User Intent in Multimedia Search: A Survey of the State of the Art and Future Challenges

CHRISTOPH KOFLER, Bloomberg L.P.
MARTHA LARSON and ALAN HANJALIC, Delft University of Technology

Today's multimedia search engines are expected to respond to queries reflecting a wide variety of information needs from users with different goals. The topical dimension ("what" the user is searching for) of these information needs is well studied; however, the *intent* dimension ("why" the user is searching) has received relatively less attention. Specifically, intent is the *"immediate reason, purpose, or goal"* that motivates a user to query a search engine. We present a thorough survey of multimedia information retrieval research directed at the problem of enabling search engines to respond to user intent. The survey begins by defining intent, including a differentiation from related, often-confused concepts. It then presents the key conceptual models of search intent. The core is an overview of intent-aware approaches that operate at each stage of the multimedia search engine pipeline (i.e., indexing, query processing, ranking). We discuss intent in conventional text-based search wherever it provides insight into multimedia search intent or intent-aware approaches. Finally, we identify and discuss the most important future challenges for intent-aware multimedia search engines. Facing these challenges will allow multimedia information retrieval to recognize and respond to user intent and, as a result, fully satisfy the information needs of users.

Categories and Subject Descriptors: H3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Algorithms, Design, Human Factors, Theory

Additional Key Words and Phrases: User intent, multimedia information retrieval, multimedia search, multimedia indexing, retrieval algorithms

ACM Reference Format:
Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User intent in multimedia search: A survey of the state of the art and future challenges. ACM Comput. Surv. 49, 2, Article 36 (August 2016), 37 pages.
DOI: http://dx.doi.org/10.1145/2954930

1. INTRODUCTION AND DEFINITION OF USER INTENT

The original challenge of multimedia was the development of the basic technologies needed for capturing, uploading, storing, and serving images, audio, and video. For example, Burgess and Green [2009] cite the main innovation of *YouTube*, at its moment of founding in 2005, to be technological in nature: removing the barriers to widespread

C. Kofler is a recipient of the Google Europe Doctoral Fellowship in Video Search. This research is supported in part by the Google Fellowship and also by the Dutch national program *COMMIT*/.
This work was carried out while C. Kofler was with Delft University of Technology. Currently, M. Larson is also at Radboud University Nijmegen.
Authors' addresses: C. Kofler, Search & Discoverability, Bloomberg L.P., 731 Lexington Ave, New York, NY 10022; email: ckofler4@bloomberg.net; M. Larson and A. Hanjalic: Multimedia Computing Group, Department of Intelligent Systems, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands; emails: {m.a.larson, a.hanjalic}@tudelft.nl.

DOI: http://dx.doi.org/10.1145/2954930

"I need to find a good video that explains ***how to do a cross knot bracelet****."*
"Where can I find a video of ***how to make an almond cake****? I need the recipe and like a video* ***to literally teach me****."*
"Where can I find a video of Flavor Flav laughing? My friend said ***I laugh like him and I just wanted to hear it****."*
"Do you know how I can find a video review of BMW X5? ***I am changing my car into a SUV and I want to buy a BMW X5*** *and I need a decent video review about this car."*

Fig. 1. Examples of real-world user information needs expressed in the domain of video search [Hanjalic et al. 2012].

video sharing. The picture today is radically different. The technical challenges have been addressed, if not completely solved: Multimedia recording equipment is widely available to users, storage capacity is plentiful, and technology has made content upload and online playback almost effortless. With these advances come the expectation that multimedia search engines are able to satisfy a wide spectrum of different user information needs. Taking YouTube again as an example: Users who were originally satisfied if they were only entertained now expect online video to support goals of a different nature, including communication, education, and problem-solving.

The concept of *information need* was developed in the field of Information Retrieval (IR) and is defined as the lack that a user is attempting to overcome by engaging in information-seeking behavior [Manning et al. 2008; Buettcher et al. 2010]. It is the abstract, often difficult-to-describe gap that a user is attempting to fill using a search engine. In order to use a search engine, users formulate queries, which reflect the underlying information needs only incompletely or indirectly. The field of Multimedia Information Retrieval (MIR) differs from conventional IR in that it goes beyond text to tackle the challenge of searching collections of multimedia content (e.g., images, spoken content, music, video). The new expectations that users bring to multimedia search engines reflect growing diversity of user information needs.

The breadth of the variety of these needs is illustrated by the list of user statements shown in Figure 1, which are questions involving video information needs that were posted to *Yahoo! Answers* [Hanjalic et al. 2012]. These examples serve to illustrate the difference between MIR and conventional text IR. Examining the individual statements, we see that text documents would not have been able to fully satisfy the information needs of these users. In some cases (e.g., the top example involving a bracelet), a video or image simply makes the process easier to understand. However, there are a large number of cases where textual information could not answer the query (i.e., the third example involving comparison of the sound of laughter). These observations are the point of departure for this survey. The MIR community clearly needs to look beyond conventional IR approaches in order to develop search engines that deliver what users are looking for.

A user information need is composed of two dimensions. The first dimension is the "what" dimension, which reflects the *topic* of the search. The second dimension is the "why" dimension and ties the need of the user directly to the task that the user is attempting to carry out. We define the "why" dimension as the *user intent*, which corresponds to the *"immediate reason, purpose, or goal behind a user's information need"* (see Hanjalic et al. [2012]). It is important to note that this definition follows the dictionary definition of intent, that is, *"the thing that [the user] plans to do or achieve; an aim or purpose."*[1] Since the "aim or purpose" here is linked to the specific search

[1]http://www.merriam-webster.com/dictionary/intent.

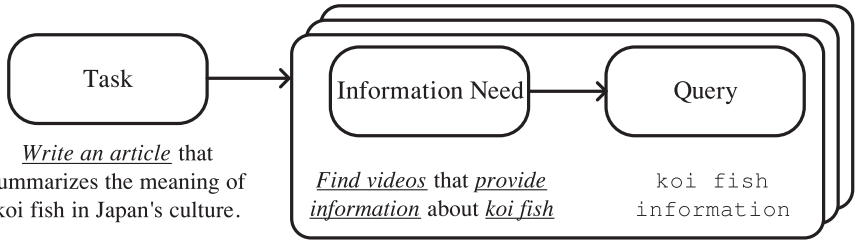

Fig. 2. The user search process: A real-world *task* is the trigger for users to consult a multimedia search engine with a set of *information needs*, which are verbalized in the form of (textual) *queries* and submitted to the multimedia search engine.

activity of the user, we will generally refer to user intent in this case as the *user search intent* (or *search intent*). We do this particularly to be able to distinguish it from other types of intent that may also play a role in optimizing multimedia search engines and that will be discussed in this article as well.

The importance of the intent dimension of the user information need is motivated by the statements in Figure 1. For example, in the top statement, the user wants to actually *produce* a cross-knot bracelet. Another user might want to learn to *identify* a cross-knot bracelet or to *examine a range* of possible cross-knot bracelet colors and styles. The videos that would best satisfy these users are different in terms of their content and their detail, although they all must generally address the topic of cross-knot bracelets. These observations provide the main motivation for this survey. Satisfying users' information needs involves providing results that are relevant to both the "what" and the "why" dimensions. MIR research must address user search intent if multimedia search engines are to fully satisfy user information needs.

In order to understand the role of intent in multimedia search in greater detail, we turn to consideration of its place in the user's larger search process, illustrated in Figure 2. Information needs arise in the context of a real-world *task* that users desire to complete. This task involves user activity that is usually completely independent of multimedia search. To understand the real-world nature of the user task, notice that the user could potentially also carry out the task without multimedia search. In the example in Figure 2, the user's task is to write an article on the meaning of koi fish in Japan, and this could also be accomplished by reading books, talking to experts, or traveling to Japan to observe.

Once users decide to consult a multimedia search engine given their initial real-world task, they approach the search engine with one or more concrete information needs. The example in Figure 2 demonstrates that the user's needs involve a desire for information *about a topic* (e.g., *koi fish*) and that also *fits the reason* why the user performs the search (e.g., *to obtain information*). Given the standard query-based interface of today's search engines, the needs are then *verbalized* in the form of *queries* (e.g., `koi fish information`). The multimedia search engine responds to the queries with a list of results. The results are considered to be relevant if the user's information need is fulfilled, and the user can move forward toward successful completion of the task.

It is important to notice that user search intent is, to a great extent, independent of topic. For example, in the first two information needs in Figure 1, both users have the goal to learn how to do something, but aim their attention at different topics ("cross knot bracelet" vs. "almond cake") (see Hanjalic et al. [2012]). We do not claim that there is complete orthogonality between the "what" and the "why" dimensions of multimedia information needs. However, treating the two dimensions as independent allows a

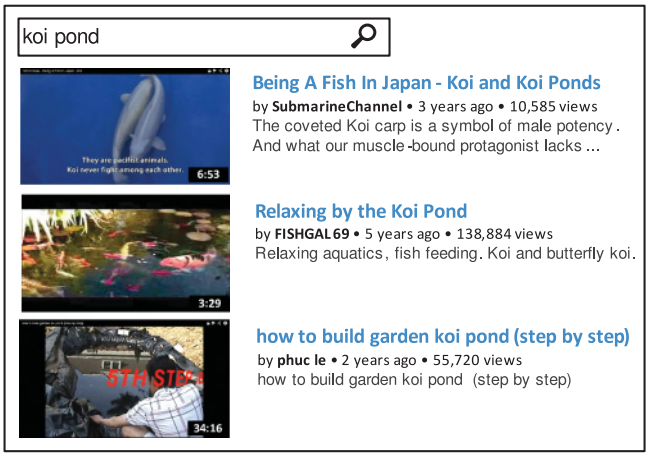

Fig. 3. Example search query with an excerpt of the produced results list illustrating the difference between the topical- and the intent-component of a user's information need: The query `koi pond` returns search results that all match the topical component of the user's information need; however, each satisfies different user intents. (This figure is adapted from a figure presented in Kofler et al. [2014]).

search engine more latitude to mix and match topics with intent types. The advantage is a potentially large improvement of the ability of the search engine to cover the full spectrum of user information needs.

The independence of "what" and "why" is further illustrated by Figure 3, which shows three videos returned in response to a user query, `koi pond`. The videos are all topically relevant to the query, but the results list has been diversified so that it covers three possible user search intents. The first result is an informational video that provides the user the opportunity to obtain knowledge, the second is a relaxing video that users might watch to change their mood, and the third result is a tutorial video that shows how to build a koi pond. These examples support the importance of considering topic and intent to make two distinct contributions to the relevance of search results to the user's information need.

To address user intent to be successful in multimedia search, it is important to understand its roots in conventional text IR. User search intent was first introduced by Broder [2002] in the field of text-based IR. Broder's work demonstrated that users' information needs are not all related to acquiring information, but rather span a much wider spectrum of intent types. Rose and Levinson [2004], Jones and Klinkner [2008], and Strohmaier and Kröll [2012] make a similar observation and characterize intent as the *underlying goal* behind a Web search. Baeza-Yates et al. [2006] examine the *"intention"* behind queries divided into topic categories (representing "what" users are searching for) and user goals (representing "why" they search). Multimedia researchers have built on the initial IR work to carry out studies of the nature of user intent in image search [Lux et al. 2010a] and video search [Hanjalic et al. 2012; Lagger et al. 2012].

At this juncture, we note that understanding of the treatment of intent in the literature requires close attention to terminology. First, although "intent" and "intention" are understood to mean "aim or purpose," there is a subtle but important difference of meaning between the two.[2] An "intention" is something that one sets about to do, but perhaps does not completely succeed in. For this reason, it is the preferred word to describe the information need as a whole. The user has the *intention* to express the

[2]http://www.learnersdictionary.com/qa/intent-and-intention.

information need as well as possible in the query, but perhaps is not completely successful. "Intent" reflects an underlying mindset and suggests great deliberation. For this reason, it is the preferred word to describe the underlying goal of the user (i.e., specifically the "why" dimension of the user need). Second, there has been little effort made to standardize the use of the term "intent" in the MIR literature. In this article, we adhere closely to the definition of intent as the user goal. In particular, "user intent" matches the "user goal" of the definition of "user intention" from Baeza-Yates et al. [2006]. However, to make the relationship between information need and intent clear, in this survey, we also cover literature in which the word "intent" has been used in the sense of "intention."

Thus far, the majority of research effort in the field of MIR has been devoted to creating algorithms and systems that analyze multimedia documents with regard to their topic. Topic is closely associated with meaning, and research on topic is regarded as addressing the semantic gap. Multimedia search engines are relatively successful at returning search results that users find to be on topic (cf. Snoek and Worring [2009] and Mei et al. [2014]). These results do not, however, completely satisfy the user's information need unless they also satisfy the user's intent. Fulfilling the intent is a challenging endeavor, particularly because user intent is often not explicitly reflected in the query. Crossing the semantic gap requires answering the question "What does this multimedia content mean?" Intent poses the problem of another type of gap, which requires answering the question "What does this multimedia content do?" Just as we cannot expect intent to be explicitly reflected in the query, we also cannot count on intent to be explicitly reflected in the multimedia content itself.

The purpose of this article is—by reviewing existing work and examining key future challenges—to discuss the enormous potential of user intent to satisfy users' information needs to their full extent. This article makes the following contributions:

—**Definitive characterization of *user intent* for multimedia search.** We survey the literature that has used the term "intent" with alternate interpretations. We show how this work relates to our definition of intent and the advantages of adopting our definition as definitive.
—**Inventory of user intent categories in multimedia search.** Many different conceptual models and typologies have been presented in the literature that enumerate and explain the reasons why users search in multimedia collections. We survey this literature and provide an analysis of the limitations and potential offered for intent-aware multimedia search.
—**Classification and analysis of intent-aware multimedia search approaches.** We provide an overview and analysis of existing approaches to multimedia search that exploit intent in addition to topic. The approaches address intent in different steps throughout the multimedia search pipeline and, overall, reveal the benefits of intent-aware multimedia search.
—**Future research challenges on intent-aware multimedia search.** Based on our analyses, we discuss key challenges that are valuable for future research and that need to be addressed by the MIR research community on the path toward truly intent-aware multimedia search engines.

The remainder of this survey is structured as follows. In Section 2, we first give an overview of the general architecture of a multimedia search engine that has been followed in the MIR community and that provides a framework for our discussion on where intent-aware approaches can be useful. In Section 3, we discuss types of approaches that aim to provide better information need satisfaction for users based on "intent" but that use an unexpected or alternate definition. In Section 4, we provide an overview of user intent models and typologies, both in text-based Web search and in

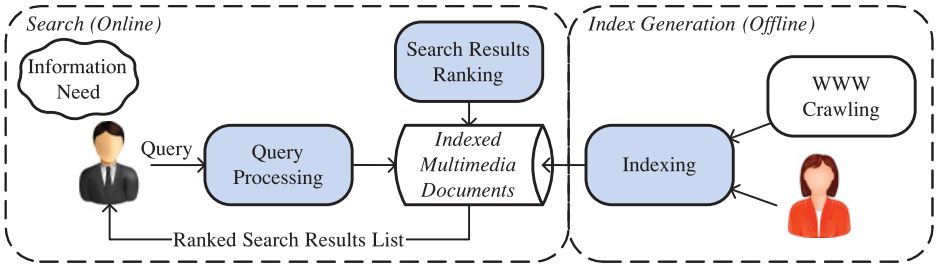

Fig. 4. High-level overview of a multimedia search engine: In the offline *index generation* step, multimedia documents that are either crawled from the Internet or drawn from a user's document collection are processed and described for scalable and efficient access. In the online *search* step, query processing prepares the user-submitted query in a way that is compatible with the data that have been indexed in the offline processing step. The query serves as input to the search results ranking component, which compares the query to the multimedia documents in the index and produces a ranked list of search results that is returned to the user.

multimedia search, and explain how these typologies have been derived. In Section 5, we analyze algorithms and approaches that build on these typologies. In doing so, we consider the entire multimedia search engine pipeline, from intent-aware indexing over intent-aware query processing to intent-aware search results list ranking. On the foundation of this analysis, we provide and elaborate on salient challenges in Section 6 that should be addressed in future work and conclude the article in Section 7.

2. MULTIMEDIA SEARCH ENGINES—A GENERAL ARCHITECTURAL OVERVIEW

A multimedia search engine is in general quite similar to a conventional text search engine, as described, for example, in Manning et al. [2008] and Buettcher et al. [2010]. It is made up of several important components that work together to provide users effective and efficient access to multimedia documents available on the Internet or in a restricted document collection. We cover the components explicitly here in order to highlight the differences with a multimedia search engine, as well as to provide a framework for the discussion that follows.

As depicted in Figure 4, a multimedia search engine consists—on a high level—of two main components: the offline *index generation* step and the online *search* step. Each of these components consists of subcomponents, which we briefly explain in turn.

Indexing. The task of the indexing step is to represent multimedia documents so that they are both effectively described and organized for efficient access. Multimedia documents have either been crawled from the Internet (e.g., discovered automatically by the search engine) or have been manually added to the system by users (e.g., uploaded images on *Flickr* or uploaded videos on YouTube). These documents typically consist of multimedia files (e.g., images, videos, audio files) paired with their user-generated textual metadata (e.g., titles, tags, description). Next to the textual metadata that is typically described using *tf-idf* or comparable representations, the actual content of multimedia files is indexed using a variety of local- (e.g., SIFT [Lowe 2004] combined with bag-of-visual-words), global- (e.g., color histograms, edge histograms etc.), and semantic- (e.g., Snoek and Worring [2009]) representations.

Query Processing. Once users have formulated a query given their information need, the purpose of this component is to prepare this query so that it is compatible with the data that have been indexed in the offline processing step (e.g., by applying compatible term weights). Additionally, this component is responsible for optimizing the query by (i) suggesting similar queries to the user (e.g., Zha et al. [2010]); (ii) classifying the query into predefined categories for which different ranking strategies are optimized

(e.g., Kennedy et al. [2005]); and (iii) expanding the initial query such that the ranking component is able to return documents that are maximally relevant to the user (e.g., Feng et al. [2010]).

Search Results Ranking. Once the query is prepared, it serves as input to the search results ranking component, whose purpose is to compare and match the query representation with the representation of the multimedia documents in the index. Output of the matching process is a relevance score for a document, used to rank documents. This component is also responsible for results list optimization (i.e., refining a results list that represents an initial or an intermediate product). Optimization is typically carried out by a reranking process that, for example, increases the homogeneity of the top ranks of results lists in terms of a particular modality as, for example, proposed in Yang and Hanjalic [2010] and Yang and Hsu [2008]. Another important optimization is diversification, which increases the heterogeneity of the top ranks of initial results lists in terms of a particular modality as, for example, proposed in Hoque et al. [2013] and van Leuken et al. [2009].

In the sections that follow, we shall see that intent-aware approaches to MIR can target many different places along the MIR pipeline to yield an overall improvement of MIR to satisfy user information needs.

3. DELIMITING THE DEFINITION OF "USER INTENT"

In recent years, a significant number of approaches to improving multimedia search have been presented in the literature. Many of these approaches cite the "intent" or the "intention" behind the user query as the target of improvement, but they do not deal with "the immediate reason, purpose, or goal behind a user's information need" (i.e., the definition of user intent introduced earlier). Our reasons for declaring our definition to be definitive were touched on in Section 1: Its origin in the text information retrieval literature and the fact that it allows us to make a clear differentiation between the "what" dimension of a user's information need and the "why" dimension. In this section, we cover MIR research that uses the term "intent" or "intention" in unexpected or alternate ways and discuss how it is related to intent-aware MIR as we have defined it here.

Covering this literature in this survey is important for two reasons. First, it serves to further clarify the concept of user intent in MIR as discussed in this survey. Second, and more importantly, the literature represents key techniques for improving multimedia search. Once it is clear that these techniques do not actually address user intent in the form of the "why" dimension, it is a fairly obvious next step to adapt them to ensure that they do.

Research in the area of text information retrieval or MIR that uses an unexpected or alternative definition of intent falls into three categories.

—*Query ambiguity*. This literature uses the term "intent" in the sense of "user intention" or "user information need." The goal of this work is to clarify the topical component of the information need expressed in the query.
—*Clarity of user needs and goals*. This literature examines the degree to which the user information need is well focused. It differentiates users who are "just browsing" from those who have highly specific goals, and it can be considered to apply to both the "what" and the "why" dimensions of the user information need.
—*Personalization*. This literature uses the term "intent" in the sense of "long-term user interests." User profiles encoding long-term interests are used to support interpretation of the current information need. Such profiles can be considered to apply to both the "what" and the "why" dimensions of the user information need.

The danger of using alternate definitions of "intent" is that the focus is set on the "what" dimension, and explicit consideration of "why" is abandoned. Such research will lead to MIR approaches that fail to cover the entire spectrum of information needs.

3.1. Query Ambiguity

In MIR, as in text IR, ambiguity in queries formulated by users is a large challenge that must be faced in order to improve search engines. It is helpful to distinguish two phenomena that can be thought of as query ambiguity. The first is "true" ambiguity and is described as related to homonymy, the phenomenon of words having multiple different and unconnected meanings. Homonyms are words that generally receive two distinct dictionary entries, such as "bar," "digest," "sewer," "sign," "wind," and "yard." The second is ambiguity of aspect. The term "aspect" is used in the IR literature to designate component topics that contribute to composing an overall topic.

The distinction is well illustrated by Santos et al. [2011]. This work provides the example of `zeppelin` as a query that is truly ambiguous: The word designates a rigid aircraft, but also a band, "Led Zeppelin." The query `Led Zeppelin`, however, does not have the same kind of "true" ambiguity. Rather, it is associated with a variety of aspects (e.g., the history of the band, website of the band, music downloads, influence on music today, recent news events involving the surviving band members). To be successful, a search engine needs to infer which aspects are of most interest to the user or present the user with a diverse mix of all of them [Clarke et al. 2008]. Aspects have been described in the literature as interpretations of the query [Santos et al. 2011] or as answers to the questions that are effectively posed by the topic [Over 1996]. We point out that topic aspects are a manner of interpretation. The definition of aspect used by the *TREC Interactive task*[3] makes it clear that the aspects of a topic will differ from topic to topic.

There is a close link between the study of aspects of a topic that are relevant to a user's query and the intent of the user. This link is pointed out by Santos et al. [2011], who propose an intent-aware approach to result diversification for conventional text information retrieval. Following the taxonomy of Broder [2002], which differentiates "navigational," "transactional," and "informational" queries (cf. Section 4 for details regarding this taxonomy), they describe the link between query aspects and user intent. Specifically, they point out that the aspect "website of the band" belonging to the topic "Led Zeppelin" is linked to a navigational intent, the aspect "music downloads" to a transactional intent, and the aspect "history" to an informational intent. The closeness of the connection between aspect and intent makes it understandable why some literature has conflated the two. This section is devoted to teasing apart work that has looked at topic, with the goal of ensuring that the "why" dimension of information needs is not overlooked due to inconsistent application of terminology.

We start by providing an example supporting our position that although topical query aspects (i.e., the "what" dimension) might appear to be closely linked to the "why" dimension, the two should be conceptualized as independent of each other. This example shows that a specific topical aspect can be associated with two distinct goals of the user, although one goal might seem, upon first consideration, to be dominant. The example returns to the query `Led Zeppelin` already discussed. The topic query aspect `Led Zeppelin music downloads` may often indeed have a transactional intent, but the goal of the user could also be informational (i.e., checking in how many different places such downloads are available). Similarly `Led Zeppelin history` might often have an informational intent, but it could be that the user has a transactional goal (i.e., wishing to buy a book or a documentary video for a friend's birthday). Our argument is that

[3]http://trec.nist.gov/data/t8i/t8i.html.

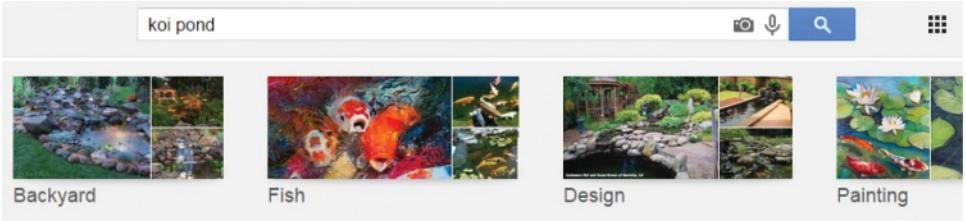

Fig. 5. Example of query disambiguation results for the query `koi pond` provided by a mainstream image search engine.

maintaining intent as a separate concept that is tightly coupled with the user's search goal and that is independent of topic will avoid the danger of missed opportunities raised by conflation of intent and aspect.

In the remainder of this section, we cover three major classes of approaches aiming at addressing query ambiguity: query suggestion, search result diversification, and query and ranking refinement.

3.1.1. Query Suggestion. In image search, Zha et al. [2010] propose a visual query suggestion approach that, given an initial textual query, suggests related textual queries along with representative query images with visual correspondence to the suggested textual query. After the user selects a query text/image pair, the retrieval approach then exploits both query types using text-based and visual search in order to make the user's query more specific and, in this way, bridge what the authors refer to as the "intent gap" and what we would call the "intention gap." Bian et al. [2012] tackle the problem of visual query suggestion for query-by-example based systems by proposing an approach that automatically suggests informative attributes for user-provided query images.

Such approaches have been proposed for video search as well. Li and Li [2011] investigate a "multimodal query suggestion method for video search" that, given an initial textual query, supplies textual query suggestions along with illustrative image examples. Feng et al. [2010] propose a "multimodal query expansion" framework that converts textual queries to visual queries using clustering techniques.

These approaches clearly provide valuable support for users engaging in multimedia search. However, they do not support intent-aware search. In other words, they do not address the reason *why* users are searching. Their shortcoming in this regard can be seen by considering Figure 5, which shows query disambiguation results for the query `koi pond` provided by a mainstream image search engine. Given a possible search intent of the user who issues the query `koi pond` (e.g., obtaining information about koi ponds or providing users with an explanation of how to build a koi pond), the suggested query aspects (e.g., `backyard`, `fish`, `design`, and `painting`) provide an overview of the topic but clearly need to be further refined if they are to address "why" the user is searching for "koi pond" (i.e., specifically support the user in carrying out a task).

3.1.2. Search Result Diversification. Diversification approaches attempt to provide users with many different types of results at the top of the results list, in hopes that at least one of the types will tightly fit the users' information needs. Previously, we mentioned that the results list in Figure 3 had been diversified with respect to intent. Search result diversification clearly serves to support users in finding multimedia content. However, much of the literature treats diversification in terms of topical query aspects, although the authors might refer to their work as involving "intent" or "intention." We briefly review some of the key contributions.

Given a textual image search query, Hoque et al. [2013] perform automatic query expansion using different topic-sensitive sets of concepts extracted from *Wikipedia*,

resulting in a broad range of results images that represent the various possible interpretations of the query. An image organization method then assists users to find images that match their "intention." van Zwol et al. [2008] propose an approach that considers diversity as a property of the initial retrieval model by sampling additional contextual terms from tags and other user-contributed image metadata that are used for automatic query expansion. Their experimental evaluation shows that diversity in image search results lists can be achieved with no negative consequences in terms of results precision. Similar investigations perform diversification that predominately relies on topical visual (dis-)similarities (e.g., van Leuken et al. [2009]).

3.1.3. Query and Ranking Refinement. Finally, we turn to look at work that attempts to refine either the query or the ranked list resulting from the query. Again, these methods clearly serve to improve the relevance fit between the user's information need and the results list. However, here again, we point out that many authors are actually carrying out refinement with respect to topical query aspects, and, although they might use the terms "intent" or "intention," are not specifically focusing on taking the "why" dimension of user search into account.

Representative examples of such work are the following: Guan and Qiu [2007] present an approach for "modeling user feedback intentions in interactive image retrieval." Given a query image, the approach considers topic-sensitive segments of the image that fit the user's intention. Zhang et al. [2013] exploit semantic descriptions and properties of visual concepts (e.g., "round," "metallic" etc.), link them in a hierarchic fashion to obtain better representations for images, and collect user feedback in an online search session to produce refined rankings.

For completeness, we mention other work targeting query and ranking refinement using click data. Yang et al. [2012] propose a similar retrieval model extracting features from local (i.e., click-through data in a user's session) and global (i.e., data comprising all images and associated Web pages in a given corpus) context. Jain and Varma [2011] provide models that are capable of promoting images in reranked search results lists that are likely to be clicked. Yang et al. [2014] combine image click-through data with several modalities in a multimodal graph which builds the basis to learn a reranking model. Pan et al. [2014] exploit click-through data to learn latent subspaces between textual queries and images.

We point out, and differentiate, work that moves toward characteristics of the image itself (e.g., image type or composition), rather than of the depicted content of the image. Cui et al. [2008] present an interactive approach that automatically reranks initial search results lists produced by user-specified query images and automatically inferred "user intentions." Here, the authors focus on high-level topical and semantic classes extracted from images returned in initial results lists. These classes are *"general object," "object with simple background," "scene," "portrait,"* and *"people"* [Cui et al. 2008]. The reranking step is performed by inferring the query classes from the provided query image and then by ranking result images according to how well they represent the inferred classes. Hua et al. [2013] represent images using eleven categories including *"black & white," "portrait photo," "clipart," "high visual quality," "line drawing," "indoor," "outdoor," "cityscape," "landscape," "human full body," and "head & shoulder."* The reranking step extracts features related to these categories from previously-clicked images representing a query and ranks initial results with respect to the co-occurrence of the extracted categories. Wang and Hua [2011] propose an interactive image search approach that allows users to specify the color distribution of the image they have in mind in addition to submitting a textual query. The sketch-based information provided by the user is then used to rerank initial search results produced by the textual query.

Such systems clearly are moving beyond the topical content of images and can support users in formulating queries that reflect their goals (e.g., a user with an intent to create a presentation might look for clipart). However, further benefit is to be gained from query and ranking refinement approaches if they can be developed explicitly in order to address user intent.

3.2. Clarity of User Needs and Goals

Work on clarity of needs and goals starts from the insight that not all users have information needs that are concretely formulated prior to beginning multimedia search. An important contribution was made by Choi and Rasmussen [2003], who investigate user needs in image search from the perspective of how specific or general they are. In particular, they manually categorize user needs into one of four classes: *"specific"* needs can be expressed in keywords and as a precise, unambiguous query. *"General/nameable"* needs can be expressed as a query, but often need to be made more concrete. *"General/abstract"* needs may include concepts that are rather abstract and that cannot necessarily be expressed as a specific query, although they typically can be expressed verbally. Finally, *"subjective"* needs are challenging to even convey verbally. Another important contribution was made by Datta et al. [2008], who study the specificity of goals (i.e., how specific the expectations of a user are concerning the images that are being searched for). They distinguish among three categories related to three degrees of goal clarity. *"Browsers"* are users browsing for pictures with no concrete idea of success. *"Surfers"* engage with search engines with a moderate clarity of a goal, while *"Searchers"* are very clear about what they are searching for. Kogler and Lux [2011] propose an adaptive retrieval approach that, similar to Datta et al. [2008], exploits the degree of clarity of goals. The authors conjecture that visual diversity of search results can be related to clarity of goals. They propose that the vocabulary size in bag-of-visual-words representations of images corresponds with the clarity of user goals (i.e., smaller vocabularies model vague intents and larger vocabularies model more specific intents).

In general, work on the clarity of users' needs and goals addresses simultaneously the "what" and the "why" dimension of user information needs. For this reason, this work can be seen as making an important contribution to intent-aware MIR, as it is defined in this article.

3.3. Personalization

IR systems often build up profiles of users using long-term search behaviors. These profiles encode characteristics of the user, which can help to serve the user better across a variety of different information needs with which the user might approach the system.

Trevisiol et al. [2012] evaluate the importance of features that have been derived from users' (long-term) browsing behavior for image ranking. Similar efforts [Cui et al. 2014; Liu et al. 2014] investigate the possibility of applying "user interests" that have been automatically derived from social media platforms in the context of search results list reranking in image search. Teevan et al. [2008] present an approach that automatically identifies queries that could be enhanced through personalization.

This work focuses on building user profiles that are related to the topics that interest users and often use the concepts *interest* and *intent* synonymously. However, a priori, personalization is not restricted to supporting the "what" dimension of user information need. It would be straightforward to also extend user profiles to encode typical reasons for which a user searches; in other words, to add the "why" dimension of user intent, which is the focus of this survey.

3.4. Summary and Insights

This section has clarified the definition of intent. In the process, it has covered techniques for addressing the ambiguity of queries and vagueness of user needs and goals, and for integrating long-term user interest. In the literature, many authors refer to "intent" or "intention," but focus on the "what" dimension of user information needs and either de-emphasize the "why" dimension or ignore it completely. We have argued that clarifying terminology, as we do here, is important to ensuring that the "why" dimension receives the attention it deserves. Furthermore, these approaches make use of various data sources (e.g., context, click-through data) and techniques (e.g., reranking, relevance feedback) that can be straightforwardly extended to apply to intent-aware MIR, that does take the "why" dimension into account.

4. CONCEPTUAL MODELS OF USER INTENT

To exploit the full potential of intent-aware algorithms, it is necessary to first fully grasp the nature of the "immediate reason, purpose, or goal behind a user's information need" that constitutes intent (see Hanjalic et al. [2012]). In other words, we need to understand the different reasons why users consult multimedia search engines and develop corresponding categories covering these reasons. In the literature, a relatively large number of different conceptual models and typologies has been proposed that cover and explain different "intent types" (i.e., reasons why users search for information). These models and typologies are composed of intent classes that attempt to cover important types of user intent that are important for search. Modeling intent types is important to provide the basis for intent-aware MIR technology that covers the whole breadth of user intent in multimedia search.

As mentioned in Section 1, there is a close connection between text-based IR and MIR, although the two must ultimately be treated as fundamentally different. For this reason, we approach conceptual models of user intent by first covering work that has been proposed in the area of conventional text-based Web search. We then move on to cover the investigations, relatively fewer in number, that have been carried out in the multimedia domain, in particular, in image and video search. Table I gives an overview of the conceptual models of user intent in the domains of text, image, and video search that we discuss in the subsequent sections. It includes information about the methodologies that have been followed to arrive at specific categories that cover different reasons why users consult search engines. For completeness, we also include a review of conceptual models of intent in general search activities in this section, going beyond search engines. We conclude by summarizing insights obtained that serve as motivation for future key challenges discussed later in Section 6.

4.1. Conceptual Models in Conventional Text-Based Web Search

The first widely known taxonomy of user intents in textual web search was proposed by Broder [2002]. As briefly mentioned in Section 1, Broder determined that users consult search engines for reasons other than simply finding information. His investigation results in a taxonomy containing three basic intent categories: *"navigational," "transactional,"* and *"informational."*

Navigational queries have the purpose of reaching a particular website that users have in mind, either because they assume that such a site exists or because they visited this site in the past. Queries in this category are closely related to "known item" search queries in the classic sense of IR, since they target a single relevant search result. A typical example of a navigational query is if the user does not type the URL of a website directly in the browser, but rather uses a search engine to get redirected to this website.

Table I. Categorization of Conceptual Models of User Intent in the Domains of Text, Image, and Video Search

Paper	Methodology	Categories
Text-based search		
Broder [2002]	Survey and Query Log Analysis	*"Informational"*; *"Navigational"*; *"Transactional"*
Rose and Levinson [2004]	Query Log Analysis	*"Navigational"*; *"Informational"* ("Directed," "Undirected," "Advice," "Locate," "List"); *"Resource"* ("Download," "Entertainment," "Interact," "Obtain")
Baeza-Yates et al. [2006]	Query Log Analysis	*"Informational"*; *"Not informational"*; *"Ambiguous"*
Morrison et al. [2001]	Survey	*"Find"* ("Find Download," "Get a fact," "Get a document," "Find out about a product"); *"Compare/Choose"*; *"Understand"*
Image search		
Lux et al. [2010a]	Query Log Analysis and Interviews	*"Knowledge Orientation"*; *"Mental Image"*; *"Navigation"*; *"Transaction"*
Fidel [1997]	User study	"*Data* pole"; "*Object* pole"
Video search		
Hanjalic et al. [2012]	Social-Web Mining and Crowdsourcing	*"Information"*; *"Experience: Learning"*; *"Experience: Exposure"*; *"Affect"*; *"Object"*
Lagger et al. [2012]	Surveys and Interviews	*"Information"*; *"Education"*; *"Entertainment"*
Cunningham and Nichols [2008]	Interviews	*"Mental state"*; *"Visual"*; *"Audio"*; *"Learning"*; *"Social"*; *"Mainstream media"*; *"Temporal"* (Note that these categories have some overlap with the topical components of users' information needs)

Transactional queries have the purpose of reaching a website that will allow the user to engage in a particular interaction (e.g., to obtain a specific product or download various types of files). Although transactional queries share similarities with navigational queries in the sense that users may have a particular website in mind, for transactional queries, any website that makes possible the transaction is considered relevant.

The goal of informational queries is to acquire information about a topic. While for navigational or transactional queries users typically engage in further interaction, informational queries typically do not trigger such interaction, except that of reading the content of the relevant results. Similar to transactional queries, more than just a single result is considered relevant. In fact, typically, a good set of different on-topic search results is preferred over a single document.

To study the allocation of queries to the three established classes, Broder [2002] makes use of user surveys that are presented to random users who submit requests to *AltaVista*. It is important to ask users about their goals immediately because this information is situation-specific and difficult to recover from the query. Broder finds that 24.5% of queries were submitted with a navigational intent and about 36% with a transactional intent. The remaining 39% are considered to be informational queries. In addition to the user surveys, Broder investigates 400 random queries submitted to AltaVista by manually assigning a category to queries without asking users who submitted them about their intents. He determines a distribution of 20%, 30%, and 48% over the three classes.

Rose and Levinson [2004] conduct a similar investigation and develop a set of goal (i.e., intent) categories in an iterative fashion, also using queries from the AltaVista search engine. The result is a taxonomy of three high-level intent categories that extends Broder's scheme: *"resource," "navigational,"* and *"informational."* Resource intent corresponds to transactional intent, but with a broader focus (i.e., including intent subtypes of *"download," "entertainment," " interact,"* and *"obtain."* The fact that the user who submits a navigational query to the search engine aims to visit a specific target

page that this user is aware of establishes the difference between the navigational category proposed by Rose and Levinson [2004] and the navigational category introduced by Broder. Informational intent is also similar to Broder's informational class, but Rose and Levinson [2004] introduce a finer-grained subcategorization to accompany different ways of obtaining information. With *"directed"* informational intent, users want to learn something particular about a topic; with *"undirected"* intent they want to learn anything/everything about it. *"Advice"* is comparable with the *"undirected"* class but focused on getting advice, ideas, suggestions, or instructions. *"Locate"* and *"list"* are informational intent types directed at getting a list of suggested websites.

To derive a distribution over the three established classes, Rose and Levinson [2004] manually label 1,500 queries based on the query string itself and the produced and clicked search results, as well as additional search session characteristics. They obtain a distribution of 13.5% navigational queries, 24.5% resource-oriented queries, and 62% informational queries.

Baeza-Yates et al. [2006] investigate 6,000 queries submitted to the *Yahoo!* search engine and propose three categories covering why users search: *"informational," "not informational,"* and *"ambiguous."* Similar to Rose and Levinson [2004], the *"informational"* category follows the definition of Broder's initial category (i.e., to obtain information independently of what topic the query is about). *"Not informational"* queries consolidate Broder's navigational and transactional queries, and for *"ambiguous"* queries no goal can be inferred directly (e.g., because users may not have clearly formulated goals). Manually labeling the queries using these three categories resulted in 61.5% of the queries being informational, 21.6% being not informational, and 16.9% being ambiguous.

For completeness, we also mention work that predates Broder [2002]. Morrison et al. [2001] investigated 100 responses to a user survey addressing why (*"purpose"*) and how (*"method"*) users search, as well as for which content they are looking (*"content"*). For the *"purpose"* category (i.e., the category we are interested in here), they establish a taxonomy of three main reasons: to *"find,"* to *"compare/choose,"* and to *"understand."* The *"find"* category is further divided into intent subtypes *"download information," "get a fact," "get a document,"* and *"find out about a product"* [Morrison et al. 2001]—informational subcategories covered by taxonomies proposed by Broder [2002] and Rose and Levinson [2004]. The *"compare/choose"* category constitutes queries that were submitted to obtain information for the user to make a choice, and *"understand"* represents queries with the goal to locate facts or documents in order to understand a topic. The distribution over the three categories is 25% *"find,"* 51% *"compare/choose,"* and 24% *"understand."*

4.2. Conceptual Models in Image and Video Search

We now turn to intent models that have been proposed for multimedia search. We start by noting the differences that have been observed in the literature between text-based IR and MIR. Several investigations [Spink and Jansen 2005; Tjondronegoro et al. 2009; André et al. 2009] point out that there is a significant difference between text-based web searching and searching for multimedia documents such as images, videos, and audio/music. In André et al. [2009], the observation is made that image search is more exploratory than Web search. These investigations are consistent with the observation in Spink and Jansen [2005]: "multimedia searching appears to require greater interactivity between users and the search engine." Compared to general web search, multimedia searching shows a significant increase in the number of query terms, search session length, query reformulations, and number of search results clicks. Similar observations have been made in Halvey and Keane [2007]. Important for our investigation of user intent, "the range of information needs in multimedia search

appears to be growing broader" [Spink and Jansen 2005], meaning that user intent schemes for general Web search, such as those covered earlier, are not necessarily applicable for multimedia search. Kofler and Lux [2009a] make similar observations and find that the taxonomies proposed by Broder [2002] and Rose and Levinson [2004] cannot be applied in multimedia search (in this particular case, Kofler and Lux [2009a] focus on image search) without adaptation: Because of the text-based context of these taxonomies, many goals in multimedia search cannot be assigned to a specific class, and the definitions of these classes are not always applicable in multimedia search. Although user intent is clearly critical if MIR is to satisfy users' information needs to their full extent, surprisingly little work has been carried out to establish intent taxonomies for multimedia search. Here, we summarize the few investigations that have been carried out.

Lux et al. [2010a] study user intent in image search by performing a small-scale query log analysis on a corpus generated by a user study. They define eight image search tasks covering different user intent types and ask study participants to find images on the *Flickr* photo-sharing platform that fulfill these goals appropriately. They record the interaction of study participants with Flickr, and the analysis of the resulting query log was confirmed by another round of user studies and interviews. The outcome of their research is a typology of four intent types [Lux et al. 2010a]: *"knowledge orientation"* (users have the goal to "obtain knowledge"), *"mental image"* (users have the goal to find a specific image they have in mind), *"navigation"* (users have the goal to locate a particular image before knowing its visual content), and *"transaction"* (users have the goal to obtain an image and reuse it as an object through, e.g., downloading).

The intent classes discovered in Lux et al. [2010a] are consistent with results obtained from earlier investigations carried out by Fidel [1997]. Fidel investigates 100 image search tasks and finds that they can be distinguished with regard to two *poles*. The "*data* pole" covers cases where the user's goal is to obtain information or knowledge from images, and the "*object* pole" covers cases where the user's goal is to retrieve actual images as objects for further use.

Hanjalic et al. [2012] investigate user intent in the domain of video search. They employ a social-Web mining approach and collect real-world video search information needs from the popular question-answering platform, Yahoo! Answers. The information needs that are collected are filtered by crowdsourcing workers, and only requests that contain an explicit expression of a user search goal are retained. These information needs (i.e., 280 needs remained after the crowdsourcing process) are then manually coded. Intent classes are established by iteratively comparing and contrasting discovered classes, merging and dividing the classes until each contains descriptions that characterize a single type of user intent.

The iterative coding process results in a taxonomy of five intent classes. *"Information"* (users focus on obtaining "declarative knowledge"; i.e., "facts or explanations that can be expressed declaratively and constitute *knowing that*" [Hanjalic et al. 2012]); *"Experience: Learning"* (users focus on obtaining "performative knowledge"; i.e., "*knowing how* to do something rather than *knowing that*" [Hanjalic et al. 2012]); *"Experience: Exposure"* (users focus on "gaining exposure to the experience of a specific real-life event, or to a certain type of experience" [Hanjalic et al. 2012]); *"Affect"* (users focus on shifting their mood); and *"Object"* (users focus on obtaining video content to be used for a specific purpose in a real-world situation).

Note that the approach applied by Hanjalic et al. [2012] is different from conventional methods such as those using query log analysis (e.g., Baeza-Yates et al. [2006]), interviews (e.g., Lux et al. [2010a]), or surveys (e.g., Broder [2002]) because it uses evidence from a wide user population, and it also directly accesses spontaneously expressed

information about why users search for videos. Using direct and spontaneously expressed search goals helps to minimize wrong interpretations in the coding of queries.

We would like to draw special attention to two aspects of the work carried out in Lux et al. [2010a] and Hanjalic et al. [2012]. First, some commonalities are to be observed between taxonomies proposed for image and video search, and text search. Second, while distinct classes were discovered, Lux et al. [2010a] point out that they are not necessarily mutually exclusive, but rather overlapping. For example, it is possible that a user approaches an image search task with a combined intent of gaining knowledge from images' contents (i.e., knowledge orientation) by already having a particular image content in mind (i.e., mental image). On the other hand, some intent classes clearly have no overlap (e.g., mental image and navigation). The same observation can be made about the classes derived in Hanjalic et al. [2012]: A given query may provide a certain level of fit with a several types of search intent.

For completeness, we mention work that goes beyond search intent to look at other forms of intent related to other ways in which users interact with multimedia. Lagger et al. [2012] carry out an exploratory user study with a similar objective, studying why users *watch* (i.e., not search for) online video. They obtain similar results and three main reasons for users to consume online videos: *"informational reasons," "educational reasons,"* and *"entertainment reasons."*

Cunningham and Nichols [2008] carry out interviews investigating video information needs and create a high-level categorization scheme that, to some extent, blends the user intent (i.e., reason behind the user's information need) with its topical component. Their scheme contains the following classes: *"mental state"* (emotional state or mood, e.g., being bored), *"visual"* (visual characteristics of the desired material), *"audio"* (audio characteristics of the desired material), *"learning"* (to learn something), *"social"* (desired material explicitly suggested by acquaintance), *"mainstream media"* (to regularly receive updates, e.g., on sports events), and *"temporal"* (references to planned events in the future; e.g., listen to music to get into the mood for attending a concert).

4.3. Conceptual Models in General Search Activity

Although this survey concentrates on multimedia search and multimedia search engines, we point out that user search behavior is not restricted to a particular form of content or to a particular way of finding that content. Instead, when users search, they look for information of content that takes any modality (e.g., spoken, text, image, video) and go beyond a particular information retrieval system or web search engine vertical. The work in this area is relatively sparse since it is difficult to study search completely independently of a certain type of data or system. However, here we mention a couple of key examples. We find it important to include these, since MIR is ultimately part of the larger picture of how users fulfill their information needs.

Jean et al. [2012] contributed significantly to this field by investigating goals of users' online search activities in general (i.e., not limited to specific scenarios). These authors perform user studies employing diary methods and ask study participants (417 unique individuals) to keep a diary of their online search activities. The user study requires participants to explain, by means of open-ended questions, their goals when engaging in online activities. The investigation covers user goals that span multimedia information needs, defined as "users' personal goals that motivates them to conduct a particular information activity" (which they refer to as "goals"), but they also look at finer grained goals and, particularly, at any "subgoal that a user has to achieve while engaging in an information activity" (which they refer to as "intentions"). Manual analysis of the open-ended questions results in 11 goal categories and 9 intention categories. Goal categories are *"Buy," "Connect with people," "Entertain," "Get employed," "Help other people," "Maintain household and electronics," "Perform*

school-related task," "Perform work-related task," "Plan for future," "Self expression," and *"Sell,"* while intention categories are *"Decide," "Evaluate," "Gather data," "Keep up to date," "Learn," "Manage personal information," "Produce," "Share,"* and *"Verify."* Note that discovered "search-related" activities, such as learn or entertain, are consistent with findings of investigations of user intent in search scenarios discussed earlier in this section.

Another important contribution is that of Chung and Yoon [2012], who investigate users' general multimedia needs and searching behavior on the Web by performing a user study and interviews with 20 participants. The authors carry out manual coding of participants' responses, a process that yields four categories: *"informational needs," "illustration," "entertainment,"* and *"download." "Informational needs"* covers cases where users search for information in order to acquire knowledge. This category applies, in particular, to both image and video search. *"Illustration"* needs cover cases where users would like to demonstrate a concept. These needs were observed particularly in the context of image search. The *"entertainment"* (to be entertained) and *"download"* (obtain a particular resource) categories are observed to be applicable in the video, image, and audio search context.

4.4. Summary and Insights

We close this section by summarizing our observations in the form of a list of insights.

(1) Both in the field of conventional text-based Web search and, to a more limited extent, in the field of MIR (image and video search), typologies and models of user intent have been introduced.
(2) Consistencies can be observed across typologies and models proposed by different authors. There are also consistencies between the cases of text and multimedia. However, it is important to treat user intent in MIR as distinct from user intent in text search.
(3) It is important to take into account that user information needs might involve overlapping intent types.
(4) The discovery of intent categories in MIR has focused mainly on the domains of images and video. Intent typologies for several other major multimedia content types, such as speech, audio, and music, have not yet been thoroughly explored.
(5) The granularity of the intent typologies or models varies. The appropriate granularity will depend on the domain and the application. We anticipate that specific applications (e.g., in education, journalism, engineering, or medicine) might benefit from finer-grained intent categories.

The insights on user intent models and typologies are a valuable basis for developing intent-aware MIR approaches (Section 5) and also for identifying future challenges on intent-aware multimedia search (Section 6).

5. USER INTENT IN THE MULTIMEDIA SEARCH PIPELINE

In this section, we build on intent as the "why" in the user information need and the typologies and models of user intent covered in the previous section to investigate the algorithms and approaches that have been proposed in the MIR research community that can contribute to intent-aware multimedia search. We discuss retrieval algorithms that have been proposed throughout the entire multimedia search pipeline presented in Section 2, spanning intent-aware indexing over intent-aware query processing to intent-aware search results list ranking. The overview provided in Figure 4 organizes the material covered in this section by position in the pipeline (indexing, query processing, or ranking) and by the type of search addressed (image or video). We consider each of the steps in the multimedia pipeline in turn. For each step, we cover approaches that

Table II. Representative Examples of User Intent-Aware Approaches Proposed in the Multimedia Informational Retrieval Research Community, Organized by Search Engine Pipeline Steps (Corresponding to the Subsections of Section 5) and Multimedia Domain

Pipeline step	Image search	Video search
Indexing	–Typologies of creation- [Kindberg et al. 2005; Lux et al. 2010b], tagging- [Ames and Naaman 2007], and uploading- [Van House 2007] intents; –Prediction of framing intent [Riegler et al. 2014]	–Typologies of creation- [Campanella and Hoonhout 2008; Lux and Huber 2012] and uploading- [Bornoe and Barkhuus 2010] intents; –Prediction of creation- [Mei and Hua 2005] and uploading- [Kofler et al. 2015] intents
Query process.	–Query classification using search session data [Kofler and Lux 2009b]; –Linking search behavior and query types with classes of search intent [Park et al. 2015]	Query classification using click-through- [Moshfeghi and Jose 2013] and ranking- [Kofler et al. 2014] data
Ranking	Results ranking using click-through data [Lu et al. 2014]	–Results filtering using intent-genre mapping [Lagger et al. 2011]; –Results reranking/diversification using initial ranking [Kofler et al. 2014]

focus on image and video search (an overview of representative examples are provided in Table II) and draw connections to intent-aware approaches introduced in conventional information retrieval. We conclude each subsection with a short summary and insights into how the individual pipeline steps can be further improved from an intent-aware perspective, thus laying the groundwork for the discussion of future challenges in Section 6.

5.1. Intent-Aware Indexing

The main purpose of the indexing component of a multimedia search engine is, as mentioned in Section 2, to represent multimedia content in a way that enables effective and efficient access. Intent-aware approaches to indexing may make use of topic-sensitive features, such as feature representations that were optimized to retrieve multimedia documents with respect to topic. However, to achieve an even closer fit between the intent behind a user's information need and the representation of multimedia documents in the index, a system should specifically make use of intent-sensitive features. The goal of intent-sensitive features is to provide representations that abstract away from the documents' content and expose characteristics of multimedia documents that are associated with intent.

One obvious approach to intent-sensitive indexing is to associate documents directly with possible intent types in the form of inferred category labels. However, this requires predicting the different types of intent with which a user might search for a multimedia document in advance. User intents are varied and unexpected, so inferring this information at the time that a multimedia document is indexed initially appears impossible. Furthermore, over time, we can expect that as information needs develop, the documents relevant to certain types of user intents will change. For this reason, we are interested in indexing documents not only with predicted search intent categories, but also with any information that might possibly be useful for matching documents with user search intent at query time. Such information has the potential of allowing search to more flexibly match user search intent with documents.

We base our discussion of intent-aware indexing on findings of, for example, Mei and Hua [2005], Riegler et al. [2014], and Kofler et al. [2015], which indicate a connection between several kinds of creation-related intents and user search intent. In this section, we give an overview of different types of multimedia intent and the contribution that

they may make to indexing documents in a way that is sensitive to the search intent behind users' information needs.

5.1.1. Image Production and Sharing. In the domain of image production, Kindberg et al. [2005] use interviews to investigate why users take pictures on mobile devices. They discover that the main reasons are *"sharing"* and *"personal"* use, as well as *"affective"* and *"functional"* use. Lux et al. [2010b] also perform interviews to investigate the reasons why people take pictures and arrive at similar results to those presented in Kindberg et al. [2005]—specific classes of creation intent are to *"preserve an emotion," "support a task," "recall a situation," "share with family and friends,"* and *"publish online."* Lux and his colleagues release a dataset containing Flickr images and their intent-annotated images [Lux et al. 2012]. Riegler et al. [2014] investigate "intentional framing"—the sum of choices a photographer makes on how to represent the main content of an image. They argue that global features derived from images capture differences in framing strategies, which are chosen by the photographer as a reflection of the reason for which the picture was taken.

In the domain of image sharing and uploading, initial work was carried out by Marlow et al. [2006], who present an investigation of the usage of tags on the Flickr photosharing platform but do not yet investigate the factors motivating users to tag photos. Ames and Naaman [2007] investigate why users tag photos uploaded to online photo-sharing platforms and encounter a classification along two main axes: *"functional"* (covering reasons to *"organize"* or to *"communicate"*) and *"sociality"* (covering *"selfish"* reasons as well *"closed group activities"*). Van House [2007] investigate the use of photos on the Flickr photo-sharing platform and discover patterns representing four main reasons for photo uploading and sharing: *"memory, narrative, and identity"*; *"maintaining relationships"*; *"self-representation"*; and *"self-expression."*

5.1.2. Video Production and Sharing. In the domain of video production, Campanella and Hoonhout [2008] investigate why users capture home videos and find that the main reasons are to *"keep memories of someone's life"* and to *"share experiences"* with family members and friends. Mei and Hua [2005] carry out a similar study and propose an approach to automatically derive "capture intentions" from home videos to ultimately support "intention-based home video browsing." Unlike the categories proposed by Campanella and Hoonhout [2008], the classes proposed by Mei and Hua [2005] go beyond "why" and also contain topical properties of the captured videos: *"static scene," "dynamic event," "close-up view," "beautiful scenery," "switch record," "longtime record,"* and *"just record."* Lux and Huber [2012] carry out semi-structured interviews to explore user intent related to video production. They arrive at five classes including *"preservation," "sharing," "affection," "functional,"* and *"technical interest."*

A few studies investigate the reason behind uploading. Bornoe and Barkhuus [2010] investigate the reasons for video microblogging and find the main goals to be *"self-expression," "entertainment,"* and *"self-presentation."* Cheng et al. [2007] carry out an analysis of properties of YouTube videos including length of videos belonging to different categories and growth trends in uploading, views, and ratings. Park et al. [2011] carry out surveys to study aspects that are related to intents in uploading videos to the Internet and find that uploading behavior is specifically associated with ego-involvement of users.

Kofler et al. [2015] investigate users' uploader intent for video-sharing platforms such as YouTube and discover, by means of social-Web mining combined with crowdsourcing five high-level uploader intent categories: *"explaining," "sharing," "promoting," "communicating,"* and *"entertaining."* The authors propose an automatic classification approach using these classes based on multimodal features extracted from textual metadata as well as visual features. The algorithm assigns a distribution of confidence

scores, each score corresponding to a particular uploader intent class, to classified videos. The intent class that has the best fit with a video is recognized to be its "dominant intent."

Although not directly related to video-sharing and uploading, some investigations have been carried out that study why users comment on multimedia documents made publicly available [Vliegendhart et al. 2013; Madden et al. 2013], which might contribute to the understanding of why users engage with these documents.

5.1.3. Connection to Conventional Text-Based Web Search. In the text domain, several investigations have been performed that study the motivation behind different actions. These include why people blog [Nardi et al. 2004], why they tweet [Java et al. 2007], and why people use Q&A sites [Gardelli and Weber 2012]. Other investigations focus, for example, on tagging resources (e.g., bookmarks) regarding their "purpose" [Strohmaier 2008]. Dai et al. [2011] investigate "link intent," categorizing links based on whether they either describe the target page's identity or its content.

5.1.4. Summary and Insights. This section has surveyed work on intent-aware indexing and allows us to arrive at a number of insights:

(1) In order for a multimedia search engine to fulfill the user's search intent, multimedia documents must be indexed with features that are sensitive to intent.
(2) One type of intent-aware feature is intent categories drawn from intent typologies that correspond to intent types and abstract away from document content.
(3) One important way of arriving at intent types is to look at different forms of user intent associated with multimedia documents, particularly image and video production and sharing.
(4) The key to using other forms of intent is to understand their relationship to user search intent.

We build further on these insights when we discuss future challenges for intent-aware indexing in Section 6.

5.2. Intent-Aware Query Processing

The main purpose of the query processing component of a multimedia search engine is, as discussed in Section 2, to prepare the user's query so that it is compatible with the data that have been indexed in the offline processing step. In light of the insights that we have gained concerning intent-aware indexing in the previous section, the main challenge of intent-aware query processing is to represent the query in a form that exposes the user's search intent. Work addressing this challenge can be divided into three major classes:

—*Query classification* matches queries to search intent classes that share common characteristics and that can be exploited by query class-dependent retrieval models and ranking strategies;
—*Query suggestion* provides queries to users that aim to better express their search intent given their initial queries; and
—*Query modification* modifies or expands initial queries such that the ranking component is able to return documents that are maximally relevant to users' intents.

We now look in turn at intent-aware image and video retrieval, discussing approaches that have been developed in each of the three classes. We then take a look at the three classes of approaches in the area of intent-aware text-based information retrieval and relate this work to opportunities in the area of MIR.

5.2.1. Image and Video Search. The query processing approaches in the context of image and video search have mainly adopted the query classification paradigm. Kofler and Lux [2009b] investigate a small-scale query log of user interactions with the photo sharing platform Flickr and derive a rule-based classifier that automatically classifies a user's intent into the predefined intent categories established in Broder [2002]. Because a previous study indicated that features derived from the query string itself are typically not indicative of the user's intent in image search [Kofler and Lux 2009a], the authors apply classification features that are derived from the user's search behavior. In particular, click-through information as well as session information is taken into account for intent classification. Evaluation is carried out in a small-scale user study and indicates better intent satisfaction in search results. However, more (accurate) features derived from the user's search session, as well as (implicit) relevance feedback are suggested to enhance classification accuracy. Park et al. [2015] carry out a qualitative user survey that aims to link search behavior of users to search intent categories established by Lux et al. [2010a]. They found the search intent classes are, to some extent, different from each other with respect to users' interaction and search behavior.

In video search, Moshfeghi and Jose [2013] propose an approach to automatically identify search task intents of user search processes. Their approach takes four distinct intent types into account: *"information seeking," "information finding," "emotion need by adjusting arousal level,"* and *"emotion need by adjusting mood."* The authors exploit features derived from interaction and click-through data for their prediction task. Kofler et al. [2014] exploit the distribution of intent types associated with videos in results lists in order to automatically classify queries into classes representing whether they satisfy one particular intent or multiple intents. Each video contained in the results list produced by the query is represented by a vector that indicates how well the video fits particular intent classes (derived during intent-aware indexing). Then, to determine to what extent each search intent class is reflected in the search engine's overall response for a list, the vectors of each video in the list are merged and used for supervised learning.

5.2.2. Connection to Conventional Text-Based Web Search. Many intent-aware query classification approaches proposed for conventional text-based Web search rely on features derived from a user's search session, such as click-through data, dwell time, and search context (i.e., formulation and reformulation behavior) as well as properties of search results returned. Lee et al. [2005] propose a supervised classification approach that automatically assigns queries to a subset of classes introduced in Broder [2002]. The classification approach exploits features derived from click-through data and websites' anchor-link distribution. Guo et al. [2009] carry out a similar classification task using click-through information as features for intent prediction. Yuan et al. [2008] derive features from click-through and website anchor data that express entropy-related information, and Kang and Kim [2003] additionally experiment with search session features such as dwell time. Li et al. [2010] perform semi-supervised learning using search click graphs for automatic query classification. Cheng et al. [2010] predict the search intent of queries based on the browsing behavior of users that was performed prior to query submission. Yoon et al. [2009] derive possible intents from queries by exploiting an external question-answering corpus.

Another feature source for intent-aware query classification is represented by the actual documents returned in search results lists as answers to queries. In particular, the structure of these documents (link distribution, etc.) can be indicative of a query's intent [Herrera et al. 2010]. Baeza-Yates et al. [2006] represent queries by a vector of terms that appeared in the documents giving an answer to a particular query. These vectors—represented by standard tf-idf weighting—are used for both supervised and

unsupervised learning in order to automatically classify queries into informational, not informational, and ambiguous search intent classes. Wu et al. [2010] employ linguistic features extracted from queries and the documents in their produced search results lists. Chang et al. [2006] infer user goals by proposing an approach that exploits snippets related to Web search results. Hu et al. [2009] make use of external data sources that go beyond returned documents and exploit Wikipedia in order to discover large quantities of intent-related features.

Some approaches exploit features derived directly from the query string for classification. Jansen et al. [2008] classify queries using intent classes proposed in Broder [2002] based on query length and the types of terms queries contain. Strohmaier and Kröll [2012] investigate a binary classification task that classifies queries in light of whether they contain an explicit goal statement. For this task, they build bags of words from the plain query string as well as from part-of-speech trigrams present in the query. While features derived directly from the query string perform somewhat satisfactorily, their performance is not comparable with the performance achieved by features that are derived from contextual data sources discussed earlier. Search intent is in many cases not properly expressed in the query [Baeza-Yates et al. 2006], [Strohmaier and Kröll 2012]. For this reason, the query string offers limited opportunities for accurate search intent prediction compared to the rich context provided by other data sources mentioned earlier.

Several approaches have been proposed for intent-aware query suggestion and modification. Guo et al. [2011] argue that query similarity prediction (which is used to suggest similar queries) should be approached from an intent-aware perspective in order to obtain a better similarity measure among queries. They employ topic modeling based on features derived from query co-clicks and search results snippets with the objective of automatically extracting search intent categories of a query. To determine how similar queries are to each other, each query is represented under the aspect of its extracted intents, and corresponding metrics are then chosen. Cao et al. [2008] propose an approach that employs context-aware features from click-through and session data. Kharitonov et al. [2013] provide query suggestions that cover several user intents that might be relevant to the user given an initial query. To this end, they exploit search behavior from the user's *short-term*, immediate search session context (i.e., features derived from the query suggestions that the system automatically provided but which the user rejected) and the previously submitted query, as well as the documents previously examined by the user. Liu et al. [2011] exploit search results snippets for intent-aware query suggestion by following the assumption that although the documents that users click are not always relevant to their information needs, the search results snippets that motivate them to click are likely relevant. Strohmaier et al. [2009] propose an "intentional query suggestion" approach that extracts queries from transaction logs that express intent for a specific topic derived from the initially submitted query (e.g., for query `used cars`, the intentional query `buy a car` would be suggested).

5.2.3. Summary and Insights. The findings collected from the intent-aware query processing research discussed in this section can be summarized in a set of insights:

(1) Intent-aware query processing is a crucial step in the search engine pipeline. Compared to the text-based Web search domain, there has only been a limited amount of work in the MIR community on intent-aware query processing. This work has focused on query classification, while query suggestion and modification have remained largely unaddressed.
(2) Intent-aware query processing using only features derived from the actual query string has limited potential. This is observable even in conventional text-based

Web search, where queries are known to carry more evidence for search intent compared to multimedia search queries [Kofler and Lux 2009a].

(3) Alternative features for intent-aware query processing are useful in the case of conventional text retrieval and stand to benefit MIR as well. These include features derived from a user's search session, such as click-through data, dwell time, and search context, as well as features derived from documents returned in search results lists as answers to queries.

We build on these insights in the discussion of future challenges in Section 6.

5.3. Intent-Aware Search Results Ranking

The main purpose of the search results ranking component of a multimedia search engine is, as discussed in Section 2, to derive the relevance between the user-submitted query and the multimedia documents stored in the index. This is done by comparing and matching the query representation with the representation of the stored documents. In the case of intent-aware MIR, this component heavily relies on approaches that address both intent-aware indexing as well as intent-aware query processing.

Two main approaches can be considered in order to incorporate user intent in search results rankings. First, the ranking produced by the multimedia search engine can directly take intent into account when performing the query matching and retrieval step. Second, user intent can be exploited to optimize a results list that has been initially generated by the search engine. That is, initially produced rankings (that are, for example, *on-topic*) can be refined using intent-aware features that are extracted both from the query as well as from the multimedia documents stored in the search engine's index. It is critical for either approach, however, that the topical focus of the query is not neglected. In other words, it would not improve the overall usefulness of the search engine to generate results lists that perfectly satisfy the user's intent but fail to be relevant to the topic of the information need.

In this section, we review the limited amount of work on intent-aware search results ranking that has been introduced in the literature so far. All of these approaches follow the second option mentioned: They optimize an initially produced search results list in a second refinement step in light of user intent. Optimization is carried out by either performing reranking (i.e., increasing the homogeneity of the top ranks of results lists in terms of user intent) or diversification (i.e., increasing the heterogeneity of the top ranks of results lists in terms of user intent). We then go on to cover approaches from the text-based information retrieval literature and to draw a connection to MIR.

5.3.1. Image and Video Search. Lu et al. [2014] exploit click-through information from query logs and combine it with the clicked images' visual information to derive users' search goals in image search. The authors build on the assumption that the visual patterns of clicked images correspond with search goals of users (i.e., visual patterns of images inspected by users with different search goals should be clearly distinguishable from each other, whereas images examined with related search goals are expected to manifest related visual patterns). The approach automatically clusters images into groups of goals using visual information extracted from images combined with session click-through data. The search results produced by ten image search queries are evaluated using a range of five scores, spanning from "not satisfied" to "satisfied." We note that the evaluation method and rating scheme used by the authors effectively serve to evaluate the system in terms of conventional topical relevance rather than "goal satisfaction" or multiple relevance criteria taking both topic and intent into account.

Lagger et al. [2011] carry out an initial user study that investigates the correlation of user intent and genre categories provided by YouTube. Based on their findings, they propose a manual mapping between four user intent types (i.e., *"learn something," "be*

entertained," "get informed," and *"solve a task"*) and genre categories. This manual mapping is then used in a retrieval step, which returns, given a textual query and an intent explicitly specified by the user, only videos that match both the query and the genre categories corresponding to the specified intent.

Kofler et al. [2014] carry out a quantitative analysis that shows that initial results lists returned by video search engines do contain videos that satisfy a user's intent but that videos with the highest potential for satisfaction are often buried within or scattered over the results list. The authors propose an approach to refine an initial list in light of intent. Results lists are reranked in case one dominant intent is detected in initial results and diversified if several intents are detected. For optimization, the search intent classes proposed in Hanjalic et al. [2012] are used. To evaluate whether users' intents are better satisfied by optimized lists compared to initial results lists, intent relevance judgments for videos are collected through crowdsourcing. Note that the evaluation approach used does not judge the search results in terms of overall information need satisfaction. In other words, instead of evaluating results based on multiple relevance criteria (i.e., intent and topic), it solely judges lists in terms of intent. In Section 6, we return to the specific challenge of evaluating intent-aware retrieval approaches.

5.3.2. Connection to Conventional Text-Based Web Search. Investigating optimization approaches in conventional text-based Web search reveals the huge potential gain that can be achieved with intent-aware methods. Santos et al. [2011], mentioned briefly already, propose a supervised intent-aware search result diversification approach that learns when to apply particular retrieval models for specific intent-aware aspects of a query. Umemoto et al. [2012] propose a two-staged approach that first analyzes the eye movements of users while they look through search results in order to derive the users' search intents. In a second step, they rerank the search results lists based on the derived intent information. Ting-Xuan and Wen-Hsiang [2011] introduce an approach that exploits features from click-through data as well as search results snippets to automatically derive search goals which are then used for reranking. Work presented in Tsukuda et al. [2013] follows a similar approach in which intent types are first estimated from the query and then optimization is carried out in terms of search result diversification.

5.3.3. Summary and Insights. We provide a list of the key insights gained by our investigation of intent-aware search results ranking:

(1) Compared to the text-based Web search domain, where many approaches perform intent-aware search results ranking based on predefined user intent typologies, only a limited amount of work has been conducted in the MIR community regarding intent-aware search results ranking.
(2) User intent can be taken into account for search results ranking by following two different approaches: The search engine can either directly exploit intent and topic in one retrieval step, or it can first produce results lists that are topically relevant and then optimize them with respect to user intent. The approaches proposed so far in the literature have focused on the latter option.
(3) From the work proposed in the field of conventional text-based Web search, it becomes clear that reranking and other optimization features derived from the user's search session and click-through data are valuable for intent-aware ranking. Since these features represent rich contextual information that is necessary to better infer the user's intent, they should be adopted in nontextual domains as well.

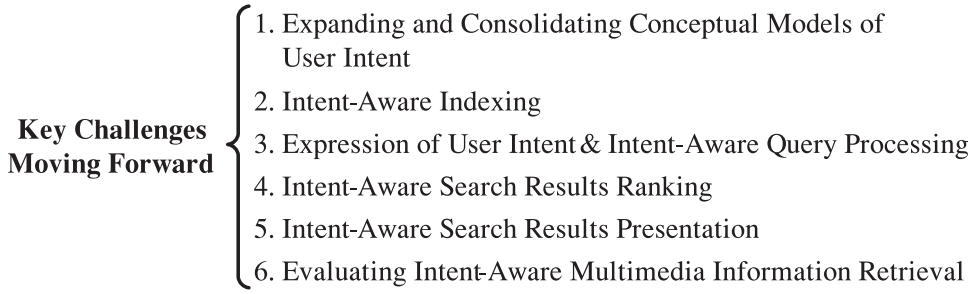

Fig. 6. Overview of key challenges to be faced in order to achieve intent-aware multimedia search engines. Points 2, 3, and 4 relate to the multimedia search pipeline and 1, 5, and 6 go beyond it.

(4) Evaluation of results lists produced by intent-aware ranking is challenging since items must simultaneously be topically relevant, and fulfill user intent. Further work in the area of evaluation procedures is necessary.

We close with the remark that producing search results that satisfy both user intent and topic—that satisfy the full information need of users and therefore multiple relevance criteria—is a nontrivial endeavor for multimedia search engines that, when not carried out appropriately, has the potential to produce suboptimal search results. For this reason, we believe it is essential to provide explanations to users on why and how ranking was performed or why the search engine was not able to provide proper results in response to users' queries. This concept of search result explanation would then further contribute to keeping the usefulness of the search engine high, even though the users' information needs might not be perfectly fulfilled (see Hanjalic [2013]). We return to this remark in the next section, which discusses the key challenges of intent-aware MIR moving forward.

6. KEY CHALLENGES MOVING FORWARD

We now move to identify a series of challenges by projecting forward from the analysis in Section 5. As seen in the overview in Figure 6, these challenges relate to particular points along the multimedia search pipeline (Figure 4), as well as stretch beyond it to other issues. Some challenges are new, and some have already begun to draw research attention. Together, they present a picture of what remains to be accomplished by the MIR research community if it is to achieve truly intent-aware search engines. Here, we address each challenge in turn, highlighting the connection to existing research and the research opportunities that it presents.

6.1. Expanding and Consolidating Conceptual Models of User Intent

Existing user search intent typologies (e.g., Lux et al. [2010a] and Hanjalic et al. [2012]) cover image and video search (cf. Section 4). New research should consolidate these typologies, as well as extend them to other multimedia content types such as speech, audio, and music. Here, we discuss individual steps of this process.

Understanding User Intent. Multimedia search involves multiple domains covering both social and professional content. Each domain requires separate investigation of user information needs, to understand domain-specific aspects of intent, and to define typologies of intent dimensions. Conventional methods such as interviews and surveys are helpful. However, reaching beyond these techniques makes it possible to understand the aspect of user intent that is either spontaneous or not consciously understood by users. Here, exploitation of data from social-Web, question-answering

platforms (e.g., Yahoo! Answers or similar), social media platforms (e.g., *Twitter*), query logs, or collected via crowdsourcing techniques are all useful. However, researchers should be aware that users will most likely not express certain needs (e.g., with respect to their private media collections) online in large public fora.

Creating User Intent Typologies for Other Major Multimedia Content Types. Multimedia search is understudied compared to conventional text search. Here, we point out that *specific types* of multimedia are also relatively neglected. Spoken audio and music are particularly affected, having received very limited research attention. Besser et al. [2010] investigate user goals in podcast search and propose corresponding intent categories by performing an approach based on online surveys, diary studies, and contextual interviews. Demetriou et al. [2016] consider music as *"as an instrument to achieve desired psychological effects within a given context"* and argue that multimedia search engines should be aware of the exact goals that users want to achieve when listening to a specific piece of music. Lee [2010] perform an iterative coding process exploiting real-life user queries to obtain a taxonomy of user needs in music information retrieval. Taheri-Panah and MacFarlane [2004] examine music-seeking behavior of users using questionnaires and interviews. Skov and Lykke [2012] study user needs in sound retrieval by means of user studies and surveys. Although sparse, this initial work provides a basis for expanding intent typologies beyond image and video search. Important tools for creating typologies will be card sorting [Rugg and McGeorge 1997] or similar coding techniques that allow typologies to be iteratively built from large amounts of information collected from users.

Understanding Fine-Grained User Intent in Specific Application Areas. Thus far, we have discussed relatively high-level intent typologies. However, users also consult multimedia search engines for highly specialized reasons. In practice, there is likely a boundless set of intent types that a user could wish to satisfy with multimedia content. We believe that finer-grained, "intra-domain" user intent typologies will be valuable in specific application areas including, but not limited to, education, journalism, engineering, and medicine.

6.2. Intent-Aware Indexing

Intent-aware indexing strives to represent multimedia content with intent-sensitive features or with explicit search intent categories (Section 5.1). Two distinct subchallenges can be identified.

Exploiting User Actions. An important step toward improving intent-aware indexing is to develop content representations that fully exploit user actions. User actions that could prove useful to derive more information for intent-aware indexing are, for example, why users *edit*, *share*, *promote*, *comment on*, or *like* multimedia documents. In particular, comments [Vliegendhart et al. 2013; Eickhoff et al. 2013] and tweets [Chen et al. 2013; Picault et al. 2013] that are related to and about particular multimedia documents might contribute to the understanding of why users engage with these documents. New audio-, image-, and video-capture devices provide new sources of context information (e.g., geographic, social, device information, etc.), which carry clues related to the reasons that users create and consume content. Specifically, we need to understand which user actions give rise to information that reflects search intent and then investigate reliable ways of using information about these actions to represent multimedia documents.

Improving Content-Based Features. The evolution of our understanding of user intent impacts the use of features for representing and classifying content. Attention should be devoted to developing new types of features, including low-level- (i.e., local and global

features) and semantic mid-level visual feature descriptors extracted from images or keyframes of videos, video shot-based features, audio- and music-related features, and features derived from automatic speech recognition transcripts.

Returning to consider papers that we have previously touched on, we mention the importance of low-level and semantic mid-level visual feature representations [Kofler et al. 2015], speech/text features [Hanjalic et al. 2012], and shot-based features that capture temporal patterns in shot length characteristics of videos [Hanjalic et al. 2012]. Mid-level semantic visual representations such as visual concepts [Snoek and Worring 2009] are, even when imperfect, promising features for intent-aware indexing. We base this projection on the success of similar approaches developed for automatic event [Jiang et al. 2013], theme [Rudinac et al. 2012], or emotion [Ellis et al. 2014] derivation from multimedia documents.

We can anticipate that, for a given intent class in a new intent typology, there will be a great deal of audio and visual variability. In some cases, it may actually be impossible to develop features that capture commonalities between content of a certain intent class. However, it is important not to take the failure of standard features as proof that no commonalities exist, but to attempt to advance state-of-the-art content-based features.

6.3. Expression of User Intent and Intent-Aware Query Processing

Future users, we anticipate, will face the same struggle as today's users when it comes to expressing intent in their queries. As observed in Section 5.2, users often fail to express their intent clearly, if they attempt to do so at all. For this reason, approaches need to address the challenge of inferring intent from the limited amount of information available in the user's queries. In light of this challenge, we discuss the following open issues that need to be attended to in future work.

Exploiting Context Information. Context provides important clues to user search intent. Users' goals can be expected to change radically, but regularly, depending on their locations or time of day. We mentioned the importance of context for intent-aware indexing, and now we return to mention the critical role that it can play in inferring user intent from a query. Church and Smyth [2009], for example, study the types of intent behind mobile information needs in terms of social interactions, activity, location, and time. A wide range of contextual data sources can be considered, and their importance for multimedia search cannot be overemphasized [Jain and Sinha 2010].

Exploiting Session-Based Information. A number of studies have been carried out that point to information gathered from user search behavior within a particular search session as highly predictive of intent. Park et al. [2015], for example, investigate user search behavior for the four image search intent classes proposed in Lux et al. [2010a] and found that they are to some extent different from each other with respect to users' interaction and search behavior. Strohmaier et al. [2007] investigate search sessions from such a log and observe that query reformulation behavior (i.e., refining, generalizing, and specializing) is a useful indicator. Downey et al. [2008] conduct a similar study by investigating URL visits extracted from a transaction log. They characterize types (and type frequency) of user intent by the last search result users clicked in a given search session. Their findings indicate that, for infrequent queries, search results clicks are less likely and that query reformulations are more indicative for particular intents. In general, the paradigm of interactive multimedia search [Thomee and Lew 2012] is worth following to exploit session-based information for understanding user search intent.

Performing Query Classification and Enrichment. The goal of intent inference is classification or enrichment of user queries to make them more effective. Additional work

is needed to transform queries so that they better reflect user intent without decreasing the topical focus (cf. Guo et al. [2011]). We anticipate that multimedia queries (e.g., query-by-example-based systems like [Zha et al. 2010]) stand to particularly benefit from classification approaches similar to those we discussed for intent-aware indexing. In particular, we believe that the classifiers applied in the indexing step in order to describe multimedia documents from an intent-aware perspective could be applied to visual queries in order to derive knowledge about the intent of the user who submitted them.

6.4. Intent-Aware Search Results Ranking

The importance of intent-aware ranking techniques is recognized by the research community [Chang 2011] but has attracted less attention than it deserves. The following subtopics provide particularly important opportunities.

Understanding the Interdependencies of User Intent. User intent in search scenarios (i.e., "Why am I searching?") is clearly connected to user intent in other areas related to the production, creation, and sharing of multimedia content (e.g., "Why am I uploading?"). More studies in this area are needed to understand and exploit this connection, including transaction log analysis and crowdsourcing studies. A preliminary example is the static crowdsourcing task that explores the connection between uploader intent and search intent in video search reported by Kofler et al. [2015]. Researchers should consider the possibility that, in a given use scenario, ranking with respect to *uploader* intent is highly transparent and valuable to users, making the specific consideration of *search* intent unnecessary.

Optimizing Search Results Ranking Taking Intent into Account. Ranking for intent-aware multimedia search involves trading off between the relevance of the search results to the topical and the intent component of the user's need. It is important to keep in mind that a results list that is highly relevant to intent will be useless to a user unless it is also relevant to topic. Both *early fusion* and *late fusion* approaches offer possibilities for combining topic and intent. Early fusion approaches combine topic and intent in one step. They calculate the relevance between the user's query and the multimedia documents in the index based on the combination of these two criteria and generate a single, final results list. Late fusion approaches first obtain search results lists satisfying the topical component of the query and then optimize these lists in terms of intent.

Regardless of which of these approaches is followed, different ranking functions are possible that consider intent-aware information. An obvious approach is intent class-dependent retrieval models, which exploit inferred intent from the query and rank the results lists based on the intent distribution derived at the indexing step from multimedia documents. However, it is important not to assume that matching inferred intent classes between query and document is the best possible approach. An alternative approach is to rely on clicks or relevance feedback to iteratively find matches between query and results (e.g., Hua et al. [2013] and Yang et al. [2012] might be applicable in intent-aware ranking scenarios as well). Furthermore, instead of performing intent-class-specific ranking using suboptimal evidence of intent, diversification approaches could be applied that diversify search results in light of the most promising matching intent types. Here, we mention approaches that combine different vertical results (e.g., from text-, image-, shopping-verticals, etc.) in one search results list to provide relevant context to user [Li et al. 2008].

Investigating Intent-Aware Multimedia Recommendation. Moving toward incorporating past click patterns for ratings moves multimedia retrieval into a domain that is

more conventionally covered by the area of recommender systems. Recommendation is sometimes characterized as *query-less search*. Independently of how widely this definition is accepted, recommender system techniques are clearly helpful for multimedia retrieval. The use of rating information and the exploitation of co-occurrence of items in user profiles (i.e., collaborative filtering) is notable.

6.5. Intent-Aware Search Results Presentation

Satisfying the intent behind users' multimedia information needs goes beyond producing an intent-aware results list. Rather, as touched upon in Section 5.3, a multimedia search engine must also convince the user that the results are actually useful in achieving her goals. Especially in the case of time-continuous multimedia (video, speech, music), the representation of the results in the interface is critical. Time-continuous content, in contrast to images, cannot be absorbed at a glance. Instead, users need to make a decision on whether they should invest the effort necessary to "listen in" to a video.

Signaling the Imperfections in Search Results. Instead of only attempting to create perfect search results lists, effort should also be invested in developing algorithms that are capable of signaling potential shortcomings. Imperfections in search results may be due to different reasons (e.g., incorrectly applied search algorithms, ineffective query formulations, or missing content [Hanjalic 2013]). User studies are needed to understand which kind of information on intent-related shortcomings in search results can be most readily understood and exploited by users. Such studies would also reveal the appropriate balance between information reflecting *that* results might be suboptimal and information reflecting *why* they are suboptimal. We point to the literature on confidence score generation and on query performance prediction [Cronen-Townsend et al. 2002] in the area of information retrieval. These are possible starting points for investigating how to create multimedia search engines that are "self-divulging" in that they provide users with information about where they fall short.

Generating Representations. An obvious way of explaining the intent-related relevance of multimedia search results to users is to tag results with intent classes or otherwise generate short textual descriptions that characterize the connection of multimedia content to intent. However, representations that allow users to understand and exploit intent relevance in results lists can also be more subtle. This can be achieved, for example, by adjusting the way search results are served to users (e.g., Kofler and Lux [2009b]) or by automatically generating textual explanations based on the overall quantified confidence score produced by the system.

Zeroing in on Relevant Time-Points in Multimedia. Multimedia information retrieval faces special challenges for time-continuous media, which users cannot absorb in one glance. As mentioned earlier, video, spoken audio, and music need special explanations of relevance in order to convince users that they are worth listening to further. Additionally, users benefit greatly if they do not need to listen to an entire multimedia file, but rather can jump directly to the point in the stream that is most useful for them, given their search intent.

To develop intent-aware multimedia retrieval systems that are able to return individual jump-in points, fine-grained indexing approaches are necessary. Specifically, we point to the existence of user comments reflecting and describing the content of specific parts of long multimedia documents (e.g., YouTube comments can be linked to specific parts of a video by including time references in the video's comments). This kind of data can be used directly in the index or can be exploited to develop features capable of representing intent-related information at the time-code level, parallel to file-level indexing discussed in Section 6.2. Work already moving in this direction includes Vliegendhart

et al. [2013] (investigating why and how timed-comments on YouTube are produced by users) and Xu and Larson [2014] (investigating "timed tags"; i.e., user tags assigned to particular points in time of a video).

6.6. Evaluating Intent-Aware Multimedia Information Retrieval

Evaluation is critical because it allows us to determine whether new techniques in multimedia search are truly satisfying user needs with respect to multimedia search intent. Research must be dedicated specifically to the development of evaluation techniques. In this section, we point out where the need is greatest.

Assessing Satisfaction of Users with Respect to Intent. The fact that user intent is specific to a user in a particular use situation, and may not be well reflected in user queries, makes intent-aware evaluation of results lists a rather delicate and difficult task. A further challenge is raised by the fact that, in order to evaluate search results in terms of the full information need satisfaction of users, it is necessary to judge these results equally in terms of both topic *and* intent. In other words, it is necessary to measure the usefulness of search results by more than a single relevance criterion. For this reason, it is crucial that, next to user-formulated queries, the actual information needs behind these queries are taken into account. In this context, we point again to the promising approach of Kofler et al. [2014], which does not collect queries, but rather directly collects user expressions of search goals in the form of questions posted to the Yahoo! Answers question-answering platform. These information needs were the starting point to derive textual queries and annotations that were used for evaluation. We suggest qualitative studies that assess the usefulness of results in response to individual queries be conducted.

Developing Intent-Aware Evaluation Metrics for Multimedia Search. Measures of the usefulness of intent-aware search engines to users must reflect multiple relevance criteria (i.e., both the topic and the intent dimension). Here, we point to work in the conventional text-based information retrieval community, where several evaluation metrics have been proposed that are adapted for user intent and assess results lists by combined relevance criteria (e.g., Sakai [2012] and Agrawal et al. [2009]). Sakai [2012], for example, proposes an evaluation metric that takes the nature of informational queries (several produced search results might be useful to the user) and navigational queries (typically only one search result is useful and the one the user is searching for) into account when designing an adapted, intent-aware version of Normalized DCG. We suggest that research effort should be invested in the design of intent-aware evaluation metrics specifically tailored to the characteristics and use contexts of MIR as well.

Moving from Offline to Online Evaluation. MIR adopts its notice of "static" (i.e., "offline") evaluation from the text-based IR community, whose highly developed frameworks and methodologies for evaluation are traced back to the *Cranfield Paradigm*. That paradigm uses a static dataset, a set of queries, and a set of annotations that reflect the relevance of the document in the dataset to the queries. Approaches following this paradigm have been used in many multimedia benchmarking initiatives such as *TRECVID*[4] and *MediaEval*.[5] Intent-aware evaluations are needed, meaning that the query set and the annotations must reflect relevance to both the topical and the intent dimensions of user information needs. An increasingly feasible alternative to static evaluation is "dynamic" (i.e., online) evaluation. "Dynamic" environments refer to evaluation environments that change over time and typically offer a larger user

[4]http://trecvid.nist.gov.
[5]http://www.multimediaeval.org.

base, as in *A/B testing* [Siroker and Koomen 2013]. Here, the algorithm to be tested is implemented as a real-world search engine and is tested by observing user interactions (most approaches concentrate on clicks). The challenge of online evaluation is reproducibility. Offline evaluation remains the only way to ensure that two algorithms are tested under identical conditions. A/B testing is useful in practice, when highly comparable conditions can be considered "close enough." Future work will include methods that use offline datasets to predict online performance. Such approaches will allow the research community to take full advantage of having users interact with algorithms directly, rather than relying on evaluation that is mediated by a static query set and annotator judgments.

7. CONCLUSION

This survey provided an overview of user intent in the area of multimedia information retrieval. We placed special emphasis on exposing the gaps in the existing research that must be addressed in order to make possible truly intent-aware multimedia search engines. Here, we provide a brief summary of the ground covered. We discussed different interpretations of user intent in the MIR research community and adopted a definition of user intent that has been successfully deployed in the field of conventional text-based Web search. According to this definition, user intent is the "immediate reason, purpose, or goal behind a user information need" and the driving motivation why users consult multimedia search engines (see Hanjalic et al. [2012]). We investigated several conceptual models of user intent, discussed approaches that have been proposed throughout the entire search engine pipeline (i.e., indexing, query processing, search results ranking), built on these models, and compared them with findings from research effort about user intent in the domain of conventional text-based Web search. On the foundation of this research, we highlighted and assessed important future challenges for developing intent-aware multimedia search engines. These challenges will likely draw in considerable research efforts and pave the way for successful research outcomes in the future.

We believe that additional challenges will emerge in the future that go above and beyond the challenges discussed in this survey. These challenges will be affected by many factors, including the availability of additional information and information sources, the development of new applications involving intent-aware algorithms (e.g., social networks, educational platforms), the growing emphasis on personalization, and the exploration of new crossovers between different types of user intent.

We close by mentioning three particular directions that arise from these considerations. First, in the future, search intent will likely span different modalities. We are already observing that when users are offered search engine verticals corresponding to particular modalities (text, maps, images, video), they satisfy their needs by mixing results. Commercial search engines are moving toward supporting such mixing. For this reason, search intent categories, currently tuned toward text or visual modality separately, will likely also need to be developed that cover simultaneously the full spectrum of modalities. Second, we find that user context will be an increasingly important source of clues concerning user intent, especially as the number of searchers using mobile devices continues to rise. Context includes user information such as time, place, activity, and environmental conditions. Third, we believe that understanding user intent is helpful for multimedia applications that go beyond search. We have already mentioned recommender systems, and we add that general media-sharing and social media platforms could be advanced by taking intent into account. Specific examples of where they could benefit are spam detection, multimedia document adaptation (e.g., applying *Instagram*-like filters given a document's inferred intent), or better selection of multimedia content, for example, for use in education (e.g., by teachers) or entertainment (e.g., by DJs)

In view of the significance of intent in multimedia search, the many challenges addressed in this article and the dynamic nature of these challenges as described herein, we believe that intent-aware MIR will remain a successful and fruitful research field with significant opportunities to achieve high impact in research, as well as in the development of major multimedia search engines and retrieval platforms.

REFERENCES

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. ACM, 5–14.

Morgan Ames and Mor Naaman. 2007. Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computer Systems (CHI'07)*. ACM, 971–980.

Paul André, Edward Cutrell, Desney S. Tan, and Greg Smith. 2009. Designing novel image search interfaces by understanding unique characteristics and usage. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II (INTERACT'09)*. Springer-Verlag, 340–353.

Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. 2006. The intention behind web queries. In *Proceedings of the 13th International Conference on String Processing and Information Retrieval (SPIRE'06)*. Springer-Verlag, 98–109.

Jana Besser, Martha Larson, and Katja Hofmann. 2010. Podcast search: User goals and retrieval technologies. *Online Information Review* 34, 3 (2010), 395–419.

Jingwen Bian, Zheng-Jun Zha, Hanwang Zhang, Qi Tian, and Tat-Seng Chua. 2012. Visual query attributes suggestion. In *Proceedings of the 20th ACM International Conference on Multimedia (MM'12)*. ACM, 869–872.

Nis Bornoe and Louise Barkhuus. 2010. Video microblogging: Your 12 seconds of fame. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems (CHI EA'10)*. ACM, 3325–3330.

Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10.

Stefan Buettcher, Charles L. A. Clarke, and Gordon V. Cormack. 2010. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press.

John Burgess and Joshua Green. 2009. *YouTube: Online Video and Participatory Culture*. Polity Press.

Marco Campanella and Jettie Hoonhout. 2008. Understanding behaviors and needs for home videos. In *Proceedings of the 22nd British HCI Group Conference on People and Computers: Culture, Creativity, Interaction - Volume 2 (BCS-HCI'08)*. British Computer Society, 23–26.

Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, 875–883.

Shih-Fu Chang. 2011. Content based multimedia retrieval: Lessons learned from two decades of research. In *Proceedings of the 19th ACM International Conference on Multimedia (MM'11)*. ACM, 1–2.

Yao-Sheng Chang, Kuan-Yu He, Scott Yu, and Wen-Hsiang Lu. 2006. Identifying user goals from web search results. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06)*. IEEE Computer Society, 1038–1041.

Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013. Understanding and classifying image tweets. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*. ACM, 781–784.

Xu Cheng, Cameron Dale, and Jiangchuan Liu. 2007. Understanding the characteristics of internet short video sharing: Youtube as a case study. *CoRR* abs/0707.3670 (2007).

Zhicong Cheng, Bin Gao, and Tie-Yan Liu. 2010. Actively predicting diverse search intent from user browsing behaviors. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, 221–230.

Youngok Choi and Edie M. Rasmussen. 2003. Searching for images: The analysis of users' queries for image retrieval in American history. *Journal of the American Society for Information Science and Technology* 54, 6 (April 2003), 498–511.

EunKyung Chung and JungWon Yoon. 2012. Analysis of multimedia needs and searching features: An exploratory study. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–5.

Karen Church and Barry Smyth. 2009. Understanding the intent behind mobile information needs. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI'09)*. ACM, 247–256.

Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st International ACM Conference on Information Retrieval (SIGIR'08)*. ACM, 659–666.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th Internatioal ACM Conference on Information Retrieval (SIGIR'02)*. ACM, 299–306.

Jingyu Cui, Fang Wen, and Xiaoou Tang. 2008. IntentSearch: Interactive on-line image search re-ranking. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*. ACM, 997–998.

Peng Cui, Shao-Wei Liu, Wen-Wu Zhu, Huan-Bo Luan, Tat-Seng Chua, and Shi-Qiang Yang. 2014. Social-sensed image search. *ACM Transactions on Information Systems* 32, 2, Article 8 (April 2014), 23 pages.

Sally Jo Cunningham and David M. Nichols. 2008. How people find videos. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'08)*. ACM, 201–210.

Na Dai, Xiaoguang Qi, and Brian D. Davison. 2011. Bridging link and query intent to enhance web search. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia (HT'11)*. ACM, 17–26.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computer Surveys* 40, 2, Article 5 (May 2008), 60 pages.

Andrew Demetriou, Martha Larson, and Cynthia Liem. 2016. On cultural, textual and experiential aspects of music mood. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016*. to appear.

Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. 2008. Understanding the relationship between searchers' queries and information goals. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, 449–458.

Carsten Eickhoff, Wen Li, and Arjen P. de Vries. 2013. Exploiting user comments for audio-visual content indexing and retrieval. In *Proceedings of the 35th European Conference on Advances in Informationa Retrieval (ECIR'13)*. Springer-Verlag, 38–49.

Joseph Ellis, Weisi Lin, Ching-Yung Lin, and Shih-Fu Chang. 2014. Predicting evoked emotions in video. In *Proceedings of the 2014 IEEE International Symposium on Multimedia (ISM).* 287–294.

Bailan Feng, Juan Cao, Zhineng Chen, Yongdong Zhang, and Shouxun Lin. 2010. Multi-modal query expansion for web video search. In *Proceedings of the 33rd International ACM Conference on Information Retrieval (SIGIR'10)*. ACM, 721–722.

Raya Fidel. 1997. The image retrieval task: Implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia* 3 (1997).

Giovanni Gardelli and Ingmar Weber. 2012. Why do you ask this? In *Proceedings of the 21st International Conference Companion on World Wide Web (WWW'12 Companion)*. ACM, 815–822.

Jian Guan and Guoping Qiu. 2007. Learning user intention in relevance feedback using optimization. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval (MIR'07)*. ACM, 41–50.

Fan Guo, Lei Li, and Christos Faloutsos. 2009. Tailoring click models to user goals. In *Proceedings of the 2009 Workshop on Web Search Click Data (WSCD'09)*. ACM, 88–92.

Jiafeng Guo, Xueqi Cheng, Gu Xu, and Xiaofei Zhu. 2011. Intent-aware query similarity. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*. ACM, 259–268.

Martin J. Halvey and Mark T. Keane. 2007. Analysis of online video search and sharing. In *Proceedings of the 18th Conference on Hypertext and Hypermedia (HT'07)*. ACM, 217–226.

Alan Hanjalic. 2013. Multimedia retrieval that matters. *ACM Transactions on Multimedia Computing, Communications, and Applications* 9, 1s, Article 44 (Oct. 2013), 5 pages.

Alan Hanjalic, Christoph Kofler, and Martha Larson. 2012. Intent and its discontents: The user at the wheel of the online video search engine. In *Proceedings of the 20th ACM International Conference on Multimedia (MM'12)*. ACM, 1239–1248.

Mauro Rojas Herrera, Edleno Silva de Moura, Marco Cristo, Thomaz Philippe Silva, and Altigran Soares da Silva. 2010. Exploring features for the automatic identification of user goals in web search. *Information Processing & Management* 46, 2 (2010), 131–142.

Enamul Hoque, Orland Hoeber, and Minglun Gong. 2013. CIDER: Concept-based image diversification, exploration, and retrieval. *Information Processing & Management* 49, 5 (Sept. 2013), 1122–1138.

Jian Hu, Gang Wang, Fred Lochovsky, Jian-tao Sun, and Zheng Chen. 2009. Understanding user's query intent with wikipedia. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, 471–480.

Xian-Sheng Hua, Linjun Yang, Jingdong Wang, Jing Wang, Ming Ye, Kuansan Wang, Yong Rui, and Jin Li. 2013. Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*. ACM, 243–252.

Ramesh Jain and Pinaki Sinha. 2010. Content without context is meaningless. In *Proceedings of the International Conference on Multimedia (MM'10)*. ACM, 1259–1268.

Vidit Jain and Manik Varma. 2011. Learning to re-rank: Query-dependent image re-ranking using click data. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. ACM, 277–286.

Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management* 44, 3 (May 2008), 1251–1266.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD'07)*. ACM, 56–65.

Beth St. Jean, Soo Young Rieh, Yong-Mi Kim, and Ji Yeon Yang. 2012. An analysis of the information behaviors, goals, and intentions of frequent internet users: Findings from online activity diaries. *First Monday* 17, 2 (2012).

Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. 2013. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval* 2, 2 (2013), 73–101.

Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, 699–708.

In-Ho Kang and GilChang Kim. 2003. Query type classification for web document retrieval. In *Proceedings of the 26th International ACM Conference on Research and Development in Information Retrieval (SIGIR'03)*. ACM, 64–71.

Lyndon S. Kennedy, Apostol (Paul) Natsev, and Shih-Fu Chang. 2005. Automatic discovery of query-class-dependent models for multimodal search. In *Proceedings of the 13th ACM International Conference on Multimedia (MULTIMEDIA'05)*. ACM, 882–891.

Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2013. Intent models for contextualising and diversifying query suggestions. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM'13)*. ACM, 2303–2308.

Tim Kindberg, Mirjana Spasojevic, Rowanne Fleck, and Abigail Sellen. 2005. The ubiquitous camera: An in-depth study of camera phone use. *IEEE Pervasive Computing* 4, 2 (April 2005), 42–50.

Christoph Kofler, Subhabrata Bhattacharya, Martha Larson, Tao Chen, Alan Hanjalic, and Shih-Fu Chang. 2015. Uploader intent for online video: Typology, inference and applications. *IEEE Transactions on Multimedia (To Appear)* (2015).

Christoph Kofler, Martha Larson, and Alan Hanjalic. 2014. Intent-aware video search result optimization. *IEEE Transactions on Multimedia* 16, 5 (Aug 2014), 1421–1433.

Christoph Kofler and Mathias Lux. 2009a. An exploratory study on the explicitness of user intentions in digital photo retrieval. In *Proceedings of the International Conference on Knowledge Technologies and Data-driven Business (I-KNOW'09)*. 7.

Christoph Kofler and Mathias Lux. 2009b. Dynamic presentation adaptation based on user intent classification. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*. ACM, 1117–1118.

Marian Kogler and Mathias Lux. 2011. Pursuing the holy grail by interrelating user intentions and bag of visual words to perform retrieval adaptation. In *Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access (SBNMA'11)*. ACM, 3–8.

Christoph Lagger, Mathias Lux, and Oge Marques. 2011. Which video do you want to watch now? Development of a prototypical intention-based interface for video retrieval. In *Proceedings of the 2011 Workshop on Multimedia on the Web (MMWeb)*. 45–48.

Christoph Lagger, Mathias Lux, and Oge Marques. 2012. What makes people watch online videos: An exploratory study. *Computer Entertainment* (2012).

Jin Ha Lee. 2010. Analysis of user needs and information features in natural language queries seeking music information. *Journal of the American Society for Information Science and Technology* 61, 5 (2010), 1025–1045.

Uichin Lee, Zhenyu Liu, and Junghoo Cho. 2005. Automatic identification of user goals in web search. In *Proceedings of the 14th International Conference on World Wide Web (WWW'05)*. ACM, 391–400.

Lusong Li and Jing Li. 2011. MQSS: Multimodal query suggestion and searching for video search. *Multimedia Tools and Applications* 54, 1 (2011), 55–68.

Xiao Li, Ye-Yi Wang, Dou Shen, and Alex Acero. 2010. Learning with click graph for query intent classification. *ACM Transactions on Information Systems* 28, 3, Article 12 (July 2010), 20 pages.

Zhiwei Li, Shuming Shi, and Lei Zhang. 2008. Improving relevance judgment of web search results with image excerpts. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. ACM, 21–30.

Shaowei Liu, Peng Cui, Huanbo Luan, Wenwu Zhu, Shiqiang Yang, and Qi Tian. 2014. Social-oriented visual image search. *Computer Vision and Image Understanding* 118, 0 (2014), 30–39.

Yiqun Liu, Junwei Miao, Min Zhang, Shaoping Ma, and Liyun Ru. 2011. How do users describe their information need: Query recommendation based on snippet click model. *Expert Systems with Applications* 38, 11 (2011), 13847–13856.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (Nov. 2004), 91–110.

Zheng Lu, Xiaokang Yang, Weiyao Lin, Hongyuan Zha, and Xiaolin Chen. 2014. Inferring user image-search goals under the implicit guidance of users. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 3 (March 2014), 394–406.

Mathias Lux and Jochen Huber. 2012. Why did you record this video? An exploratory study on user intentions for video production. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th Int'l Workshop on*. 1–4.

Mathias Lux, Christoph Kofler, and Oge Marques. 2010a. A classification scheme for user intentions in image search. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems (CHI EA'10)*. ACM, 3913–3918.

Mathias Lux, Marian Kogler, and Manfred del Fabro. 2010b. Why did you take this photo: A study on user intentions in digital photo productions. In *Proceedings of the 2010 ACM Workshop on Social, Adaptive and Personalized Multimedia Interaction and Access (SAPMIA'10)*. ACM, 41–44.

Mathias Lux, Mario Taschwer, and Oge Marques. 2012. A closer look at photographers' intentions: A test dataset. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia (CrowdMM'12)*. ACM, 17–18.

Amy Madden, Ian Ruthven, and David McMenemy. 2013. A classification scheme for content analyses of YouTube video comments. *Journal of Documentation* 69, 5 (2013), 693–714.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. 2006. HT06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the 17th Conference on Hypertext and Hypermedia (HYPERTEXT'06)*. ACM, 31–40.

Tao Mei and Xian-Sheng Hua. 2005. Intention-based home video browsing. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA'05)*. ACM, New York, 221–222. DOI:http://dx.doi.org/10.1145/1101149.1101186

Tao Mei, Yong Rui, Shipeng Li, and Qi Tian. 2014. Multimedia search reranking: A literature survey. *ACM Computer Surveys* 46, 3, Article 38 (Jan. 2014), 38 pages.

Julie B. Morrison, Peter Pirolli, and Stuart K. Card. 2001. A taxonomic analysis of what world wide web activities significantly impact people's decisions and actions. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems (CHI EA'01)*. ACM, 163–164.

Yashar Moshfeghi and Joemon M. Jose. 2013. On cognition, emotion, and interaction aspects of search tasks with different search intentions. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 931–942. http://dl.acm.org/citation.cfm?id=2488388.2488469

Bonnie A. Nardi, Diane J. Schiano, Michelle Gumbrecht, and Luke Swartz. 2004. Why we blog. *Communications of the ACM* 47, 12 (Dec. 2004), 41–46.

Paul Over. 1996. TREC-5 interactive report. In *Proceedings of the 5th Text Retrieval Conference (TREC-5)*. NIST Special Publication 500-238.

Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. 2014. Click-through-based cross-view learning for image search. In *Proceedings of the 37th International ACM Conference on Information Retrieval (SIGIR'14)*. ACM, 717–726.

Jaimie Y. Park, Neil O'Hare, Rossano Schifanella, Alejandro Jaimes, and Chin-Wan Chung. 2015. A large-scale study of user image search behavior on the web. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. ACM, New York, 985–994. DOI:http://dx.doi.org/10.1145/2702123.2702527

Namkee Park, Younbo Jung, and Kwan Min Lee. 2011. Intention to upload video content on the internet: The role of social norms and ego-involvement. *Computers in Human Behavior* 27, 5 (2011), 1996–2004.

Jerome Picault, Myriam Ribiere, and Yann Gaste. 2013. Indexing video segments using microblogs. In *Proceedings of the 11th International Workshop on Content-Based Multimedia Indexing (CBMI 2013)*, 155–160.

Michael Riegler, Martha Larson, Mathias Lux, and Christoph Kofler. 2014. How 'how' reflects what's what: Content-based exploitation of how users frame social images. In *Proceedings of the ACM International Conference on Multimedia (MM'14)*. ACM, 397–406.

Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04)*. ACM, 13–19.

Stevan Rudinac, Martha Larson, and Alan Hanjalic. 2012. Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval. *International Journal of Multimedia Information Retrieval* 1, 4 (2012), 263–280.

Gordon Rugg and Peter McGeorge. 1997. The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems* 14, 2 (1997), 80–93.

Tetsuya Sakai. 2012. Evaluation with informational and navigational intents. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, 499–508.

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2011. Intent-aware search result diversification. In *Proceedings of the 34th International ACM Conference on Information Retrieval (SIGIR'11)*. ACM, 595–604.

Dan Siroker and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks into Customers* (1st ed.). Wiley Publishing.

Mette Skov and Marianne Lykke. 2012. Unlocking radio broadcasts: User needs in sound retrieval. In *Proceedings of the 4th Information Interaction in Context Symposium (IIIX'12)*. ACM, 298–301.

Cees G. M. Snoek and Marcel Worring. 2009. Concept-based video retrieval. *Foundations and Trends in Information Retrieval* 2, 4 (April 2009), 215–322.

Amanda Spink and Bernhard J. Jansen. 2005. *Web Search: Public Searching of the Web – Multimedia Searching*. Information Science and Knowledge Management, Vol. 6. Springer Netherlands. 161–177 pages.

Markus Strohmaier. 2008. Purpose tagging: Capturing user intent to assist goal-oriented social search. In *Proceedings of the 2008 ACM Workshop on Search in Social Media (SSM'08)*. ACM, 35–42.

Markus Strohmaier and Mark Kröll. 2012. Acquiring knowledge about human goals from search query logs. *Information Processing & Management* 48, 1 (2012), 63–82.

Markus Strohmaier, Mark Kröll, and Christian Körner. 2009. Intentional query suggestion: Making user goals more explicit during search. In *Proceedings of the 2009 Workshop on Web Search Click Data (WSCD'09)*. ACM, 68–74.

Markus Strohmaier, Mathias Lux, Michael Granitzer, Peter Scheir, Sotirios Liaskos, and Eric Yu. 2007. How do users express goals on the web? In *Web Information Systems Engineering WISE 2007 Workshops*. Lecture Notes in Computer Science, Vol. 4832. Springer Berlin, 67–78.

S. Taheri-Panah and A. MacFarlane. 2004. Music information retrieval systems: Why do individuals use them and what are their needs. In *Proceedings of the International Symposium on Music Information Retrieval*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.7318&rank=2.

Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. 2008. To personalize or not to personalize: Modeling queries with variation in user intent. In *Proceedings of the 31st International ACM Conference on Information Retrieval (SIGIR'08)*. ACM, 163–170.

Bart Thomee and Michael S. Lew. 2012. Interactive search in image retrieval: A survey. *International Jouranl of Multimedia Information Retrieval* 1, 2 (2012), 71–86.

Wang Ting-Xuan and Lu Wen-Hsiang. 2011. Identifying popular search goals behind search queries to improve web search ranking. In *Information Retrieval Technology*. Lecture Notes in Computer Science, Vol. 7097. Springer Berlin Heidelberg, 250–262.

Dian Tjondronegoro, Amanda Spink, and Bernard J. Jansen. 2009. A study and comparison of multimedia web searching: 1997-2006. *Journal of the American Society for Information and Science Technology* 60, 9 (Sept. 2009), 1756–1768.

Michele Trevisiol, Luca Chiarandini, Luca Maria Aiello, and Alejandro Jaimes. 2012. Image ranking based on user browsing behavior. In *Proceedings of the 35th International ACM Conference on Informationl Retriveal (SIGIR'12)*. ACM, 445–454.

Kosetsu Tsukuda, Tetsuya Sakai, Zhicheng Dou, and Katsumi Tanaka. 2013. Estimating intent types for search result diversification. In *Informational Retrieval Technology.* Lecture Notes in Computer Science, Vol. 8281. Springer Berlin Heidelberg, 25–37.

Kazutoshi Umemoto, Takehiro Yamamoto, Satoshi Nakamura, and Katsumi Tanaka. 2012. Search intent estimation from user's eye movements for supporting information seeking. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI'12)*. ACM, 349–356.

Nancy A. Van House. 2007. Flickr and public image-sharing: Distant closeness and photo exhibition. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems (CHI EA'07)*. ACM, 2717–2722.

Reinier H. van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. 2009. Visual diversification of image search results. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, 341–350.

Roelof van Zwol, Vanessa Murdock, Lluis Garcia Pueyo, and Georgina Ramirez. 2008. Diversifying image search with user generated content. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR'08)*. ACM, 67–74.

Raynor Vliegendhart, Babak Loni, Martha Larson, and Alan Hanjalic. 2013. How do we deep-link? Leveraging user-contributed time-links for non-linear video access. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*. ACM, 517–520.

Jingdong Wang and Xian-Sheng Hua. 2011. Interactive image search by color map. *ACM Transactions on Intelligent Systems Technology* 3, 1, Article 12 (Oct. 2011), 23 pages.

Dayong Wu, Yu Zhang, Shiqi Zhao, and Ting Liu. 2010. Identification of web query intent based on query text and web knowledge. In *Proceedings of the 1st International Conference on Pervasive Computing Signal Processing and Applications (PCSPA'10),* 128–131.

Peng Xu and Martha Larson. 2014. Users tagging visual moments: Timed tags in social video. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM'14)*. ACM, 57–62.

Linjun Yang, Bo Geng, Alan Hanjalic, and Xian-Sheng Hua. 2012. A unified context model for web image retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 8, 3, Article 28 (Aug. 2012), 19 pages.

Linjun Yang and Alan Hanjalic. 2010. Supervised reranking for web image search. In *Proceedings of the International Conference on Multimedia (MM'10)*. ACM, 183–192.

Xiaopeng Yang, Yongdong Zhang, Ting Yao, Chong-Wah Ngo, and Tao Mei. 2014. Click-boosting multimodality graph-based reranking for image search. *Multimedia Systems* (2014), 1–11.

Yi-Hsuan Yang and W. H. Hsu. 2008. Video search reranking via online ordinal reranking. In *Proceedings of the 2008 IEEE International Conference and Expo on Multimedia*. 285–288.

Soungwoong Yoon, Adam Jatowt, and Katsumi Tanaka. 2009. Intent-based categorization of search results using questions from web q&a corpus. In *Web Information Systems Engineering - WISE 2009*. Lecture Notes in Computer Science, Vol. 5802. Springer Berlin, 145–158.

Xiaojie Yuan, Zhicheng Dou, Lu Zhang, and Fang Liu. 2008. Automatic user goals identification based on anchor text and click-through data. *Wuhan University Journal of Natural Sciences* 13, 4 (2008), 495–500.

Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, Zengfu Wang, Tat-Seng Chua, and Xian-Sheng Hua. 2010. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Transactions on Multimedia Computing, Communications, and Applications* 6, 3, Article 13 (Aug. 2010), 19 pages.

Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*. ACM, 33–42.

Received July 2015; revised January 2016; accepted May 2016