above, MCA has, however, not delivered practically deployable solutions, except for some narrow application domains. Due to the incapability to fully incorporate human (domain) knowledge into automatic procedures for interpreting multimedia data at the semantic level, bridging of the semantic gap is difficult in the general case [27].

### 2.2 Content access via social media

A popular activity of people involved in social media is to collectively tag, annotate, comment on and rate the multimedia content found there. By doing so, they provide objective cognitive information and subjective opinions and in

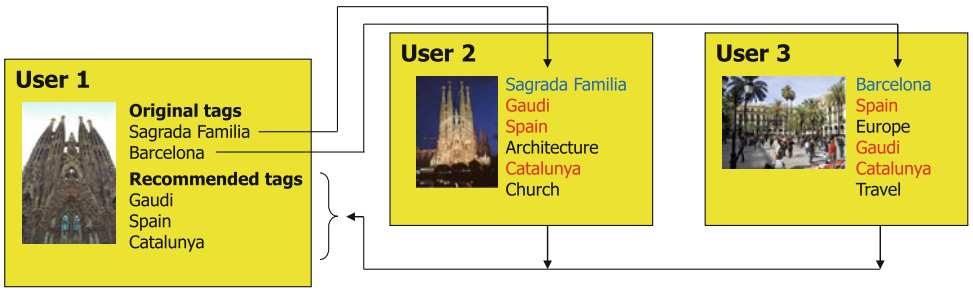


**Fig. 2** Illustration of a tag recommendation process following the method proposed in [57]. Two original tags assigned to the first image by User 1 are enriched by new tags implicitly recommended by Users 2 and 3 based on the analysis of the lists of candidate tags collected from the images that share the original tags from the first image

this way generate, explicitly or implicitly, metadata at the abstraction level that can typically not be handled by MCA procedures. Realizing this, Davis et al. [11] stated that the information derived from these interactions may in many cases be more powerful in facilitating multimedia content access in the social media context than the semantic inference techniques based on MCA. This initiated an alternative movement in MIR research that we refer to in this paper as the *social media movement*. MIR approaches belonging to this movement rely solely on the information drawn from the interactions of people with content and each other in social media and without doing any content analysis whatsoever. Such approaches can prove useful particularly in those indexing and relevance ranking scenarios that cannot be handled successfully by the semantic inference methods [62].

A typical example of the usability of social media for facilitating the multimedia content access is illustrated in Fig. 2. There, following the method proposed by Sigurbjörnsson and van Zwol [57], two original tags assigned by User 1 to the first image are enriched by adopting new tags from the tag sets generated by Users 2 and 3 for the images in their own collections. Based on the analysis of the lists of candidate tags collected from the images from other collections that share the original tags from the first image, new related tags can automatically propagate to the first image. An enrichment of the access to an image collection is here achieved using the information acquired simply by sharing images and tags among the users. High visual diversity of the images in Fig. 2 indicates, however, that achieving the same by means of semantic inference, e.g. by trying to link related images together based on their visual commonalities and then use these relations for tag propagation, would not be likely to succeed. A further illustrative example is the method for web image classification into adult versus non-adult images [47]. In view of the semantic complexity (high abstraction level) of the classification criterion, it was not surprising that exploitation of the social context, or more specifically the graph built out of hyperlinks among the web sites, showed a clear advantage in achieving the right classification compared to semantic inference.

In general, however, the MIR paradigms relying on social media are also insufficient for effective multimedia content access if considered in isolation. For instance, in order for the example method illustrated in Fig. 2 to succeed, large quantities of relevant images and tags are needed, preferably also accompanied by reliable links among the users indicating their similarity in interests, tastes and preferences. We have indeed witnessed a rapid rise of the image sharing, tagging and user-linking phenomena over the past years, which could be attributed to the enormous power of the social dimension of the modern Internet. However, large quantity of information is in this case hard to match with high quality; a vast variety of *social incentives* [1] moving the people to share content and interact may cause the tags to be noisy, vague and even misleading and therefore not necessarily helpful for MIR [40].

### 2.3 Convergence of content analysis and social media

While the potential of each of the two movements is too valuable to be ignored, neither of them alone is capable of solving the MIR problem. The discussion above clearly indicates that the future of MIR lies in a convergence of the two movements. It is intuitive that user-provided tags or ratings can complement MCA in expanding the content access possibilities, and also confirm or correct the low-confidence MCA results. Possibility for enhancement in the opposite direction is obvious as well. While tagging of small media volumes and simple multimedia items (e.g. through an online image annotation game [72]) might be doable by users, this becomes much more difficult for large volumes of audio-visual content, such as those found in typical video collections. Then, MCA can be used to improve the effectiveness and efficiency of the tagging process. For example, MCA could first identify a raw set of content items that may be of interest for the users and then this set could be refined and enriched through rating, recommending and tagging of the items by natural human actions on online social platforms [66].

Increasing convergence between the content and context movements within MIR have resulted in a new wave of scientific contributions. These contributions can be brought under the social media retrieval research direction introduced earlier in this paper that has marked the developments in the MIR

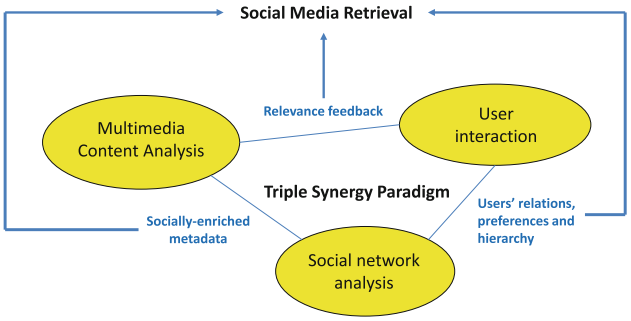


**Fig. 3** An illustration of the *social media retrieval* idea integrating the information resources available in a social network context and combining them with multimedia content analysis. Connections among the nodes in the graph indicate three main categories of benefits to enhancing the content access experience by means of social media retrieval: (a) relevance feedback regarding content and metadata from the side of the user, (b) social enrichment and verification of metadata through combining community-added and automatically extracted metadata and (c) inferring knowledge on users' preferences and relations in the social media context

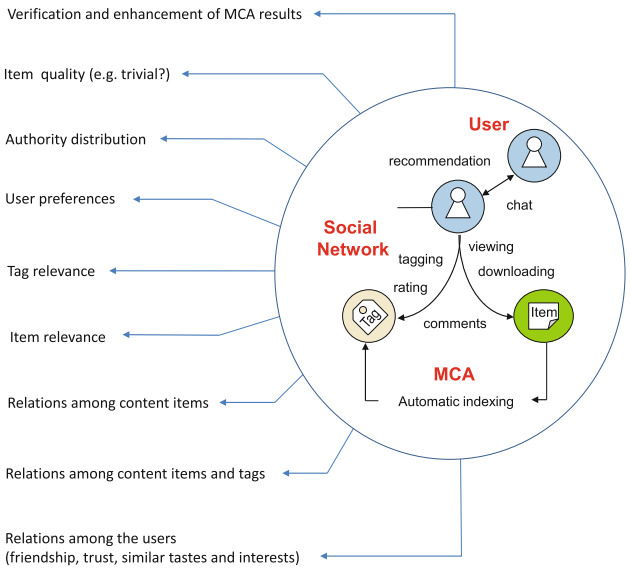


**Fig. 4** Illustration of the *triple synergy paradigm (TSP)* for integrating information resources in the social network context. The goal is to have different pieces of information about the multimedia content derived from MCA, user interaction and the analysis of processes in a social network complement each other to infer new information that can facilitate multimedia content access substantially better than any of the individual information sources taken in isolation. Examples of new inferred information are listed in the *left part* of the figure

field over the past several years. Essentially, social media retrieval combines the information resources available in the social media context with MCA to improve the efficiency and effectiveness of multimedia content access.

More specifically, the principles underlying the methods and algorithms of social media retrieval can be said to largely follow the *triple-synergy paradigm (TSP)* [22,53]. As illustrated in Fig. 3, this paradigm provides a conceptual framework for realizing MIR solutions by integrating three fundamental categories of information-generating processes:

- **Multimedia content analysis (MCA)**, which stands for automatically assigning metadata to multimedia data using signal processing, data mining, computer vision and pattern recognition,
- **Individual user interaction** with content and other users, such as adding tags or ratings, commenting uploads or ratings of other users and explicitly recommending content to them,
- **Analysis of networks of users, content and metadata**, which stands for analyzing information flows in such (typically very large) networks and deriving implicit characteristics of the users (e.g. trustability, authority), items and metadata (e.g. relevance), as well as the implicit relations among users, content and metadata.

Connections among the nodes in the TSP graph in Fig. 3 indicate three main categories of benefits to enhancing the content access experience by means of social media retrieval:

- Acquiring **relevance feedback** regarding content and metadata from the side of the user,
- **Social metadata enrichment and verification**, through combining community-added and automatically extracted metadata,
- Inferring **knowledge on users' preferences, relations and hierarchy** in the social media context.

The goal of social media retrieval is to let these benefits optimally complement each other in order to be able to infer new information that can facilitate multimedia content access substantially better than any of the individual information sources (i.e. the nodes in Fig. 3) taken in isolation. Examples of possible new inferred information about multimedia items, users and metadata are listed in the left part of Fig. 4. In the remainder of this section we illustrate the possibilities for a realization of the social media retrieval approach on a number of examples from recent literature.

The objective of the method proposed by Li et al. [39] is to estimate the relevance of the tags previously assigned to the target image. For this purpose, tag lists are analyzed accompanying the images available in the same context as the target image (e.g. on a social content sharing platform), but that are visually similar to the target image. The rationale behind this approach is that occurence of a tag at many images of the same or similar visual content would imply a high relevance of that tag with respect to the visual content (e.g. a scene, object, people) depicted in those images. Therefore, visual similarity criteria are deployed to identify the images being the visual neighbors of the target image and it is analyzed in how many neighboring images a given tag of the target tag

list occurs. The number of occurences of a tag in the visual neighborhood determines the relative relevance of the tag in the target list.

Having a similar objective, Liu et al. [41] propose a probabilistic tag relevance estimation mechanism that infers the relative tag relevance based on how often a tag is used to annotate the same visual content in the collection. This estimation is then refined by means of a random walk along the tag-similarity graph. The graph takes all tags as the nodes, while tag similarities deployed to weight the graph edges are obtained based on tag co-occurrence in the collection and the similarity of visual neighbors of the tags (i.e. the groups of visually similar images sharing each of the tags).

More complex systems, relying on more information resources and looking beyond metadata enrichment only can be found in the works of Zha et al. [78], Bu et al. [4] and Rudinac et al. [55]. Zha et al. [78] combined user interaction, content analysis and the information derived from a social content sharing platform to devise an iterative multimedia search mechanism. This mechanism first guides the user in query specification using multimodal cues and then uses the same cues to refine the final result by means of reranking the initial text-based search results. Bu et al. [4] developed a collaborative music recommendation system that infers the affinity between users and music items by means of a hypergraph. The hypergraph integrates music items, users and social metadata and deploys MCA to provide additional information about the similarity of music items. Rudinac et al. [55] proposed an algortihm for selecting optimal user-generated images summarizing a given location. Given a geo-location, the algorithm collects all images from a social content sharing platform taken in the vicinity of that location. It then integrates the images, visual features, textual metadata and user relations into a graph structure. Finally, a sophisticated image clustering is performed using explicit and implicit image relations derived from the graph, after which the obtained clusters are used to select images for the visual summary. This is done by trading off the representativeness of the images and the diversity of the generated summary.

## 3 Technology-inspired versus utility-centered research approach

The TSP-enabled integration of resources available in a social media context has already established itself as the necessary condition to boost the development of MIR solutions towards the desired level of sophistication. Since the related efforts draw from the existing MIR technology of semantic inference, interactive search and network analysis and aim at exploring new possibilities for solving MIR problems through different combination of these technology components, we refer to them as being *technology-inspired*. Vast number of ideas and methods proposed over the past years, some of which mentioned in the previous section, have substantially advanced the state of the art in MIR by exploring the TSP-enabled solution concepts that have not been possible before and by revealing first insights in how rich information resources and interdisciplinary expertise in the social media context can be deployed to make progress.

While the emphasis on the necessary condition has been rather strong in the past, not so much attention has been given to the sufficient condition for improving sophistication of MIR solutions. This condition states that MIR research must maintain a firm connection with the users and the use scenarios arising in their daily lives [17,43]. In other words, in order to achieve full practical impact, MIR research must orient itself towards achieving increased utility for a user.

The main consequence of neglecting the sufficient condition is that the existing MIR solutions are still insufficiently helpful to the users:

> *Insufficient utility can be seen as the main bottleneck preventing MIR to achieve as high a practical impact as the traditional text search.*

What we mean by insufficient utility can be illustrated again using the examples of the methods mentioned in the previous section. Tag relevance estimation and ranking for images is still too biased towards objects visually depicted in images. This bias prevents these methods to properly handle and evaluate (in terms of relevance) semantically rich tags, like those describing an image in terms of its general topic or the emotion it elicits in users. Furthermore, while novel socially enriched recommendation approaches have shown more potential than the traditional collaborative filtering approaches in improving the rate of retrieved relevant results, still the fundamental insights are missing on how to make the recommendation maximally useful (e.g. non-trivial, diversified).

It can be said that, in general, the technology-inspired MIR solutions have largely focused on maximizing the objective relevance of the top-N retrieved results, while neglecting the following important questions that explicitly address the usefulness of the results, namely

1. Which of the many relevant results best match the user's information need in a given search case, and
2. How the informativeness and helpfulness of the retrieval output can be maximized even if its relevance is suboptimal.

Insufficient awareness regarding utility is also observable from how the MIR solutions have been evaluated and how 'blind' deployment of common evaluation criteria has negatively biased the development of MIR solutions in terms

of utility. In this analysis of the precision-recall metric [63], Slaney revealed fundamental deficiencies of this metric as an evaluation criterion for MIR systems. He emphasizes the complexity of the information need and consequently urges for more sophistication in defining the evaluation criteria addressing this need in all its facets relevant to a given use case. We illustrate this on another typical example from the past research on MIR, namely the methods for visual search reranking. Various visual search reranking approaches [30] have shown potential to improve the text-based image and video search results using the visual modality. Reranking performance has been evaluated statistically, over a large multimedia collection and averaged over a vast number of queries, typically using the *mean average precision* (MAP) criterion. While the performance of reranking in terms of MAP has often shown significant improvement compared to the baseline text-based search, the results are still largely unpredictable per individual query—possibly the only query that is actually of interest for the user. This is not visible in MAP-based evaluation since MAP only reveals the overall statistical performance and not the performance for individual query. In this case it can be said that a too early focus on MAP optimization leads to reranking solutions that are statistically better, but of low utility for the individual user. We could almost paraphrase Slaney [63] and pose the following question here: 'If MAP is the metric, are you solving a relevant problem?'.

Although MAP remains a valid evaluation criterion, abandoning the MAP-bias during the method development could free the place for the *utility criteria* to take over the steering role in the design and realization of MIR solutions. In the case of visual search reranking, this would mean that, instead of statistical optimization, the design goal should be to devise a mechanism that first estimates automatically whether reranking is likely to improve the initial result and then, based on this, decides whether to rerank or not.

In the next section, we elaborate in more detail on the possibilities for rethinking the overall MIR research approach towards better utility. Due to the need to have the utility criteria play the central role in the development and evaluation of MIR solutions, we refer to the related research effort as being *utility-centered*.

## 4 The utility gap

The discussion in the previous section indicates that the research in the MIR field may be facing a new grand challenge in the form of the *utility gap*:

> *Utility gap is the gap between the expected and de facto usefuleness of MIR systems.*

Utility gap is a grand challenge as it requires a change in the mindset that underlies the design, development and evaluation of MIR solutions. In order to successfully pursue this challenge, the utility-centered criteria must become fundamental and embedded deep in the MIR system foundations. We take these considerations into account in the new *utility-by-design* paradigm that we introduce in this paper and that could serve as the basis for the new generation of MIR technology of improved utility.

While it is difficult to map the abstract notion of utility onto a limited set of criteria to steer the MIR research, we believe that a first substantial step towards improved utility could be taken by addressing the following, as yet barely explored four criteria that need to be imposed on the next generation of MIR systems:

- **Understanding the user's information need** in all its aspects,
- **Extracting semantically complex information from data** to respond adequately to semantically complex search requests,
- **User-centered search results ranking** that distinguishes truly relevant from objectively relevant search results, and
- **Facilitating further actions of the user**, even if the relevance of the search results is sub-optimal.

Since each criterion addresses a particular utility aspect, we will further refer to them also as different *facets of the utility-gap challenge*. Furthermore, as these criteria cover all major aspects of MIR, namely search request formulation and analysis, indexing, ranking and presentation of retrieval results, addressing these four facets leads to a radical shift in the MIR research focus.

In terms of scientific challenges, the first criterion listed above essentially requires insights and solutions that will enable a MIR system to expand the analysis of the information need of the user. Typically, this analysis has focused in the past on inferring 'what' the user is searching for. We believe, however, that automatically **inferring the search intent** of the user, namely the 'why' behind the user's search request, is at least as important for being able to adequately respond to this request as inferring the 'what'. The second criterion seeks beyond the current potential of MCA, but also the current generation of socially enriched solutions for tag relevance estimation or tag propagation, to perform **advanced semantic inference**. The goal is to push the limits of the current possibilities in bridging the semantic gap and automatically infer, assess and recommend tags or labels that also cover complex aspects of the multimedia content, like 'aboutness' (e.g. a general topic of a video), or those that are 'orthogonal to topic' and thus reflect issues like 'affectiveness' (emotion elicited in users through the video). The third criterion demands **increasing sophistication of the multimedia search results** towards the actual needs of the user.

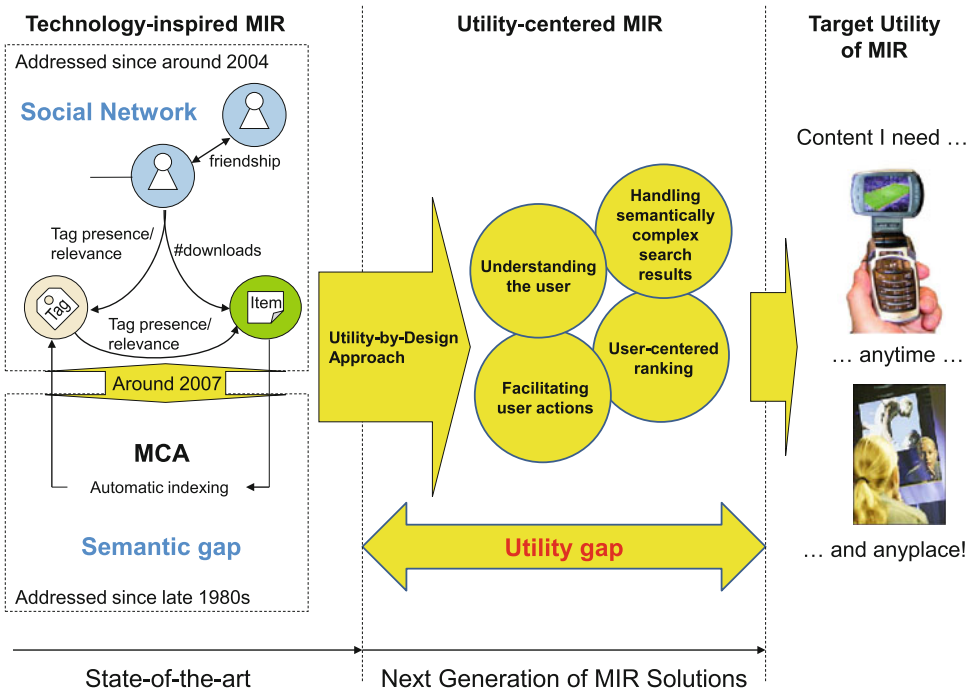


**Fig. 5** Positioning of the proposed utility-by-design approach with respect to the state of the art in the MIR field. The approach builds on the current TSP-enabled research momentum and aims at redirecting the MIR research to make it more utility-centered. The four facets of the utility-gap challenge indicated in the middle of the scheme can be seen as the first milestones in the realization of this approach

With the growing data collections, in which the number of relevant items grows rapidly, the retrieval needs to be reoriented towards truly useful items, where usefulness can be assessed both objectively (e.g. non-triviality, diversity) and subjectively (matching the search intent). Finally, the fourth criterion opens the question whether usefulness of the retrieval output can be kept high even if the user's search request is not optimally fulfilled. This requires from a MIR system to **explain the retrieval output** with respect to user's needs and indicate where the imperfections in the retrieval results stem from, i.e. either from the search system (e.g. imperfect retrieval algorithm), user input (e.g. suboptimal query formulation) or external factors (e.g. deficient or missing metadata or missing relevant, diverse or non-trivial multimedia content).

Figure 5 positions the proposed utility-by-design approach with respect to the state of the art in the MIR field. The approach builds on the current TSP-enabled research momentum and aims at redirecting the MIR research to make it more utility-centered. The four facets of the utility-gap challenge described above can be seen as the first milestones in the realization of this approach.

## 5 Strengthening the TSP foundations

The utility-by-design approach requires solid TSP foundations. While much has been achieved so far, more effort needs to be invested to make these foundations as strong as possible. In the following subsections, we identify the main challenges related to the three nodes in the TSP diagram in Fig. 3 and their relations.

### 5.1 Making the most of MCA

Semantic inference algorithms are well-known for their imperfections. Experience from the TRECVID evaluation benchmark [50] has shown that pushing the current performance limits of MCA solutions is not easy. However, alternative ideas have been proposed as well that do not focus on further improvement of the existing MCA concepts, but rather on combining of weak MCA components to produce new added value.

Preliminary experiments reported by Rudinac et al. [54] in the context of enabling topic-based video search have indicated that distributions of the responses of semantic concept detectors aggregated across different shots of a long video may be helpful in determining whether two videos cover the same general theme (e.g. topic, subject matter). In fact, these results point to the conclusion that the emphasis in the development of semantic concept detectors should not be on expanding the scope of entities to be detected, but rather on improving the detection of a limited number of most informative concepts. Furthermore, effective and efficient learning mechanisms would be required that investigate and automatically infer coocurrence and other dependency patterns among the visual concepts in various content domains. Having this information would make it possible to deploy the inherently weak individual visual concepts jointly in a much more productive manner.

In general, the fact that MCA results will always remain imperfect calls for a research approach that does not incorporate endless incremental improvement of MCA solutions towards the ideal (and unrealistic) performance level. Instead, the focus should be on determining the maximum possible added value (*best educated guess*) of a given MCA solution in a given use context. This added value would then serve as input into a more complex inference schemes, like the one described above and that involves aggregation of MCA results, or hybrid inference schemes, like the one in which imperfect MCA results will be corrected and enhanced through user interaction (e.g. [66]).

### 5.2 Making the most of the user

Users provide many sorts of explicit or implicit inputs regarding the content and other users they interact with in a social media context. Examples are tags, comments, ratings, but also the way video, music or image collections are browsed through. These inputs need to be collected and exploited to reveal the underlying preferences towards particular content items and the profile characteristics of the users in general. While the availability of such input is often generally assumed, more insight is required into the processes generating this input and this, in view of the strongly varying amount and quality of the input, preferably across different use cases and applications. Insufficient insight into these issues may lead to wrong assumptions regarding the usability of the user input, which may weaken the foundations for developing reliable and useful MIR solutions. On the other side, a sufficient insight into these issues may help create guidelines for optimizing the user input per use case and maximizing its value for a given MIR application.

Typical examples of the types of research the MIR field could substantially benefit from in the coming years therefore include research on tagging incentives (e.g. [1]), tagging games (and *games with the purpose* in general [72]), implicit (human-centric) tagging (e.g. [51]), feasibility of tagging and quality of tags in a MIR use case (e.g. [66,29]) and human-based computation, or *crowdsourcing* (e.g. [38,68,74]).

In particular, the role of crowdsourcing in a general MIR research context is becoming substantial. While the platforms like the Amazon's Mechanical Turk have served in the past to organize large-scale user tests evaluating the performance of MIR theory and algorithms, their potential in helping MIR research is much larger. They can be deployed to generate information, using which the focus, design rationale, operating points and the required performance quality of MCA algorithms could shift significantly compared to the past. Vliegendhart et al. [74] already showed how crowdsourcing can be deployed to discover new notions of user-perceived similarity between near-duplicate multimedia items, while Larson et al. [38] elaborated on the value of crowdsourcing as the means for multimedia benchmark dataset development.

Another relevant research line that needs deeper exploration encompasses modern means of interactions among the users that are formally out of the MIR context, but can still be deployed as a useful source of information to facilitate MIR, make it more reliable and improve its utility. For instance, pointers to potentially interesting events in video could be collected from spontaneous user reactions in the form of live chats on the web [48] or from Twitter [59].

### 5.3 Making the most of the network

Integrating together heterogeneous information collected from social media about the users, multimedia items and metadata has often been approached by means of a *social graph*, having as nodes the entities mentioned above and where the edges represent the explicit relations among the nodes. For instance, two images, music or video items can be compared directly, e.g. by means of audiovisual signal analysis, and the result of the comparison can then be mapped onto a value characterizing the edge between the two item nodes and representing the level of their relatedness.

The discussion in Sect. 2.3 indicated, however, that the main added value of social media retrieval lies in exploiting implicit information about the nodes, examples of which are given in Fig. 4. Therefore, after a social graph has been modeled using the available explicit information about the nodes, typically a graph analysis would follow to derive as much implicit information characterizing the nodes and their relations as possible. Coming back to the example of the item nodes, implicit relations between the multimedia items can be inferred through indirect links connecting two items in the graph, e.g. via shared metadata or users who uploaded, downloaded, commented, tagged or rated these items. The same holds for acquiring information about implicit relations among users that can be deployed to enrich the findings from the analysis of explicit user input as explained in the previous section. Aggregation of the explicit and implicit information encoded in the social graph can help acquire insights into the true relations among the nodes and the information flows through the network. This, again, enables validation, propagation and enrichment of annotations throughout the collection, bringing related users to each other and bringing the right content to the right users. Many recent works in MIR have adopted this approach for a wide variety of applications, as illustrated by the examples referred to earlier in this paper [4,41,55].

Substantial effort is, however, still required to make the social-graph and related models suitable for practical use cases and applications. One of the most obvious bottlenecks is the scalability of this modeling approach that may rapidly become non-tractable in view of the immense richness of

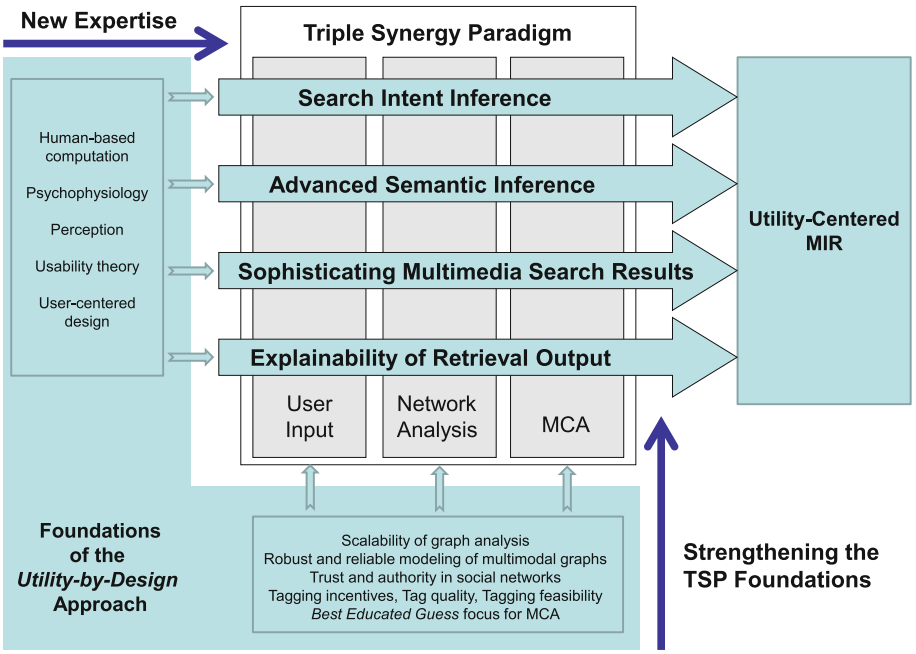


**Fig. 6** Illustration of the theoretical framework for realizing the proposed utility-by-design approach. The framework builds on the strengthened TSP foundations and a number of new enabling theories that should serve to bring the MIR research into the right utility-centered focus and steer it towards new generation of MIR solutions of improved utility

information in a social media context that is to be captured by nodes and edges of a graph. Another critical bottleneck results from still insufficient insights on how to handle multimodal information encoded in a typical social graph model. Addressing this bottleneck would typically involve research on cross-modal analysis, normalization and fusion of information [7,55]. Finally, MIR should rely more strongly on recent works in the domain of recommender systems, and, in particular, on the works on trust-aware recommendation (e.g. [44,45,60]). From there, insights can be drawn on how to most effectively model user relations in social graphs. An interesting research direction for MIR would then be to investigate to which extent these relation models could provide added value for typical MIR applications.

## 6 How to bridge the utility gap?

Figure 6 illustrates the theoretical framework that we envision for realizing the proposed MIR focus shift by means of the utility-by-design approach. The framework builds on the strengthened TSP foundations as described in the previous section and a number of new enabling theories. The theories mentioned in the figure, namely the human-based computation, psychophysiology, perception, usability theory and user-centered design, should serve to bring the MIR research into the right utility-centered focus and steer this research towards a new generation of MIR solutions of improved utility. We can say that these theories, together with the theories underlying the TSP, form the theoretical foundations of the utility-by-design approach.

In this section, we elaborate on a number of research directions for realizing this approach building on these foundations and following the four facets of the utility-gap challenge introduced in Sect. 4.

### 6.1 Advanced semantic inference: inferring the aboutness of the multimedia content

We believe that an important goal of the research on semantic inference should be to enable automatic annotation and search of videos in terms of general topics covered by their content. For instance, a query formulated to encode a topic, such as *court hearings*, *youth*, *politics* or *archaeology*, should lead to a list of videos that treat relevant subjects. The importance of solving this problem is large, as topic-based searches belong to the most natural ways of interacting with a multimedia collection.

The problem addressed here has multiple facets. First, the queries currently dominating the MIR research field are largely drawn from the ontologies typically focusing on the semantic concepts corresponding to, e.g. named persons, named objects, general objects and visual scenes (e.g. [27,49,65]). These concepts are strongly biased towards the visual channel of the video and therefore often referred to as the *visual concepts*. If taken individually, they are not representative of a general topic of a video. Second, the conventional supervised learning approach to building visual concept detectors is not effective here. While such detectors already suffer from a vast visual content variance indicating high semantic richness of many visual concepts, the richness and variation in the content corresponding to the same gen-

eral topic is much too high to be captured by a reasonably large set of visual examples and handled by a straightforward classification mechanism. Third, with the increasing abstraction level of the search request, information for the inference process needs to be extracted from a broad (longer) video. This is because a general topic is very difficult, if impossible, to extract from a single video frame, or even a shot, what has been the main practice in the past. Research approach here will therefore require an analysis of longer video clips, if not entire videos, for the purpose of indexing, which is a problem as yet hardly addressed in the MIR field. The main scientific challenge lies therefore in modeling high-level semantic similarity relations between long videos and assigning topic-level labels in a scalable automatic fashion.

One of the first attempts to address these challenges has been reported by Rudinac et al. [54]. There, topic-based video search was performed by combining the *query-performance prediction* (QPP) principle (e.g. [9,26,77]) with the results of many visual concept detectors aggregated across a video into a meta-level video representation. This representation is deployed by a QPP framework to evaluate the coherence (e.g. [28]) of the candidate video search list and to select the list that is most likely to respond to a given topical query. The potential power of this hybrid solution can be observed from the fact that the proposed approach is able to select the most suitable video search list for 30 % more queries than in the cases where only textual information is used to compare the videos. However, we are not there yet, which can be observed from the fact that the performance in the absolute sense (e.g. in terms of AP) is still too low.

While substantial improvement along the lines of the approach proposed in [54] could be achieved by refocusing the research on visual concept detection, as discussed in Sect. 5.1, alternative approaches are needed as well that may open new possibilities for addresing this problem in an innovative fashion to bring the performance to an acceptable level. A productive avenue of exploration here could be to guide the MCA processes by the input of users, e.g. collected through crowdsourcing. This input could help acquire insights not only about the different abstraction levels at which topics are typically specified in MIR scenarios, but also about various human-perceived topic-related video similarity and relevance criteria. Here, we can build on the initial works in this direction (e.g. [74]).

### 6.2 Advanced semantic inference: inferring the affectiveness of the multimedia content

In addition to the 'aboutness' as a retrieval criterion discussed in the previous section, another important aspect of the content of multimedia items that should be revealed by semantic inference is their *affectiveness*, i.e. the affective states (emotions, moods) they elicit at users. There have been several attempts to automatically infer affective labels (e.g. 'happy', 'sad', 'romantic', 'exciting') for a video that employed supervised classification methods (e.g. [32,35,67]). Successful realization of this approach for arbitrary audiovisual content remains, however, a difficult task. Finding a representative training data set is difficult due to the subjective nature and context dependence of affective responses to a given image, a piece of music or a video clip. Furthermore, the audiovisual signal variety of the 'happy', 'sad' or 'exciting' content is practically unlimited. Also, steering the classification process is difficult in a general case due to the abstract nature of affect categories, overlap and fuzzy boundaries between them. In addition, like in the problem of inferring aboutness, affective responses of the user are hardly elicited by a single frame or a shot. This results in the need to infer affect by analyzing longer video segments.

As a well-established alternative, a *dimensional approach* to affective video content analysis and modeling could be followed [24]. According to this approach, MCA models could be developed by characterizing the affective aspects of a video clip along two main affect dimensions, namely the *valence* (V) and *arousal* (A) [12], representing the type (positive, negative) and intensity (calm, excited) of an affective response, respectively. Although a third dimension (*control* or *dominance*) exists as well, it can be neglected here since it has been shown that A and V account for most of the independent affect variance [19]. The A and V models can then be combined into the model of a two-dimensional (2D) *affect curve* [24] as an elegant video representation that mimics the evolution of the affective state of the user while watching a video. It was shown in [21] how affect curve could be used to automatically index different video segments based on the expert-defined affective labels matching different regions of the 2D valence-arousal space and how compact representations of affect curves for different videos in the collection could be deployed to achieve affect-based video genre classification and recommendation.

Although the dimensional approach to affect described above has been widely adopted in recent literature (e.g. [33,70,75,76,79], the theory and algorithms for inferring affect-related information from multimedia items have remained largely undeveloped in terms of solid theoretical models and reliable, widely deployable algorithms. We believe, however, that a new momentum could emerge in this research direction by benefiting from some recent works (e.g. [37,58,71]), which provided new insights on affective audiovisual feature extraction, and by relying on social information resources discussed earlier in this paper. In particular, such resources could help improve the quality of affective video annotations and video accessibility at the affective level for individual users and user groups. As an example, a standard set of professional labels annotating different parts of the 2D valence-arousal space could first be mapped on a subset indicating

the labels being of interest for the users' community. This is also important as professionally added affective labels do not necessarily correspond to typical search terms inserted by the users. This label acquisition could be achieved either by letting an individual user interact with the system, possibly in a transparent fashion via a game, or from many users through crowdsourcing. Depending on the segments the affect curve passes through, a distribution of video segments across the labels could be obtained indicating the affective nature of video content. This distribution could then be refined by integrating the video with other videos and metadata in a social graph and evaluating the relevance of the labels per video using rich, graph-informed video similarity relations. Video comparison and recommendation based on affective criteria could then be performed either as a next step building on a socially enriched affective video representation, or in parallel with the enrichment in a social graph by building on related initial works in this direction (e.g. [4,55]).

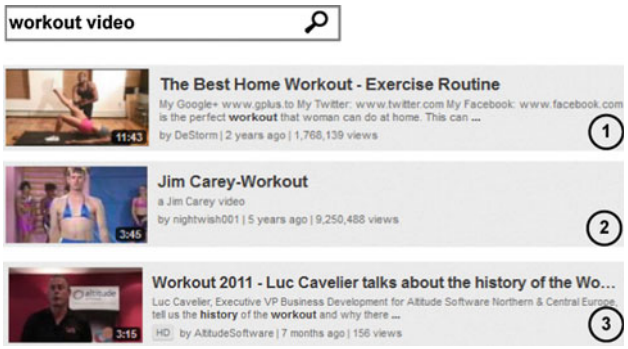


**Fig. 7** Illustration of the *search intent* concept. While all three videos returned for the query "workout video" are technically relevant, each of them may still satisfy different information need of the user, and as such not all of them would be relevant given a specific intent aspect of the information need. Video (*1*) shows how to work out, video (*2*) is related to work-out, but serves to entertain, and video (*3*) is on the history of the work-out and as such serves to inform

### 6.3 Search intent inference

If we want to move toward the ultimate goal of multimedia search engine development, namely to fulfill the information need of the users, we first need to stand still at the question what the notion of information need (or the *need behind the query* [3]) actually stands for. If one searches for 'airplanes', any image or a video related in some way to airplanes could be considered relevant to the search request. However, it is what the user had in mind before entering the search request, which makes a difference among the returned search results. We refer to this hidden dimension of the information need as *search intent* and emphasize its importance for evaluating the results list returned by a search engine.

We illustrate the notion of search intent on the example in Fig. 7. While all three videos returned for the query 'workout video' are relevant, each of them may still satisfy different information needs of the user, and as such not all of them would be relevant given a specific intent aspect of the information need. The first video shows how to work out, the second one is related to work-out, but serves to entertain, and the third one is a video on the history of the work-out and as such serves to inform.

If we refer to the objective relevance criteria as 'what' the user is searching for, the search intent can be linked to the reasons 'why' the user entered the search request in the first place. Previous MIR research analyzed earlier in this paper has largely focused on the 'what' aspect of the information need. This research branch still has some challenging goals to reach. However, in parallel, the research on the 'why' dimension should be intensified as well in order to allow the multimedia search engines to produce better articulated results and to link the results lists tighter to user's needs.

Search intent, just like the 'affectiveness' discussed in Sect. 6.2, can be said to be 'orthogonal' to the actual content of a multimedia item and as such also requires a dedicated research approach to be inferred with. Recent attempts in this direction [6,10,18,42] largely focused on expanding the general taxonomy-oriented approach from the 'textual' web (e.g. [3,14,34]). Taxonomy expansions have rightfully been considered necessary due to the fact that images, video and music items fulfill a different role on the web compared to general websites. Furthermore, multimedia items are much more information rich than text documents, which make them suitable to match a much wider (and partly also different) scope of intents than text. However, research on multimedia search intent has been rather small in scale and limited to specific types of multimedia items (largely images). It has also been insufficiently systematic to lead to well-justified dimensions of search intent that would result in productive discussions in the community and serve as the basis for setting up a productive intent-related research line.

To bring this research a significant step further, substantial effort in the community is required. We believe that this effort should explore the search intent space in its largest possible scope and exhaustively across different classes of multimedia content to obtain a well-founded set of intent categories. We also believe that this challenge requires an original research approach that breaks with the traditions of the past, strongly data-driven MIR research. Search intent is orthogonal to multimedia content and has a strong bias towards users. This implies that main insights underlying the search intent inference should ideally be acquired by relying on the input coming from the users, for instance generated by analyzing interactions in a social graph or through crowdsourcing.

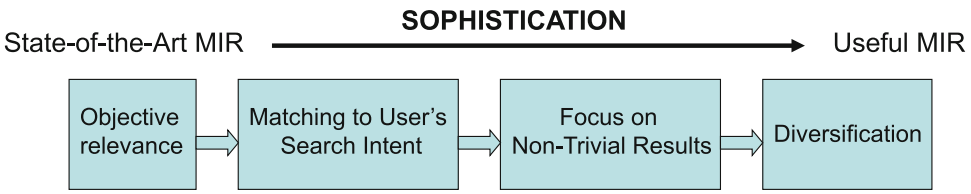


**Fig. 8** Illustration of the proposed *sophistication line*. Following this line, a typically large set of objectively relevant search results could be reduced to a smaller set of the results that are maximally informative and helpful to the user

### 6.4 Sophisticating the multimedia search results

Once the 'what' and 'why' aspects of the user's information need are properly addressed, the following intriguing facet of the utility gap challenge is to map a typically large set of technically relevant search results onto a more sophisticated set of the results that are maximally useful. Here, in addition to the need to match the multimedia content in the collection with the search intent of the user, other dimensions play an important role as well, such as whether they are diversified enough or trivial.

In order to help steer the process of developing useful multimedia search results, we propose the *sophistication line* illustrated in Fig. 8. The line departs from the general set of multimedia items being objectively (technically) relevant to the query (i.e. addressing the right 'what' aspect of the information need). A first step towards sophistication is then to reduce this set to a subset that fits the intent of the user (i.e. constraining the 'what' by the 'why'). Then, within this subset, items should be kept that are non-trivial in order to make the results list as informative to the user as possible. Finally, the remaining set of items should be diversified to present to the user all different aspects of the collections that match the information need.

Following the sophistication line is not trivial, as many obstacles are found underway for which no mature solutions have emerged from the past MIR research. The first problem that needs to be resolved is to automatically infer the search intent of the user. If this is successful, the user's query can be answered by the multimedia items that match the proper intent categories. Scarce works that go in this direction can mainly be found in the domain of interactive search [31]. However, no method has been proposed yet explicitly aiming at understanding the search intent. The second obstacle on the sophistication line is the detection of (non-)trivial multimedia items. Triviality of search results has not been addressed extensively, and then mainly in the domain of collaborative recommendation [8,61]. Finally, the problem of diversifying the multimedia search results remains largely unsolved as well. Although this problem has been addressed frequently in the past (e.g. [36,73,52,55,56,69]), the problem is not solved yet. The diversfication criteria need to move further from being too technical (e.g. diversity in the preselected audiovisual feature space) towards being more user-centered (e.g. diversity in semantics or elicited emotions). A critical ingredient that has been missing so far impeding the quality of diversification is the user feedback. Such feedback should be acquired not only to verify the diversification approaches proposed so far but also to inform further refinement of the existing approaches in terms of truly user-centered criteria. We believe that crowdsourcing could here also be a productive avenue to explore in the coming years.

### 6.5 Improving the explainability of the retrieval output

We conclude this paper by addressing the question whether the usefulness of the retrieval output could be kept high even if its quality in terms of the criteria defined in Sect. 6.4 is suboptimal. Could we still benefit from a search results list that contains trivial multimedia items, that is rather monotonous and does not directly reflect what we had in mind when approaching the search engine in the first place?

We believe that the ability of *any* search result list to guide the user in her further actions could be improved significantly *if* the information is provided to the user about the level at which the results can be relied on and about the probable causes of their imperfections. Is the imperfect result due to the imperfect retrieval algorithm, suboptimal query formulation or simply due to missing or inadequately indexed content? With such information, the user could either stop searching (no relevant content) or refine her search in an informed manner, through which the search session could be successfully completed in much fewer iterations. This information could, however, also be deployed internally by the search engine, as in the example given in Sect. 3 regarding the dilemma whether to rerank the initial results list or not before presenting it to the user. One of the important goals set to the MIR research community is therefore to develop theoretical and algorithmic concepts that evaluate the multimedia search results in view of the criteria discussed in Sect. 6.4, maximize the quality of the search result list based on the acquired insights and make the results explainable from the user's point of view.

The mean values currently available in the MIR field to address the above are rather scarce. Closest we can get are the methods mentioned earlier in the paper and based on

the QPP principle that can be deployed for evaluating top-N results lists given a query and the properties of the collection. The information acquired in this way can be used to make a prediction regarding the quality of the results list in terms of relevance. While this prediction can be deployed to directly select the best possible results list, the rationale underlying the prediction can be used to explain the 'absolute' quality of each considered results list to the user. In addition to the fact that it is already a challenge to extrapolate these QPP-based methods to work well on non-textual data [54], additional retrieval models and criteria are required in order to be able to expand the evaluation beyond the objective relevance only, namely to also evaluate the match of the MIR performance with the search intent, diversity and non-triviality requirements and reveal the probable causes of imperfections.

## References


1. Ames M, Naaman M (2007) Why we tag: motivations for annotation in mobile and online media. ACM CHI
2. Boll S (2007) MultiTube—where multimedia and Web 2.0 could meet. IEEE Multimedia 14(1):9–13
3. Broder A (2002) A taxonomy of web search. ACM SIGIR Forum 36(2):3
4. Bu J, Tan S, Chen C, Wang C, Wu H, Zhang L, He X (2010) Music recommendation by unified hypergraph: combining social media information and music content. In: ACM Multimedia, pp 391–400
5. Casey MA, Veltkamp R, Goto M, Leman M, Rhodes C, Slaney M (2008) Content-based music information retrieval: current directions and future challenges. In: Hanjalic et al (eds) Proceedings of the IEEE special issue on advances in multimedia, information retrieval, vol 96, no. 4
6. Choi Y, Rasmussen EM (2003) Searching for images: the analysis of users' queries for image retrieval in American history. J Am Soc Information Sci Technol 54(6):498–511
7. Clements M, de Vries AP, Reinders MJT (2010) The task-dependent effect of tags and ratings on social media access. ACM Trans Information Syst 28:21:1–21:42
8. Cremonesi P, Koren Y, Turrin R (2010) Performance of recommender algorithms on top-n recommendation tasks. In: ACM RecSys
9. Cronen-Townsend S, Zhou Y, Croft WB (2002) Predicting query performance. In: ACM SIGIR
10. Datta R, Joshi D, Li J, Wang JZ (April 2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60
11. Davis M, King S, Good N, Sarvas R (2004) From context to content: leveraging context to infer media metadata. In: ACM Multimedia
12. Dietz R, Lang A (1999) Aefective agents: effects of agent affect on arousal, attention, liking and learning. In: Cognitive technology conference
13. Divakaran A (ed) (2009) Multimedia content analysis, theory and applications. Springer, Berlin. ISBN: 978-0-387-76567-9
14. Donato D, Donmez P, Noronha S (2011) Toward a deeper understanding of user intent and query expressiveness. In: ACM SIGIR, query representation and understanding workshop
15. Downie JS (2008) The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. Acoust Sci Technol 29(4):247–255
16. Downie JS, Byrd D, Crawford T (2009) Ten years of ISMIR: reflections on challenges and opportunities. In: 10th International Society for music information retrieval conference (ISMIR)
17. Ebrahimi T (2009) Quality of multimedia experience: past, present and future. In: Keynote. ACM Multimedia
18. Fidel R (1997) The image retrieval task: implications for the design and evaluation of image databases. New Rev Hypermedia Multimedia 3
19. Greenwald MK, Cook EW, Lang PJ (1989) Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. J Psychophysiol 3:51–64
20. Hanjalic A (2004) Content-based analysis of digital video. Kluwer, Dordrecht. ISBN: 1-4020-8114-6
21. Hanjalic A (March 2006) Extracting moods from pictures and sounds: towards truly personalized TV. IEEE Signal Process Mag 23(2):90–100
22. Hanjalic A (2007) Content you like anytime and anyplace: multimedia research for new TV concepts. In: Keynote at the Pacific Rim conference on multi-media
23. Hanjalic A, Lienhart R, Ma W-Y, Smith JR (eds) (2008) Proceedings of the IEEE special issue on advances in multimedia information retrieval, vol 96, no.4
24. Hanjalic A, Xu L-Q (2005) Affective video content representation and modeling. In: IEEE transactions on multimedia, pp 143–154
25. Hanjalic A (2012) A new gap to bridge: where to go next in social media retrieval? In: Keynote international conference on multimedia modeling, Klagenfurt
26. Hauff C, Murdock V, Yates RB (2008) Improved query difficulty prediction for the web. In: ACM CIKM
27. Hauptmann AG, Christel MG, Yan R (2008) Video retrieval based on semantic concepts. In: Proc IEEE special issue advance multimedia information retrieval, vol 96, no 4, pp 602-622
28. He J, Weerkamp W, Larson M, de Rijke M (2009) An effective coherence measure to determine topical consistency in user-generated content. Int J Document Anal Recogn 12(3):185–203
29. Hildebrand M, van Ossenbruggen J (2012) Linking user generated video annotations to the web of data. In: MMM, pp 693–704
30. Hsu W, Kennedy LS, Chang S-F (2007) Reranking methods for visual search. IEEE Multimedia 14(3):14–22
31. Huang TS, Dagli CK, Rajaram S, Chang EY, Mandel MI, Poliner GE, Ellis DPW (2008) Active learning for interactive multimedia retrieval. In: Proceedings of the IEEE special issue on advances in multimedia information retrieval, vol 96, no 4
32. Irie G, Satou T, Kojima A, Yamasaki T, Aizawa K (October 2010) Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. IEEE Trans Multimedia 12(6):523–535
33. Jaimes A, Nagamine T, Liu J, Omura K, Sebe N (2005) Affective meeting video analysis. In: IEEE ICME
34. Jansen B, Booth D, Spink A (2008) Determining the informational, navigational and transactional intent of Web queries. Inf Proc Manag 44(3):1251–1266
35. Kang H-B (2003) Affective content detection using HMMs. ACM Multimedia
36. Kennedy LS, Naaman M (2008) Generating diverse and representative image search results for landmarks. In: Proceedings of WWW, pp 297–306
37. Lang P, Bradley M, Cuthbert B (2008) International affective picture system (IAPS): affective ratings of pictures and instruction manual. University of Florida, Gainesville, Tech. Rep. A-8
38. Larson M, Soleymani M, Eskevich M, Serdyukov P, Ordelman R, Jones G (2012) The community and the crowd: multimedia benchmark dataset development. IEEE Multimedia 19(3):2–10
39. Li X, Snoek CGM, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. In: ACM MIR

40. Li Q, Lu SC-Y (2008) Collaborative tagging applications and approaches. IEEE Multimedia Mag 14–21 (July–September)
41. Liu D, Hua X-S, Yang L, Wang M, Zhang H-J (2009) Tag ranking. In: World Wide Web conference, pp 351–360
42. Lux M, Kofler C, Marques O (2010) A classification scheme for user intentions in image search. In: International conference human factors in computing systems
43. Ma W-Y (2011) Intent, knowledge, and the cloud: towards a new search paradigm. In: Keynote. International conference on multimedia modeling
44. Ma H, Lyu MR, King I (2009) Learning to recommend with trust and distrust relationships. In: ACM RecSys, pp 189–196
45. Massa P, Avesani P (2007) Trust-aware recommender systems. In: ACM RecSys, pp 17–24
46. Müller M, Ellis DPW, Klapuri A, Richard G Signal processing for music analysis. IEEE J Selected Topics Signal Process 5(6):1088–1110
47. Mahajan DK, Slaney M (2010) Image classification using the web graph. In: ACM international conference on multimedia, pp 991–994
48. Miyamori H, Nakamura S, Tanaka K (2005) Generation of views of TV content using TV viewers' perspectives expressed in live chats on the web. ACM Press, ACM Multimedia, New York
49. Natsev A, Haubold A, Tesic J, Xie L, Yan R (2007) Semantic concept-based query expansion and re-ranking for multimedia retrieval: a comparative review and new approaches. In: ACM Multimedia
50. Over P, Awad G, Michel M, Fiscus J, Kraaij W, Smeaton AF (2012) TRECVID 2011—an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2011. NIST, March
51. Pantic M, Vinciarelli A (2009) Implicit human-centered tagging. IEEE Signal Process Mag 26
52. Paramita ML, Sanderson M, Clough P (2009) Diversity in photo retrieval: overview of the ImageCLEFPhoto task. In: CLEF
53. Ramzan N, Larson M, Dufaux F, Cluver K (2010) The participation payoff: challenges and opportunities for multimedia access in networked communities. In: ACM MIR
54. Rudinac S, Larson M, Hanjalic A (2010) Exploiting noisy visual concept detection to improve spoken content based video retrieval. In: ACM Multimedia
55. Rudinac S, Larson M, Hanjalic A (2011) Finding representative and diverse community contributed images to create visual summaries of geographic areas. In: ACM Multimedia
56. Sanderson M, Tang J, Arni T, Clough P (2009) What else is there? Search diversity examined. In: ECIR
57. Sigurbjörnsson B, van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: WWW, pp 327–336
58. Shan MK, Kuo FF, Chiang MF, Lee SY (September 2009) Emotion-based music recommendation by affinity discovery from film music. Expert Syst Appl 36(4):7666–7674
59. Shamma D, Kennedy L, Churchill EF (2009) Tweet the debates. In: ACM Multimedia workshop on social, media, pp 3–10
60. Shi Y, Larson M, Hanjalic A (2010) Towards understanding the challenges facing effective trust-aware recommendation. In: ACM RecSys '10 workshop on recommender systems and the social web
61. Shi Y, Serdyukov P, Hanjalic A, Larson MA (2012) Personalized landmark recommendation based on geotags from photo sharing sites: towards alleviating data sparseness and making non-trivial recommendations. ACM Trans Intell Syst Technol (accepted)
62. Slaney M (2011) Web-scale multimedia analysis: does content matter? IEEE Multimedia 18(2):12–15
63. Slaney M (2011) Precision-recall is wrong for multimedia. IEEE Multimedia 18(3):4–7
64. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans PAMI 22:1349–1380
65. Snoek CGM, Worring M (2008) Concept-based video retrieval. Foundations Trends Inf Retrieval 2(4):215–322
66. Smits EAP, Hanjalic A (2010) A system concept for socially enriched access to soccer video collections. IEEE Multimedia 17(4):26–34
67. Soleymani M, Kierkels JJM, Chanel G, Pun T (2009) A Bayesian framework for video affective representation. In: International conference on affective computing and intelligent interaction (ACII 2009) September
68. Soleymani M, Larson M (2010) Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus. In: ACM SIGIR, pp 4–8
69. Song K, Tian Y, Gao W, Huang T (2006) Diversifying the image retrieval results. In: ACM Multimedia
70. Sun K, Yu J, Huang Y, Hu X (2009) An improved valence-arousal emotion space for video affective content representation and recognition. In: IEEE ICME
71. Tkalcic M, Burnik U, Kosir A (2010) Using affective parameters in a content-based recommender system for images. User modeling and user-adapted interaction. J Pers Res 20(4):1–33
72. von Ahn L (2006) Games with a purpose. IEEE Comput 39(6):96–98
73. van Leuken RH, Garcia L, Olivares X, van Zwol R (2009) Visual diversification of image search results. In: Proceedings of WWW, pp 341–350
74. Vliegendhart R, Larson M, Pouwelse J (2012) Discovering user perceptions of semantic similarity in near-duplicate multimedia files, CrowdSearch 2012 workshop. In: WWW, pp 259–262
75. Xu M, Jin JS, Luo S, Duan L (2008) Hierarchical movie affective content analysis based on arousal and valence features. ACM Multimedia, ACM Press, New York
76. Yazdani A, Kappeler K, Ebrahimi T (2011) Affective content analysis of music video clips. ACM Multimedia, MIRUM workshop. ACM Press, New York, pp 7–12
77. Yom-Tov E, Fine S, Carmel D, Darlow A (2005) Learning to estimate query difficulty. In: ACM SIGIR
78. Zha Z-J, Yang L, Mei T, Wang M, Wang Z (2009) Visual query suggestion. In: ACM Multimedia
79. Zhang S, Tian Q, Jiang S, Huang Q, Gao W (2008) Affective MTV analysis based on arousal and valence features. In: IEEE ICME