When Video Search Goes Wrong: Predicting Query Failure Using Search Engine Logs and Visual Search Results *

Christoph Kofler †, Linjun Yang ‡, Martha Larson †, Tao Mei ‡, Alan Hanjalic †, Shipeng Li ‡
† Multimedia Information Retrieval Lab, Delft University of Technology, Delft, The Netherlands
‡ Microsoft Research Asia, Beijing, P. R. China
{c.kofler, m.a.larson, a.hanjalic}@tudelft.nl, {linjuny, tmei, spli}@microsoft.com

ABSTRACT
The recent increase in the volume and variety of video content available online presents growing challenges for video search. Users face increased difficulty in formulating effective queries and search engines must deploy highly effective algorithms to provide relevant results. Although lately much effort has been invested in optimizing video search engine results, relatively little attention has been given to predicting for which queries results optimization is most useful, i.e., predicting which queries will fail. Being able to predict when a video search query would fail is likely to make the video search result optimization more efficient and effective, improve the search experience for the user by providing support in the query formulation process and in this way boost the development of video search engines in general. While insight about a query's performance in general could be obtained using the well-known concept of query performance prediction (QPP), we propose a novel approach for predicting a failure of a video search query in the specific context of a search session. Our *context-aware query failure* prediction approach uses a combination of *user indicators* and *engine indicators* to predict whether a particular query is likely to fail in the context of a particular search session. User indicators are derived from the search log and capture the patterns of query (re)formulation behavior and the click-through data of a user during a typical video search session. Engine indicators are derived from the video search results list and capture the visual variance of search results that would be offered to the user for the given query. We validate our approach experimentally on a test set containing 1+ million video search queries and show its effectiveness compared to a set of conventional QPP baselines. Our approach achieves a 13% relative improvement over the baseline.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process, Information filtering, Query formulation*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing;

General Terms
Algorithms, Experimentation, Human Factors, Performance

Keywords
Video search, query failure, transaction log analysis, query performance prediction, visual relatedness

* This work was performed when Christoph Kofler was visiting Microsoft Research Asia as a research intern.

MM'12, October 29–November 2, 2012, Nara, Japan.

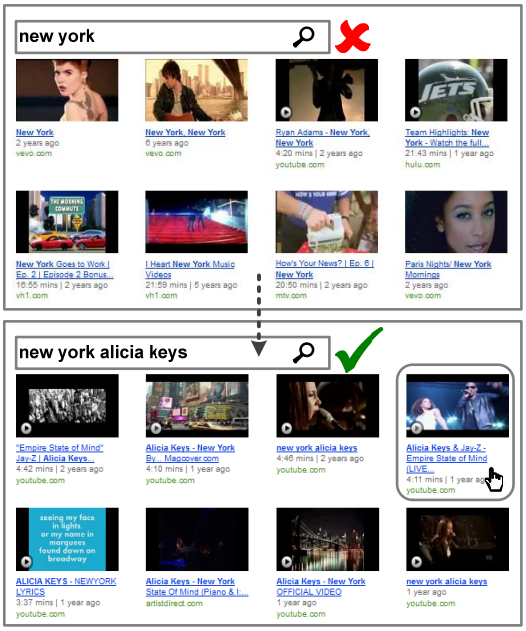

Figure 1: A visualization of an excerpt from an ongoing search session from the transaction log. The query *'new york'* was not successful, gets reformulated to *'new york alicia keys'* leading to a query success.

1. INTRODUCTION
As the volume and diversity of video data on the Internet increases, the interaction between users and search engines becomes increasingly complex. Users, on the one hand, can be expected to have a less complete overview of the full scope of the content available, resulting in increased difficulty in formulating successful video search queries. Video search engines, on the other hand, can be expected to have more difficulty pinpointing content that matches user information needs. Recently, a number of approaches have been developed that improve the performance of video search engines operating on web-scale data. Key examples of these techniques include concept-based video search [26], reranking of video search results [9, 30] and query suggestion [31]. Such approaches could be more effectively deployed if it were possible to predict at which points in the sequence of user interactions with a video search engine (i.e., the video search session) the user is likely to be confronted with non-relevant search results.

In this paper, we introduce an approach for the prediction of *context-aware query failure* in video search. We consider a query, issued by a user in the context of a video search session as failed if it produces a results list containing no search results the user considers relevant enough to click. We adopt methodology from transaction log analysis including assumptions about the relationship between clicks and relevance from the field of text retrieval

[3, 15]. Hereby we solely take the 'effective' part of a results list for relevance into account, i.e., results so low in the list that they are never reached are not regarded [3]. We approach query failure prediction as a multimodal binary classification problem and introduce a technique that combines two information sources: *user indicators* (derived from search engine logs recording user interaction) and *engine indicators* (derived from the results list that the search engine delivers in response to the user query).

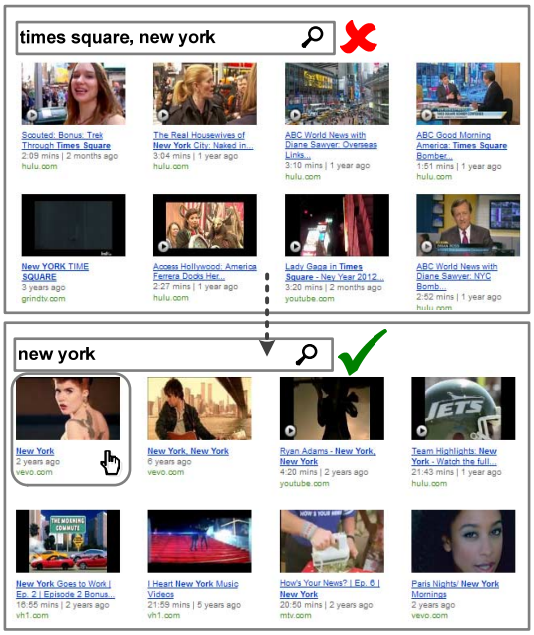

Figure 2: A visualization of an excerpt from an ongoing search session from the transaction log. The query *'times square, new york'* was not successful, gets reformulated to *'new york'* leading to a query success.

The context-aware query failure prediction problem is a challenging one. Figure 1 illustrates this problem with an excerpt from an ongoing search session consisting of a failed query followed by a query expansion that leads to a successful query. The user is trying to find a video of the song 'Empire State of Mind' and the first query depicted is *'new york'*, a more common designation for 'The Empire State'. The resulting query is too broad, returning a results list in which the user does not click. The user then adds the name of the soloist 'Alicia Keys' to the query, which leads to a results list containing a video the user does click. The challenge of query failure prediction lies in the enormous range of variability in user query patterns and engine responses, and the link to success is not straightforward. The contrast between Figure 2 and Figure 1 provides a simple illustration of the unexpectedness of user query patterns. In Figure 2, the first query *'times square, new york'* does not lead to a result that the user finds to be worth a click, but a reduction of the query to *'new york'* does. Possibly the user is looking for information about New York City entertainment, concentrated at, but not limited to, the Times Square area. Clearly, the success of the query '*new york*' is strongly dependent on the local context of the search session, which supports the need for context-aware query failure prediction. The fact that either query expansion or reduction can lead to query success illustrates the difficulty of query failure prediction and motivates us to propose a query failure approach that goes beyond simplistic characteristics of query sequences within a search session when designing our approach.

A critical insight is that the problem of context-aware query failure prediction cannot be reduced to the problem of predicting the failure of specific queries independently of their context in the search session (e.g., the query string *'new york'*). Figure 3 depicts statistics for different queries calculated over ~108K queries drawn from users' interactions with several video search engines. The relative frequency of queries with different success rates is shown in terms of bars color coded according to the video search engine where they were submitted. The success rate of a query is defined as the relative proportion of cases in which a user submits that query to the search engine and subsequently clicks on a result in the returned results list. A query with a success rate of 0% never results in a user click and a query with a success rate of 100% always results in a user click. The majority of the queries across all engines can be seen in Figure 3 to fall midway between the two extremes. The distribution provides evidence that a single query can correspond to different underlying information needs, corresponding to different user decisions to click on a result. This evidence provides support for our position that context-aware query prediction is the most productive form of query failure prediction. This form of prediction does not treat all instances of a query equally, but rather takes the context of the query within the search session into account when predicting search failure.

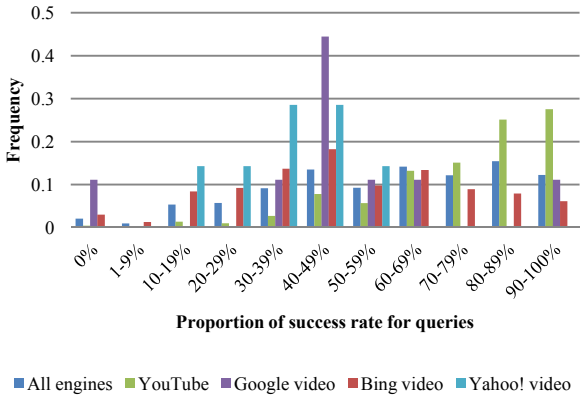

Figure 3: Distribution of success rates for queries submitted to four major video search engines.

The starting point of our proposed approach are *query performance prediction* (QPP) techniques that have been developed in the area of conventional text-based information retrieval [2, 7, 11, 32]. These techniques treat the query independently of its context. Our approach builds on QPP and extends it with two innovations. First, we extend conventional, context-independent QPP with user indicators that are derived from session information drawn from a video search engine log. These features allow us to take the context of a query into account when predicting failure. Second, we demonstrate the potential of adapting context-independent QPP specifically to video search by extending a conventional text-based predictor with visual features drawn from the video search results list (specifically, from thumbnails accompanying individual video results). To the best of our knowledge, this work is the first to treat the prediction of query failure for video search engines. Here, we restrict ourselves to tackling the failure prediction problem, but note that the ultimate value of our approach will be its usefulness to steer the application of already existing search engine performance enhancing techniques, thereby maximizing their benefit. The suitability of our query failure prediction approach to steering other methods will be critically dependent on its ability to operate online during a user search session. For this

reason, we direct our attention at information derived from the ongoing search session, i.e., we exclude the possibility of learning user profiles or global search patterns in order to predict failure. By limiting ourselves to sessions we ensure that our approach can be tested or applied without access to the large amounts of transaction log data that are used to carry out the investigations reported here.

The main contributions of this paper are twofold:

- We carry out a qualitative analysis of search engine logs to build an understanding of why video search engine queries fail.
- Building on observations made during the analysis, we propose and validate a framework for query failure prediction.

Validation is carried out in the form of an extensive series of experiments designed to answer the following research questions: RQ1: Do context-aware user indicators derived from transaction logs improve the performance of query prediction over conventional context-independent QPP? RQ2: Which user indicators extracted from transaction logs are particularly helpful? RQ3: What is the potential of visual features to improve conventional text-based QPP in the case of video search? RQ4: Can QPP integrating visual features be combined with context-aware user indicators to further improve search results? The experimental evaluation is completed with a detailed analysis of results that demonstrates its particular usefulness for long-tail queries.

The remainder of the paper is organized as follows. After positioning our proposed approach with respect to the related work in Section 2, we carry out a qualitative analysis of search engine logs in Section 3 and explain our approach in detail in Section 4. Section 5 reports the experimental validation of our approach. We discuss our results in Section 6 and conclude the paper with an outlook on future work in Section 7.

2. RELATED WORK

2.1 Search Failure

Our definition of query failure adopts widely-applied assumptions from the field of text retrieval. Joachims [15] showed that a click on a search result indicates a relevance judgment: a user scans the top search results returned by a search engine for a particular query from the top to the bottom and must have noticed result 1 before clicking result 2, making a decision not to click result 1. Fox et al. [3] collected and compared both implicit and explicit relevance judgments of search results and showed that implicit judgments, including clicked search results, are indicative for actual relevance. Using these observations, for our investigations we consider a query instance *successful* if it leads to a search results click and, consequentially, *failed* if it does not lead to a click in the generated video search results list. The same assumption concerning the connection of clicks and relevance has been applied in [8, 10].

Work investigating search failure has been limited, both in the text-based information retrieval community as well as in the multimedia retrieval community. The existing work can be categorized according to the exact definition of failure that is adopted: a failed query can either be a query that performs poorly with respect to a retrieval metric that measures results relevance or be a query that returns no results (i.e., a zero-hit query). Our work falls into the former category, in contrast to important recent work, which falls into the latter category. Important recent work from the field of information retrieval are Mastora et al. [21], who showed that failed queries represent 36% of overall submitted queries, and Kim and Can [16], who investigated failure as a factor characterizing queries in different search tasks. In the field of multimedia retrieval, Pu [22] compared characteristics of successful and failed image queries and Kofler et al. [17] investigated video search failure search using expressions of failed user needs posted on a social forum.

2.2 Query Performance Prediction

Query performance prediction (QPP) originated in the area of conventional text information retrieval and aims to predict whether a particular query will have a high or low retrieval performance. In our work, we make use of the most widely-used and highly successful post-retrieval QPP predictor, the clarity score [2], which captures the topical focus of the results list returned by the search engine in response to the query. We adopt this post-retrieval QPP predictor to inspire our engine indicators, applying the assumption that a low clarity score reflects a high likelihood of failure. Conventional text retrieval also uses pre-retrieval QPP predictors [7, 11, 32], although they are generally overall less successful than post-retrieval predictors. Pre-retrieval predictors do not take the actual results list produced by a search engine into account but rather analyze different aspects of the query, such as specificity, ambiguity, and term relatedness of queries. We adopt both a post-retrieval predictor [2], as well as specific pre-retrieval predictors [7, 11, 32] as baselines to test our proposed approach.

2.3 Transaction Log Analysis

Transaction log analysis (TLA) was established in the field of text information retrieval. In our work, we use transaction logs both as a source of ground truth concerning which queries have failed and also as a source of information on user interactions, which we use to create user indicators. The literature has shown user interactions derived from transaction logs to be a reliable source of information that can be used to make predictions on aspects of user search sessions. Key examples from the area of text retrieval include White et al. [28], who predicted a user's short-term interests based on context derived from the current query and pre-query session activity, and Kotov et al. [18], who predicted which queries and tasks are related to each other across sessions using features of individual queries and long-term user search behavior. Xiang et al. [29] used information from query generalization, reformulation and specialization to improve ranking. Our work makes use of similar indicators, but as part of a much broader inventory of user indicators. The ultimate benefit of our work is also improved ranking, but because we target query failure and not exclusively ranking, we leave the possibility open for other applications (such as improved query suggestion). Guo et al. [4] can be considered to be the work most similar to our own since they use the query, search results and user interaction to predict query performance. They found that the way in which users interact with search results provides a strong indication of the quality of those results. However, this work differs from our approach in that query performance is predicted independently of the other queries constituting the context of the query within the search session. Comparable little work has been carried out applying TLA for multimedia search. Jansen et al. [14] and Tjondronegoro et al. [27] show the form of multimedia queries differs from text queries. These differences provide motivation for our own work, which is dedicated specifically to video search.

3. LOG ANALYSIS OF FAILED QUERIES

In this section, we report on a qualitative analysis of query failure in video search engine logs, which identified useful categories of context-aware user indicators (i.e., features which capture the context of a query within a search session) in our query failure prediction approach.

3.1 Transaction Log

The investigations reported in this paper make use of a large-scale anonymized transaction log of the Bing video search engine collected via the Internet Explorer 8 Internet browser used by millions of users world-wide. The transaction log contains, in addition to information on search engine-based interactions, all interactions of a user with the Internet browser and provides a realistic picture of the video search behavior of Internet users. Entries of the transaction log contain a timestamp for each interaction with the Internet browser, the anonymized URL of each request, and the URL of the preceding request. Since there is one log entry for each transaction, not every transaction has a query. If there is a query present, the transaction contains both the query string and the search engine vertical on which the query was executed (e.g., Web, image, video).

The investigations and experiments described in this paper are based on sessions logged from June 1, 2011 until June 7, 2011. Typical of transaction logs [12, 27], our logs contain corrupted data entries like missing fields or information, resulting from errors in the logging procedure and were subjected to a standard cleaning process [12]. To remove variability introduced due to regional or linguistic differences, we take into account exclusively user interactions of US-based English speakers. Sessions start with the first interaction of the user with the Internet browser, terminate if there is an inactivity of the user exceeding 30 minutes and have at least one search engine interaction. Session data did not contain information about different Internet browser tabs or windows used by users while surfing; therefore, we ordered user actions within each session chronologically.

For the purposes of this paper, we focus on sessions containing at least one query submitted to the video vertical of the search engine. This results in 174,955 sessions having 1,218,936 total and 445,859 unique queries. The average session duration was 11 minutes and 18 seconds containing an average of 13 actions per session. The video search engine does not adapt to any context of the user. We can therefore assume that users, independently of how they interact with the video search engine or where they are located, will receive similar search results for the same query. Further details, including the separation of the development and test sets, are explained in Section 5.1 'Experimental Framework'.

3.2 Observations

We performed an exploratory investigation to extract observations decisive for failure, where we zero in on user indicators from the transaction log. We sampled the 100 most-submitted queries having both, successful and failed query instances in our development set and manually investigated the search session in which they were submitted. The result of our transaction log analysis was a number of observations concerning sources of query failure.

The first observation concerns other interactions within the search session that may not have proven satisfactory to a user. In our observations we discover cases in which clicked videos are watched for a relatively short amount of time, resulting in either a query reformulation or further investigation of the original search results list. Similar observations have been made in [3]. Continuous switching behavior between query formulation and watching results for a short amount of time has implications for how long the search session lasts and how many interactions the user has with the search engine and where in the local context of the search session a query was submitted [5]. We define this observation as:

Previous dissatisfaction: If the user is dissatisfied with results thus far in the session, the current query is more likely to fail.

The second observation concerns query reformulation behavior. A user might reformulate a query after it has failed to return acceptable results: the query was either too specific or too general. According to [13, 29] query reformulation follows four strategies: formulating a *new query*, and *generalizing*, *specializing*, or *reformulating* a query. With the exception of 'new query,' all other strategies involve the same query topic with variations depending on whether the user wants to get more general (e.g., *'spongebob lost episode'* → *'spongebob'*), specific (e.g., *'spongebob'* → *'spongebob dance'*), or other results (e.g., *'spongebob lost episode'* → *'spongebob house party'*). It is not necessarily decisive that the more specific a query gets, the higher the probability that the query instance will be successful: the video search engine might return no useful results for a too specific query, leading the user to submit generalizations of the same query. Figure 4 contains an example that ultimately leads to a successful query. Similar observations have proven useful in related investigations [4, 18].

how to make a napoleon total war comentary	✗
how to make a commentary video for total war game	✗
how to make a commentary video for napoleon total war	✗
how to upload a video from a game	✗
how to upload a video from a videogame	✓

Figure 4: Excerpt of a search session showing iterations of query reformulation.

Lack of clicking behavior during reformulation was also revealed as important by our investigation. We summarize our findings on interactions with an observation on query iterations:

Query iterations: The more query reformulations (without search result clicks), precede a query in a search session, the more likely the query will not fail.

How clear a user expresses his search goal is crucial for query failure. A search goal is the high-level goal a user wants to achieve using video search and results in a set of user actions [23]. We observe that users with different types of search goals engage in different interactions (i.e., users who freely browse vs. others pursuing a specific goal). Figure 5 contains an example from the log of a sequence of queries that are not clearly related and all fail.

body snatcha ent	✗
hood to hood atlanta	✗
Facebook	✗
Utube	✗

Figure 5: Excerpt of a search session showing a sequence of video search queries that are not clearly related.

We observe that browsing users tend to switch to other verticals of the search engine and vary in how often a particular search result was clicked. We summarize these findings with the following observation:

Goal-lessness: If the user is browsing randomly without a specific search goal, the more likely that the current query does fail.

In contrast to browsing behavior, our analysis revealed that in some cases, users will enter successively more focused queries, suggesting that these users are not browsing, but rather pursuing a specific goal. Figure 6, illustrates a series of queries from the log that is increasingly specific and ultimately leads to success.

sons of anarchy	✗
sons of anarchy season premier	✗
sons of anarchy season 4 premier	✓

Figure 6: Visualization of a search session showing the development of a search goal and its query expressions.

We conjecture that how directed the goal formulation of a user is, is represented by topical focus or topical overlap expressed in consecutive queries. We define this observation as:

Goal-directedness: If the current query expresses a more specific search goal formulation than the previous queries, the current query is less likely to fail.

Our final observation concerns how familiar a user is with interacting with a search engine. This observation concerns whether the search session reveals evidence that the user understands the possibilities offered by the search engine, including the search engine verticals. We define this observation as:

User familiarity: Within the bounds of a given search session, the greater the familiarity with the video search engine demonstrated by the user, the more likely that the current query does not fail.

4. APPROACH

4.1 Overview

Figure 7 contains an overview of our approach. We define a set of features (Sections 4.2 and 4.3) corresponding to each of the observations obtained from the transaction log analysis of failed queries. In an offline step, we extract these features from the local search session and use them to train classifiers (Section 4.4), which predict whether the current query a user is submitting to a video search engine will fail or be successful.

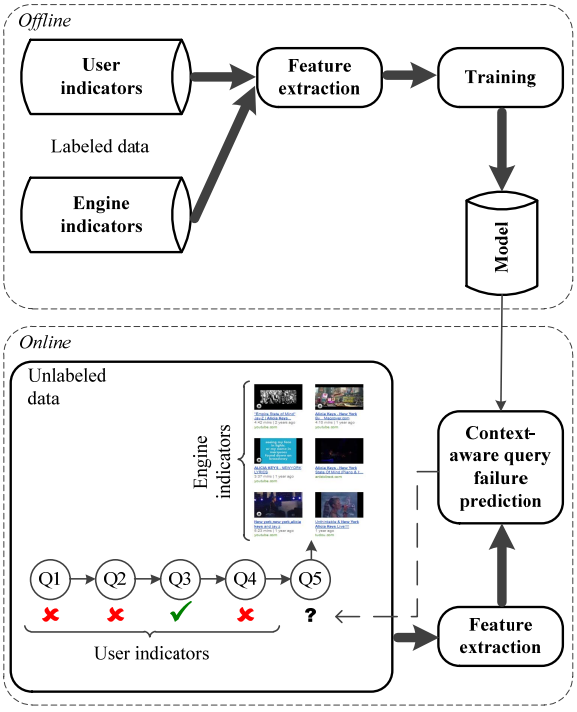

Figure 7: Overview of our context-aware query failure prediction approach for video search including model training in an offline step and the actual prediction in an online step.

In an online step, we again extract features based on user indicators (in the example above: query instances Q1-Q4 and corresponding click-through- and session-data) and engine indicators (in the example above: visual search results of the current query instance Q5) from a user's search session and perform the context-aware query failure prediction step for the current query instance (in the example above: Q5).

4.2 User indicator features

Since our context-aware query failure prediction approach aims to predict the outcome of the current query instance the user submits to the video search engine, we extract features from the local search session relative to this query. Previous queries are considered queries submitted chronologically before the current query. We differentiate between two types of pre-query session histories: the *session query history* is the set of all queries submitted in a search session before the current query; the *query-specific reformulation history* (*QRH*) is the local set of queries submitted in a search session *before* the current query. It is a sequence of topically related queries within a session and a subset of the session query history.

Table 1 gives an overview and description of features derived from user indicators and grouped by observation. We extract features related to the different sources of information associated with the transaction logs, which can be characterized by the following three groups (i) general Internet browser session and search session statistics, (ii) query (re)formulation behavior and clarity of search goal expressiveness, and (iii) click-through data in the video search results lists generated by the queries in the search session. Note that we do not claim that the features are independent of each other. We investigate a variety of features corresponding to each observation and later use feature selection to choose the most useful subsets for classification.

For most features, the information provided in Table 1 is sufficient to explain their generation. However, a few cases require additional details, which are explained here. We use features calculated on the QRH and also on the whole session in order to supply the classifier with information at different local resolutions. A QRH boundary is set by classifying subsequent queries into query-reformulation classes and then defining a boundary whenever the class type changes. The class types that we use are: reformulations, generalizations, and specializations and they are calculated using the algorithm proposed in [13]. The algorithm takes query statistics such as overlapping terms and query length into account. The feature *Count{Reformulated, Generalized, Specialized}Queries[QRH]* uses counts of query-reformulation types generated by the algorithm.

Query clarity [2] is a conventional post-retrieval query performance predictor based on the assumption that a query for which the search engine performs well will enjoy topical coherence among the top items (referred to as documents due to the origin of the clarity score in text retrieval) returned in the results list. The predictor estimates a language model from search results for each query and a language model of the overall document collection as shown in Eq. 1 (R is the set of retrieved documents, D is a document, w is a term in the overall vocabulary V, and Q is a query).

$$P_{QM}(w) = \sum_{D \in R} P(w|D)P(D|Q) \qquad (1)$$

The similarity between these language models, usually defined as the *Kullback-Leibler* (KL) divergence, categorizes the clarity of the query (Eq. 2).

$$QC = D_{KL}(P_{QM} \| P_{CM}) = \sum_{w \in V} P_{QM}(w) \log \frac{P_{QM}(w)}{P_{CM}(w)} \qquad (2)$$

We take the difference between the clarity scores as an indicator (*QueryClarityDifferenceToAllQueries[QRH]*) for parts of the observation about goal-directedness.

Table 1: Overview of the user indicator features based on our observations from the log analysis.

Feature name	Feature description
Previous dissatisfaction	
CountRequests	Number of requests in the session query history excluding the current query.
SessionDuration	Duration of session until current query in seconds.
TimeToPreviousQuery	Time difference in seconds between current query and previous query.
RankSubmitQuery[QRH]	Current number of total submitted queries in the session query history/query-specific reformulation history.
RatioResultsClicked{Videos, Web, Other}	Average number of {video, web, non-web- and non-video-vertical} results clicked per unique {video, web, non-web- and non-video-vertical} query.
CountVideosWatchedDuration[QRH]	Number of videos watched for a duration less than k seconds (k set here to 15) in the session query history/query-specific reformulation history.
Query iterations	
Count{Reformulated, Generalized, Specialized}Queries[QRH]	Number of {reformulated, generalized, specialized} queries in the session query history/query-specific reformulation history.
RatioUnclickedQueriesVideos[QRH]	Average number of unclicked video search results per unique video query in the session query history/query-specific reformulation history.
Goal-lessness	
CountVideoQueries{Total, Unique}	{Total, Unique} number of video-vertical queries submitted in the query session history.
CountMutualExclusiveQueryTopics	Number of mutually exclusive query topics in the query session history.
CountWebQueries{Total, Unique}	{Total, Unique} number of web-vertical queries submitted in the query session history.
CountOtherQueries{Total, Unique}	{Total, Unique} number of non-web- and non-video-vertical queries submitted in the query session history.
RatioVideo{Unclicked, OneClicked, MoreClicked}Queries	Average number of {unclicked, once-clicked, more-than-once-clicked} video search results per unique video query.
Goal-directedness	
QueryClarityDifferenceToAllQueries[QRH]	Difference between query clarity score of the query, calculated using the top n most relevant video vertical search results to all queries in the session query history/query-specific reformulation history.
QueryTermsOverlap{BothQueries, PreviousQuery, CurrentQuery}Expanded	Number of terms which occur {in both the current query and the previous query, only in the previous query, only in the current query} using query expansion based on co-occurring terms of top n most relevant video vertical search results.
User familiarity	
CountVerticalSwitches	Number of vertical switches before current query.
CountUniqueDomainsVisited	Number of unique domains visited in the overall session history.
CountSearchEngineSwitches	Number of search engines used before current query.
QueryAlreadySubmitted	Flag indicating whether the current query was already submitted.
RatioWeb{Unclicked, OneClicked, MoreClicked}Queries	Average number of {unclicked, once-clicked, more-than-once-clicked} web search results per unique video query.

Indicative for a query and its development with increasing session duration is how well topics overlap expressed in consecutive queries. We expand queries by using the top n videos returned on the video vertical of the search engine. If, within these video search results, a term had a higher-than-average co-occurrence with a query term, it was used to expand the query. For calculating the query clarity score and performing query expansion we use the title information of the top n = 25 video search results returned by the search engine.

4.3 Engine indicator features

Our engine indicators are features that extend insights from conventional text QPP to video search results by exploiting visual information of thumbnails associated with each search result. We assume that the consistency of the visual content of top search results will reflect the topical focus of the results list. High consistency should then indicate that the search engine has achieved good performance on the query that generated the results list.

Visual thumbnails can be represented and the visual content consistency can be evaluated at different levels of abstraction, such as the pixel-, visual object-, semantic concept-, scene composition- or scene interpretation-level. Insights acquired in [24] regarding the potential of the visual channel to help retrieve videos using topical queries showed that representation of the visual channel at higher abstraction levels, and in this particular case using distributions of outcomes of a selected number of semantic concept detectors aggregated from long videos, are helpful for this purpose.

While we conjecture that a similar approach could be developed for evaluating the results lists consisting of static visual thumbnails, in this work we do not search for the appropriate representation of the thumbnails, but limit our goal solely to showing, in a light-weight fashion, the potential of the visual information contained in the thumbnails to be helpful for the purpose of such evaluation. We conjecture, namely, that if engine indicators solely based on low-level features (i.e., pixel-level representation) are capable of contributing to query failure prediction, then thumbnail representations at higher abstraction levels will likely make an even larger contribution. We therefore represent video search results visually by extracting standard low-level visual features from thumbnails.

Each visual thumbnail has representations based on local and global visual features. We use a Pyramid Histogram Of visual Words (PHOW) [1] representation for appearance and extract SIFT descriptors from each image on a dense grid. We quantize the descriptors using a vocabulary of 1,000 clusters and compose them in a pyramid histogram. For shape information of the entire image we use a Pyramid Histogram Of Gradients (PHOG) [1] representation. We use a Pyramid Self-Similarity (PSS) [25] representation for capturing the layout of shape context and, similarly to PHOW, quantized descriptors into 1,000 clusters. In general, pyramid histogram construction is done using three levels. Global features were extracted using the open source implementation Lucene Image Retrieval (LIRe) [20]. We extracted the following features: color histograms; MPEG-7 descriptors scalable color, color layout and edge histogram; Tamura texture fea-

tures coarseness, contrast and directionality; Gabor; color and edge directivity descriptor; fuzzy color and texture histogram; and auto color correlation feature. Visual features are extracted from thumbnails at indexing time and the variance among these thumbnails is calculated in real-time at query time.

The consistency among the top `n` visual search results is calculated in terms of the variance *VD* among the visual thumbnails, represented by visual descriptors. Visual variance within a results list has also been used in [19] to weigh feature vector components and can be formulated as

$$VD = \frac{1}{|RL_n|} \sum_{i=1}^{|RL_n|} \left(\overrightarrow{V_{v_i}} - \vec{M}\right)^2 \quad (3)$$

where RL_n is the set of top `n` video search results and where `n = 25` in our case. Further, $\overrightarrow{V_{v_i}}$ is the i-th visual thumbnail and $\vec{M}$ is the mean of the thumbnails' descriptor components.

4.4 Model Training and Prediction

In the offline step, we apply supervised learning to train generic classifiers on our development set using the extracted features. Ground truth for query instance failure was generated using click information from the transaction log on search sessions. We formulate the context-aware query failure prediction task as binary classification: query instances falling in the class query instance failure `+qif` (query instance success `-qif`) are considered failed (successful).

We trained one binary classifier for each feature set representing a particular observation from user and engine indicators and combine all features of user and engine indicators in generic classifiers trained on all features. For combination of features we use both early and late fusion. For early fusion, all involved features are combined in one single feature vector used for training. For late fusion, different classifiers are first trained and their prediction results are combined in one final feature vector, which is again used for training a final classifier.

We adopted well-known classifiers for our task. Random Forest classifiers are sets of decision trees that return the most likely class label based on the results of the individual internal decision trees. They run efficiently on large datasets and estimate what features and feature sets are crucial in the classification process. We further investigated the usage of rule-based classifiers, pure decision trees, and Naïve Bayes classifiers. Their performance was comparable with the one achieved by the Random Forest classifier, therefore we do not report specific experimental results using these additional classifiers. We use the classifier implementations from the Weka data mining framework [6]. For feature selection we use the *CfsSubsetEval* implementation combined with *Best First search*.

Although we make use of supervised learning, similar to [30], our approach is not restricted and applicable to predict query failure for unseen queries: in an online search session, we extract features from the user indicators and engine indicators and use our trained models in order to make a prediction.

5. EXPERIMENTS

5.1 Experimental Framework

Dataset: From the set of total queries we randomly sampled 108,692 queries for our development set (56,396 unique queries) and 1,110,244 queries for our test set (429,482 unique queries), resulting in a session-based separation of 24,734 and 150,221 sessions for the development and test set, respectively. The development set and test set contain mutually exclusive search sessions. 392,809 unique queries exclusively occur in the test set. Additionally to the user actions available in this transaction log, we collected information from the search results produced by the search engine for each query in our development and test set. We collected textual information—the titles—and visual information—the visual thumbnails—of the 25 most-relevant search results returned by the video search engine. We assume that the first results page of search results contains around 25 documents which are most decisive for the user's satisfaction with the query [3].

Baselines: In order to show that our prediction results using user indicators and engine indicators are non-trivial, we compare them against a set of baselines represented by existing query performance prediction methods: post-retrieval methods are represented by query clarity (QC) [2], and pre-retrieval methods are represented by normalized similarity score between query and collection (NSCQ) [32], query scope (QS) [7], normalized co-degree score (NCDS) [11], and query length (QL) [7]. These baseline scores are calculated using the title information of the top `n = 25` video search results returned by the video search engine. We further calculate the dominant class (DC) baseline—the larger class of `-qif` and `+qif` on our ground truth and assume the worst case that all query instances belong to this class. For the total number of 1,110,244 query instances in our test set, 667,600 are labeled as `-qif` in the ground truth and 442,644 as `+qif`.

Evaluation: The classification results are generated using *10-fold cross-validation* on our test set. Results are reported in terms of Weighted F-measure (WF), defined as F-measure—the harmonic mean of precision and recall—weighted by class size of classes `-qif` and `+qif`. F-measure (F) performance for individual classes `-qif` and `+qif` are also provided in order to see the differences between prediction performance for the two classes. Statistical significant tests were performed using the Wilcoxon signed-rank test ($p<0.05$).

5.2 Baseline Performance

All baseline performances are reported in Table 2.

Table 2: Performance overview of QPP baselines.

Baseline	WF	F (`-qif`)	F (`+qif`)
DC	0.4516	0.751	0.253
QL	0.5453	0.727	0.271
NCDS	0.5669	0.751	0.289
QS	0.6523	0.731	0.533
NSCQ	0.6810	0.744	0.586
QC	**0.6862**	**0.748**	**0.593**

Both, NSCQ and QC baselines achieve a good balance between correctly classified instances of `-qif` and `+qif`, however QC outperforms NCSQ. The relatively strong performance of the conventional QPP baseline demonstrates the potential and the strength of the text-retrieval methods to transfer to video retrieval problems. For the remainder of the experiments we compare performance against the best-performing conventional QPP baseline achieved by the query clarity score.

5.3 Prediction Using User Indicators

Here we report results of our methods based on user indicators from the search log. A performance overview is presented in Table 3. We (i) perform feature selection using feature sets derived from our observations of the transaction log, and (ii) combine these feature sets in an early and late fashion. We perform classification derived from features from individual observations and report performance for each observation.

Table 3: Performance overview of user indicator methods.

Features	WF	F (-qif)	F (+qif)
Previous dissatisfaction	0.7647	0.821	0.680
Qu. iterations (all)	0.7153	0.771	0.631
Qu. iterations (QRH)	0.7257	0.783	0.639
Goal-lessness	0.7322	0.783	0.655
Goal-directedness	0.7161	0.778	0.623
User familiarity	**0.7694**	**0.826**	**0.684**
Early fusion	0.7620	0.821	0.673
Late fusion	0.7678	0.820	0.688

Our user indicator-based query failure prediction methods statistically significantly outperform the conventional QPP baseline (QC in Table 2) and achieve an 8% improvement in absolute performance solely by taking local search context into account. The best-performing method is the classifier built on features derived from 'User familiarity'. Another strong performer is 'Previous dissatisfaction', reflecting previous failures in the session. For the observation 'Query iterations', using local features from the query-specific reformulation region of the search session increases the performance compared to using the entire query history results, suggesting the value of using narrow local context. The relatively poor performance achieved by observation 'Goal-directedness' suggests that search goal clarity evolving over a search session is not consistent. Early and late fusions perform well but do not succeed in outperforming individual well-performing observations. Looking at F-measure values of individual classes shows that classifying `+qif` using the proposed classifiers is more conservative than classifying `-qif` instances. Observations clearly achieve a much better result for `-qif` than for `+qif`. The characteristics of successful queries are presumably more stable, most likely reflecting the relatively greater stability of the characteristics of the successful query.

These results allow us to give a positive answer to our first and second research questions RQ1 and RQ2. Context-aware user indicators do outperform conventional context-independent QPP and certain individual observations emerge as more helpful. Because we are interested in investigating a general combined classifier that is robust to the incidental failure of any one particular feature, we continue our experiments on the whole set of features, rather than singling out the top performing individual features.

5.4 Prediction Using Engine Indicators

Here we report results of our methods based on engine indicators from the relatedness of the video items in the search results list. For our experiments, we use the variance among the visual search results (Eq. 3) inspired by the same principle as clarity score. A performance overview is presented in Table 4. We experiment with (i) global and (ii) local visual descriptors of the visual thumbnails in the search results list and (iii) further combine these descriptors using early and late fusion. We performed exploratory experiments with feature selection, which did not show appreciable differences.

Table 4: Performance overview of engine indicator methods.

Features	WF	F (-qif)	F (+qif)
Global visual feat.	0.7348	0.787	0.657
Local visual feat.	0.7349	0.787	0.656
Early fusion	0.7350	0.787	0.657
Late fusion	**0.7356**	**0.788**	**0.656**

All classifiers built using engine indicators statistically significantly outperform the best-performing QPP method (QC in Table 2). Combining existing QPP methods with the variance of the video items in the search results list expressed in terms of visual variance adds 4% improvement in absolute performance. The best performance is achieved using the late fusion approach of classifiers based on visual variance among the search results list represented by global and local visual features extracted from the visual thumbnails.

Since adding information of visual results list relatedness in terms of variance among search results for video search increases classifier performance, we can give a positive answer to our third research question RQ3: visual features have the potential to improve conventional text-based QPP in the scenario of video search. The improvement for query failure predication is quite modest, however, consistent with our expectations for our relatively simple visual representations. However, the demonstration that even a simple representation is able to outperform conventional text-only QPP approaches is a clear sign that the visual component of video search results should not be ignored, but rather potentially makes an important contribution to query failure prediction.

5.5 Prediction Using Combined Indicators

Here we report results on the integration of user indicators and engine indicators achieved by combining all features from both indicators using early and late fusion. A performance overview is presented in Table 5.

Table 5: Performance overview of combined methods.

Features	WF	F (-qif)	F (+qif)
Early fusion	**0.7744**	**0.830**	**0.690**
Late fusion	0.7692	0.819	0.694

The results achieved by our combined methods statistically significantly outperform the best QPP baseline (QC in Table 2). The best performance is achieved using early fusion which achieves almost 9% absolute (13% relative) performance improvement over the baseline. Combining both indicators is helpful for context-aware query failure prediction since the combination exceeds the top performance of user indicators and engine indicators used separately (cf. Table 3 and Table 4).

These results allow us to answer RQ4 positively: there is indeed a potential for engine indicators encoding the visual variance to further improve query failure prediction beyond the performance achievable with only transaction log-derived user indicators.

6. DISCUSSION

To attain a better understanding of our approach and gain additional insights we investigate the prediction results in more detail. Our first investigation focuses on the performance of individual queries and their local search session context in order to understand the usefulness of user indicator features. For each query in our test set, we calculated the performance improvement achieved by our late fusion approach over our best-performing QPP baseline. Three classes of queries emerged as being particularly dramatically improved by user indicator features. We specify the observation corresponding to most important features used by the random forest classifier for each class.

- Queries such as *'youtube'* that returned results that could technically be considered relevant, but did not satisfy the user. We conjecture that users do not click in these results lists because they were actually looking for personalized recommendations and not general results. (Important features from observations 'Goal-lessness' and 'Goal-directedness'.)
- Queries such as *'free movies'* that do not return interesting results due to lack of availability of content. (Important features from observations 'Previous dissat.' and 'Query iterations'.)

- Queries that were submitted to the video vertical of the search engine, but were probably intended for another vertical (*'yahoo mail login'*) and/or were misspelled or otherwise malformed (*'rkellyvideo', 'facebok'*). (Important features from observations 'User familiarity' and 'Previous dissatisfaction'.)

We turned to investigate queries for which user indicator features fail to outperform the baseline and discovered in many cases that these were 'cold-start' queries, i.e., queries which were submitted in the beginning of a user's search session. Their failure is understandable since the session must accumulate sufficient information about query (re)formulation behavior and click-through data to make an appropriate decision. In general, features based on the query-specific reformulation history perform better than features extracted from the whole session query history, implying that the locality of the search session context makes an important contribution for our approach and that only very little information is needed to address the cold start issue.

Our second investigation focuses on the use of representations of visual variance. Visual variance among video search results can, to a certain extent, positively influence poor predictions based on user indicators alone. Figure 8 gives examples of the two queries for which the combination of user indicators and engine indicators outperforms the prediction of our best-performing baseline and helps making correct predictions.

(a) Query *'justin bieber'* (`-qif`) correctly predicted

(b) Query *'transformers autobot'* (`+qif`) correctly predicted

Figure 8: Two example queries for which the combination of user indicators and engine indicators positively influence query failure prediction.

For (a), all visual thumbnails show similar content (Justin Bieber in different situations), while for (b) completely different visual thumbnails are presented to the user (scenes from both the comic and the movie of 'Transformers' and a thumbnail from an interview). Based on the idea of visual variance among thumbnails in the search results list, the thumbnails help making a correct prediction. We emphasize that the effect is not dramatic, but that our experimental results clearly serve to demonstrate its potential.

Our final investigation involves the performance of our context-aware query failure prediction approach for video search carried out on long-tail queries, i.e., queries that are very rarely submitted to the video search engine. The long tail is important, as witnessed by the fact that in our test set 36% of all video queries were submitted only one time. We visualize the performance improvement between our best-performing approach using both user indicators and engine indicators and the best-performing conventional QPP baseline and average this improvement over sets of queries submitted with the same frequency (i.e., equally often) to the video search engine. Figure 9 contains this averaged improvement ordered by decreasing query frequency (queries submitted ~8,000 times down to queries submitted 1 time).

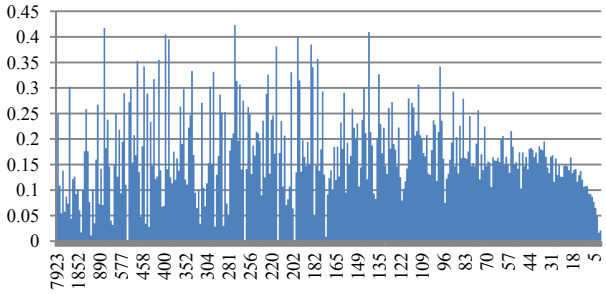

Figure 9: Performance increase between our best-performing method and best-performing baseline, averaged and ordered (from left to right) by decreasing query frequency.

We observe that improvement of the combined approach over the QPP baseline occurs throughout the full range of query frequencies. This distribution demonstrates that the contribution of the search session context features is independent of the frequency of query submission. In other words, the approach performs well not only for general and often submitted queries (e.g., *'youtube music videos', 'clarice taylor', 'seinfeld'*), but also for long-tail queries (e.g., *'number one hit with sade', 'micheal jackon', 'boku no atama'*). Figure 9 also reveals that the improvement over the conventional QPP baseline is more consistent for the long tail, implying that our approach delivers particularly reliable improvement in cases where conventional QPP reaches its limitations, presumably due to too few search results or unexpected diversity. This observation demonstrates the performance our context-aware query failure prediction approach achieves in a specific situation in which the conventional context-independent approach is inadequate.

7. CONCLUSION AND OUTLOOK

We have presented a novel context-aware approach able to predict query failure in video search using user indicators and engine indicators. We performed an exploratory investigation in order to derive observations indicative for context-aware query failure prediction and designed a scalable approach using features derived from these observations. Our results, obtained via experiments performed on a large test set of 1+ million video queries, yield a series of important insights. First, they show that applying conventional QPP methods using textual information from search results lists provided by a video search engine yields reasonable results. We are able to confirm that the basic principle of QPP is transferable from conventional text retrieval to multimedia search and use the best-performing conventional text QPP as baseline against which to compare our proposed method. Second, our results show that conventional QPP is too rigid and that user indicator features derived from the local context of the search session is useful and improves the prediction of query failure. Third, the basic success of QPP motivates us to further adopt a QPP-related consistency criterion to visually capture the quality of the video search results list using the thumbnails associated with each result. This relatively simple engine indicator using low-level thumbnail presentations exploits visual variance of results lists and is capable of delivering a modest, however promising improvement that confirms the potential of visual features. Finally, our results show that the good performance already achieved using prediction approaches based on user indicators can further be improved by combining them with engine indicators.

Our future work will address several aspects. First, we will work on the improvement of engine indicators using visual information

from the video search results list. As suggested in Section 4.3, a higher-level representation of thumbnails, such as using the scene composition level of abstraction rather than operating solely on the pixel level is likely to further improve our initial findings. Following the same reasoning, including additional sources of visual information such as motion thumbnails, which are played back once the user hovers over a static thumbnail of a search result, can be assumed to further make a positive contribution to engine indicators. Second, we would like to integrate query performance prediction into video search engines, since the experiments reported here reveal that the performance of our query failure prediction approach is already strong enough to be useful in applications. A promising place to start would be the optimization of ranking, since, as mentioned above, initial work from the text retrieval field has demonstrated that a restricted set of search-session-derived indicators can be successfully exploited to optimize ranking [29]. Ultimately, however, we anticipate that the ability to predict when a video search query would fail will serve to enhance the performance of an entire range of video search engine optimization techniques. For example, a video search engine can attempt to back-off to concept-based video retrieval for queries for which other forms of video retrieval are predicted to fail. Alternatively, query suggestion could be adapted based on the goal of bringing users more quickly to highly successful queries. Further, since search engines are steadily improving their performance and functionality—e.g., the implementation of motion thumbnails, which allows users to briefly preview video without actually clicking it—we aim to work on adaptive query failure prediction for video search which can easily accommodate improvements of search engines and provide consistently good prediction results.

8. ACKNOWLEDGMENTS

This publication was supported by the Dutch national program COMMIT. The authors would like to thank Bing Lang for providing the transaction logs and Yang Yang and Guangxin Ren for their constructive suggestions and support.

9. REFERENCES

[1] Bosch, A., Zisserman, A. and Muoz, X. Image Classification using Random Forests and Ferns. 2007.

[2] Cronen-Townsend, S., Zhou, Y. and Croft, W. B. Predicting query performance. In SIGIR (2002). ACM, 299-306.

[3] Fox, S., Karnawat, K., Mydland, M., Dumais, S. and White, T. Evaluating implicit measures to improve web search. ACM Trans. Inf. Syst., 23, 2, 2005, 147-168.

[4] Guo, Q., White, R. W., Dumais, S. T., Wang, J. and Anderson, B. Predicting query performance using query, result, and user interaction features. In Adaptivity, Personalization and Fusion of Heterogeneous Information (2010), 198-201.

[5] Guo, Q., White, R. W., Zhang, Y., Anderson, B. and Dumais, S. T. Why searchers switch: understanding and predicting engine switching rationales. In SIGIR (2011). ACM, 335-344.

[6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. The WEKA data mining software: an update. SIGKDD Explor. Newsl., 11, 1, 2009, 10-18.

[7] He, B. and Ounis, I. Inferring Query Performance Using Pre-retrieval Predictors. 2004.

[8] Hollink, V. and Vries, A. d. Towards an automated query modification assistant. ACM, 2011.

[9] Hsu, W. H., Kennedy, L. S. and Chang, S.-F. Reranking Methods for Visual Search. IEEE MultiMedia, 14, 3, 2007, 14-22.

[10] Huang, J. and Efthimiadis, E. N. Analyzing and evaluating query reformulation strategies in web search logs. In CIKM (2009). ACM, 77-86.

[11] Imran, H. and Sharan, A. Co-occurrence based predictors for estimating query difficulty. In International Conference on Data Mining Workshops (2010). IEEE Computer Society, 867-874.

[12] Jansen, B. J. Search log analysis: What it is, what's been done, how to do it. Library & Information Science Research, 28, 3, 2006, 407-432.

[13] Jansen, B. J., Booth, D. L. and Spink, A. Patterns of query reformulation during Web searching. J. Am. Soc. Inf. Sci. Technol., 60, 7, 2009, 1358-1371.

[14] Jansen, B. J., Spink, A. and Pedersen, J. The effect of specialized multimedia collections on web searching. J. Web Eng., 3, 3, 2004, 182-199.

[15] Joachims, T. Optimizing search engines using clickthrough data. In SIGKDD (2002). ACM, 133-142.

[16] Kim, J. and Can, A. Characterizing Queries in Different Search Tasks. In Hawaii International Conference on System Sciences (2012). IEEE Computer Society, 1697-1706.

[17] Kofler, C., Larson, M. and Hanjalic, A. To seek, perchance to fail: expressions of user needs in internet video search. In ECIR (2011). Springer-Verlag, 611-616.

[18] Kotov, A., Bennett, P. N., White, R. W., Dumais, S. T. and Teevan, J. Modeling and analysis of cross-session search tasks. In SIGIR (2011). ACM, 5-14.

[19] Leuken, R. H. v., Garcia, L., Olivares, X. and Zwol, R. v. Visual diversification of image search results. In WWW (2009). ACM, 341-350.

[20] Lux, M. and Chatzichristofis, S. A. Lire: lucene image retrieval: an extensible java CBIR library. In MM (2008). ACM, 1085-1088.

[21] Mastora, A., Kapidakis, S. and Monopoli, M. Failed Queries: a Morpho-Syntactic Analysis Based on Transaction Log Files. In (2011)

[22] Pu, H.-T. An analysis of failed queries for web image retrieval. J. Inf. Sci., 34, 3, 2008, 275-289.

[23] Rose, D. E. and Levinson, D. Understanding user goals in web search. In WWW (2004). ACM, 13-19.

[24] Rudinac, S., Larson, M. and Hanjalic, A. Exploiting noisy visual concept detection to improve spoken content based video retrieval. In MM (2010). ACM, 727-730.

[25] Shechtman, E. and Irani, M. Matching Local Self-Similarities across Images and Videos. 2007.

[26] Snoek, C. G. M. and Worring, M. Concept-Based Video Retrieval. Found. Trends Inf. Retr., 2, 4, 2009, 215-322.

[27] Tjondronegoro, D., Spink, A. and Jansen, B. J. A study and comparison of multimedia Web searching: 1997-2006. J. Am. Soc. Inf. Sci. Technol., 60, 9, 2009, 1756-1768.

[28] White, R. W., Bennett, P. N. and Dumais, S. T. Predicting short-term interests using activity-based search context. In CIKM (2010). ACM, 1009-1018.

[29] Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E. and Li, H. Context-aware ranking in web search. In SIGIR (2010). ACM, 451-458.

[30] Yang, L. and Hanjalic, A. Supervised reranking for web image search. In MM (2010). ACM, 183-192.

[31] Zha, Z.-J., Yang, L., Mei, T., Wang, M., Wang, Z., Chua, T.-S. and Hua, X.-S. Visual query suggestion: Towards capturing user intent in internet image search. ACM Trans. Multimedia Comput. Commun. Appl., 6, 3, 2010, 1-19.

[32] Zhao, Y., Scholer, F. and Tsegay, Y. Effective pre-retrieval query performance prediction using similarity and variability evidence. In ECIR (2008). Springer-Verlag, 52-64.