Master Thesis

Radboud University

Photo Collection Summarization

Author:
Luuk Scholten
s4126424

First supervisor/assessor:
prof. dr. ir. A.P. de Vries (Arjen)
Second assessor:
prof. dr. M.A. Larson (Martha)

October, 2017

Abstract

There is a need for the automation of photo collection summarization, which is a process that is often time consuming. Photo collection summaries for photo products should be relevant to the needs of the customer, cover all different aspects of the photo collection and avoid redundancy.

This work formulates creating an optimal photo collection summary as a ranked information retrieval diversification problem. The diversity of photo collections is estimated by detecting duplicates, segmenting photo collections into 'photo series' and temporal event clustering. Unique to the field of photo collection summarization, a customer dataset is extracted from a service for the creation of photo collection summaries. This dataset is used in a learning-to-rank setting to train a model to predict the relevance of a photo to a summary of its collection. High level features such as aesthetic quality, naturalness and indicators of important people are extracted. These features, along with photo collection and customer features, are fed in a pairwise manner to a machine learning model. The relevance and diversity estimations are used to balance the relevance and diversity of the final ranking. Two re-ranking methods which promote novelty in the result set are evaluated. Coverage over the aspects of the photo collection is achieved by proportionally re-ranking the novelty-optimized ranking over the detected temporal events.

Contents

Chapter 1

Introduction

In the modern era of smartphones and the internet, photos are one of the most important pieces of media being consumed. Popular services like Snapchat[1] and Instagram[2] actively promote the taking and sharing of photos. In 2016 alone 2.5 trillion photos are estimated to be shared or stored online (Sallomi & Lee, 2016). Every single one of these photos contains something that is valuable to the photographer or the person consuming the photo.

The days of photography being an obscure hobby and the days of waiting for film rolls to develop are long gone. Instead, with a single press of the button a photo gets captured and instantly stored in the cloud. The important memories, things, events and people encaptured in the photos are pushed away to the internet to be forgotten.

One way of remembering these important moments, people and events is to create photo collection summaries out of photo collections. However, with the ease and simplicity of capturing photos with modern devices, photo collections get larger and larger. It is a hassle to wade through all the photos and choose the ones that best summarize the photo collection as a whole. There is a need for automatic photo collection summarization to combat this hassle.

Which photos to use in these summaries is a difficult decision. The photos need to be selected according to their sentimental value. Of course, the selected photos need to be most beautiful pictures that depict these needs.

Next to this, it is important that the summary does not contain redundant content. Besides a lack of redundancy, the summary of photos should be an accurate summary of the whole collection, where all needs are satisfied. These needs can be different aspects of the photo collection.

Photo collection summaries need to contain beautiful photos that are relevant to the sentimental needs of a person. The summarization should remove redundancy, and should cover the different aspects of the photo collection.

The amount of photos used for a summary of the collection can not be pinpointed to one specific number or percentage of the full collection. It depends on several properties of the photo collection and the target use of the summary. For example, a photo collection with many redundant photos leads to less photos in

[1] `https://www.snapchat.com/`
[2] `https://www.instagram.com/`

a summary than a photo collection with no redundant photos. If many aspects are present in the photo collection, then the summary also requires more photos to cover all the aspects. Next to this, the use of the summary determines the amount of photos that are used.

Because of these reasons, the photo collection summarization task is formulated as a ranking task. This means that the summary of the whole collection is actually a permutation of all photos in the collection. The photos in the top of the ranking are most likely to be good candidates for a summary, while the photos in the bottom of the ranking are least likely.

Automatic summarization tasks have mainly been researched in a setting of text mining and natural language processing. The most common definition of summarization is the task: *Given a large document, how can the important points be conveyed in only a few sentences?* (Zhai & Massung, 2016). This is similar to the task of photo collection summarization where given a large photo collection, the important points have to be conveyed in only a few photos. The task of text summarization is usually categorized into *extractive summarization* and *abstractive summarization* (Zhai & Massung, 2016). Abstractive summaries are created by analyzing the text, and writing a completely new but shorter text of the contents. This is much the same as the way humans create summaries. Extractive summarization simply retrieves the most relevant elements, such as passages or sentences out of a larger text. Extractive summarization precisely aligns with photo collection summarization, where the most relevant photos need to be retrieved from the larger collection. Most work on extractive summarization uses well-known information retrieval based techniques (Zhai & Massung, 2016, p.692) to estimate the relevance of a sentence or passage to the whole text. Information retrieval deals with finding material of an unstructured nature that satisfies an information need from within a large collection (Manning, Raghavan, & Schütze, 2008). Selecting the most relevant photos out of a photo collection fits within this frame. The task of ranking photos for an extractive photo collection summary is therefore defined as an information retrieval problem.

As stated before, photo summaries need to contain relevant photos, but should also minimize redundancy and maximize the coverage over different aspects of the photo collection. In other words, diversity is required. Search engines, arguably the best examples of information retrieval, heavily deal with diversity as well. The search results need to avoid redundant contents, and should cover all potential information needs of the query. Because of the importance of search engines in modern life, diversity has become a well-researched topic in the field of information retrieval.

This abundance of research and ideas is leveraged in this work, by approaching photo collection summarization as a ranked information retrieval diversification problem.

Chapter 2

Background

This research project uses customer feedback for the task of photo collection summarization. An information retrieval approach is used for estimating the relevance and diversity of photos in a photo collection. This chapter describes the most important work and methods in the fields of photo collection summarization, learning to rank and diversification for information retrieval. First, related work on problems related to photo collection summarization is presented. After this, an introduction is given on neural networks and the varieties that are used in this project. An explanation of the most common group of methods for learning from user feedback in the context of information retrieval follows. After reviewing diversification in information retrieval, the chapter concludes with an overview of evaluation metrics that are relevant for the task of photo collection summarization.

2.1 Related work

In the context of photo collection research, many authors use a combination of aesthetics estimation with other features. Shen and Tian (2016) describe a method for photo collection summarization where photos are represented by multi-modal features including time, location, color texture and deep learning features reduced with PCA. These features are then clustered using a Gaussian Mixture Model and Expectation Maximization to create clusters of events. From these events, the key photos are selected using a ranking algorithm based on quality, representativeness and popularity. Wu et al. (2016) build on the work of Shen and Tian (2016), and extend the photo selection with music video generation for automated story telling. The authors define a method for selecting events to sample key photos from, in the form of event uniformity. These selected events are the 'events' that result from the clustering method by Shen and Tian (2016). Kim and Lee (2016) describe a photo collection summarization method where photos selected must be of high aesthetic quality, interesting and memorable. Aesthetic quality is measured by a rule based system based on color composition, brightness and the rule of thirds. Interestingness deals with excluding similar contents, where visually similar images in a short time range are removed from the selection. Memorableness is set to be proportional to the number of people in the photo and the number of photos taken at around the

same time of the photo under inspection.

Other authors take a different approach and consider a subset of the photo collection summarization problem. Chang, Yu, Wang, Ashley, and Finkelstein (2016) describe a method for triage of photos in a series. This is based on the observation that people often take a series of nearly redundant pictures to capture a moment or a scene. The authors do not create a method for absolute rating of photos, but rather focus on 'better' or 'worse' in the context of a photo series. A dataset of 15000 photos in photo series is published, where one image is the 'best' photo of that series. A Siamese neural network is used to select the best photo in the series. While this is not a full solution for photo collection summarization, it does handle photo selection for photo series, an important sub-problem of photo collection summarization.

Lidon et al. (2015) describe a system for the semantic summarization of egocentric photo stream events in the context of lifelogging systems. Given a set of events, the authors describe an information retrieval inspired optimization method. First of all, a convolutional neural network based informativeness estimator is proposed. Semantic relevance of images is estimated by their saliency, "objectness" and detected faces. The resulting ranked lists are fused in order to promote diversity, where fusion weights are optimized. Finally, the result list is re-ranked in order to promote novelty in the result set. This method is similar to the method presented in this work, as it is mainly inspired by advances in information retrieval.

Some authors add textual features to their method. The dataset of these authors is often based on photo collections from social media. Camargo and González (2016) present a method for automatically selecting a set of prototypical images from a large set of images given a query. Both textual and visual contents are combined in the same latent semantic space, which allows to find a subset of images from which the whole collection can be reconstructed. The quality of the summary is measured by its ability to reconstruct the whole set, along with its ability to represent semantic diversity. Samani and Moghaddam (2017) present a semantic knowledge-based approach for image collection summarization. Ontology based features are created to measure the amount of semantic information contained in each image. A semantic similarity graph was made based on this information. Summary images are then selected based on graph similarity. The quality of images is not taken into account.

2.2 Neural networks

Neural networks, and more generally deep learning, has become very popular in recent years, especially when considering image recognition. This project uses neural networks as its main machine learning modeling approach. While this work assumes basic knowledge about deep learning, a small recap is presented in this section. The book by Goodfellow, Bengio, and Courville (2016) provides more in-depth information about neural networks, convolutional neural networks and deep learning in general, if more information is desired.

2.2.1 Basic neural network

The term 'neural network' has its origins in its attempts to find a mathematical representation of biological systems (Bishop, 2006). While originally inspired by biological systems, neural networks are not models of these systems. Rather, they are a class of machine learning models that consist of weighted computation graphs. A neural network takes a set of *input features* and passes it through a series of *layers* in order to calculate a representation of the input features. Each *layer* consists of multiple neurons (or nodes), which all have an associated *weight* and an activation function. These weights are the trainable parameters of the model, and they are trained by iteratively optimizing a *loss function* using a gradient descent approach.

The basic neural network model is the following model, adapted from the work of Bishop (2006). First, M linear combinations of the input $x_1, \ldots, x_D$, called activations, are constructed following

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{2.1}$$

where $j = 1, \ldots M$, and the superscript (1) denotes that the parameters are in the first 'layer' of the network. The parameters $w_{ji}^{(1)}$ are the weights, and parameter $w_{j0}^{(1)}$ is a trainable bias parameter. These activations a_j are transformed using a differentiable nonlinear activation function h to give $z_j = h(a_j)$. The resulting z_j are the learned representations of the input features. Given a two layer network, these representations are once again linearly combined to produce the output unit activations

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{2.2}$$

An output activation function that maps the output activation to the target domain is used as the prediction value.

This simple neural network with one 'hidden layer', the layer that is neither the input or the output layer, is the basis for all neural network models. A graphical representation of one instance of this basic network is shown in figure 2.1. In principle, the number of layers, number of weights in each layer and activation functions in each layer are all up to the choice of the researcher creating the model.

2.2.2 Convolutional neural networks

Research on machine learning on images has been booming in recent years and this work is one of many examples. Bishop (2006) describes several properties of image recognition that create the baseline for the unique architecture of convolutional neural networks.

Given an example task of detecting whether an image contains a cat, several properties of image features (pixels) can be leveraged. First of all, the detection of cats is location invariant. If the task is to detect whether an image contains a cat, it does not matter where this cat is located in the image. It is even invariant to many other transformations such as scaling, small rotations, morphological

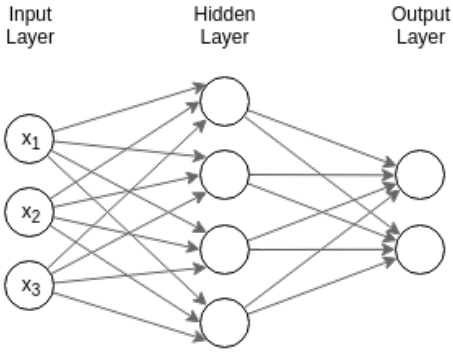


Figure 2.1: Schematic overview of a basic neural network with one 'hidden layer'

stretching etc. Next to invariance, images have the unique property that pixels are highly correlated. Even though it does not matter where in the image a cat is located, a cat always consists of many pixels which are closely related.

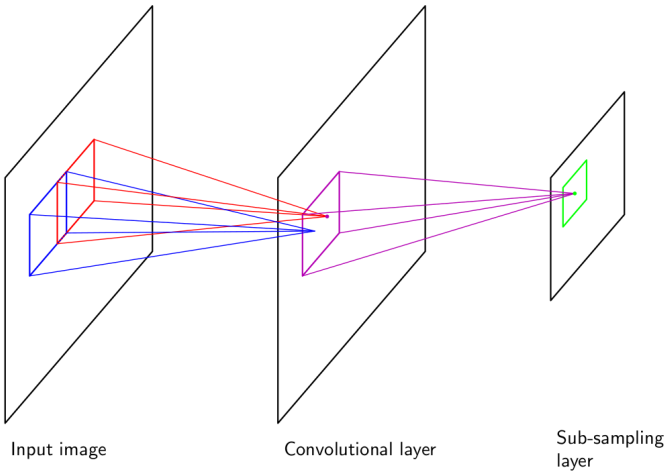


Figure 2.2: A part of a convolutional neural network

Convolutional neural networks use local receptive fields, weight sharing and sub-sampling in order to deal with this structural information and invariance. Figure 2.2, adapted from Bishop (2006), shows a convolutional layer and a sub-sampling layer, which are the key components of convolutional neural networks. A convolutional layer contains several planes, which are called *feature maps*. Each unit in a feature map takes its input as a linear combination of a small subregion of the original input image. All units in the same feature map are constrained to share the same weight values. This makes sure that a cat in the top left corner is handled the same as a cat in the top right corner. A sub-sampling layer takes the representations in the feature maps, and computes a downscaled version of this feature map using small receptive fields. The most important features in the feature maps are therefore aggregated over a region in the input image, which makes the network relatively insensitive to small transformations.

In practice, many of these convolutional layers and sub-sampling layers are stacked together, to create sequentially more abstract feature maps which are more invariant to input transformations compared to the previous layers. Finally, the image is compressed into a small sized feature map, after which 'normal' neural network layers are used to draw conclusions about the feature representations and therefore the original image. The number of layers makes the network 'deep', which is the origin of the term 'deep learning'.

2.2.3 Siamese neural networks

The idea of weight sharing can be leveraged in other problem domains as well. Bromley, Guyon, LeCun, Säckinger, and Shah (1994) describe a network architecture which they call Siamese neural networks. The network has two input fields to compare two patterns and one output whose state value corresponds to the similarity between the two patterns. Two seperate sub-networks act on each input pattern to extract representations of the input features, and a distance value is calculated. All weights in the network can be trained using back-propagation (Hecht-Nielsen, 1988), but the two sub-networks are constrained to have identical weights. This type of network architecture is usually chosen to learn a distance model between two inputs of the same type. Identical weights make sure the same representation is learned for both inputs, embedding each input into a low-dimensional space. These lower dimensional representations are then compared using a chosen distance measure. Bromley et al. (1994), for example, use the cosine of the angle between the two representations. The end-to-end model then describes a distance model between two inputs, which makes use of a feature representation step and a step that uses a manually chosen distance measure operating on these representations.

2.3 Learning to rank

Many information retrieval problems are ranking problems by nature (Liu, 2009). The typical examples of document retrieval, such as web search, can easily be formulated as ranking problems, where the documents that most likely satisfy the user's information need are assigned the highest position in the ranked list. The task of photo collection summarization can also be formulated as a ranking problem. If there is no restriction on the number of photos used in a summary, then the summarization task becomes a ranking problem. The goal is then to find a ranking where each summarization of length k gives the optimal summarization of the whole set with k photos.

Traditionally, many heuristic ranking models have been proposed and used in information retrieval literature (Liu, 2009). However, user interaction with search engines, recommendation engines and other information retrieval products caused enormous amounts of potential training data. This training data can be used by machine learning methods in order to make a prediction of the relevance, instead of just a heuristic approach. In general, all methods that use machine learning technologies to solve the problem of ranking can be called "learning-to-rank" methods (Liu, 2009). These methods all have in common that the relevance of a document to a query of a user is the prediction objective.

The term 'document' is used in the descriptions of the learning to rank methods, since the majority of research on learning to rank is focused on document retrieval. It is important to note that most ideas and methods are not specific to retrieving textual documents. The term is used to describe more than just documents with text, but contains any piece of content that a user is attempting to retrieve. This includes the 'retrieval' of photos out of a full photo collection for summarization.

2.3.1 Gathering labeled data

Gathering labeled data is typically done by one of two methods. The first method is based on explicit labeling of a (randomly) sampled dataset and the second on implicit evidence (Liu, 2009).

The first method takes randomly sampled queries and collects the documents associated with that query. After this, the relevance judgment for each document is assessed by human raters, which creates a set of explicit feedback. Relevance judgments are gathered using three different strategies. The first strategy asks human raters to assign a relevance score of each document to its associated query. These judgments can be binary relevance judgments, or more precisely specified relevance judgments such as a five-point Likert scale (Likert, 1932). The second strategy deals with pairwise comparison. A query and a pair of documents is presented to human raters, which choose the most relevant document out of the two. The third strategy asks the human raters to create a total ordering out of the documents associated with the query. These relevance judgments are closely related to the three learning to rank approaches described below.

The second method uses implicit feedback as relevance judgments. This data can be gathered from a bootstrapped system, where a baseline method presents retrieval results to the user and the user interacts with that retrieval system. The interactions can be analyzed to gather implicit user feedback. For a photo collection context, for example, one could argue that all photos that were included in the final photo product are relevant and all the others are not. From this implicit user feedback, relevance judgments can be extracted in the same three formats that were described for explicit relevance judgments.

Implicit feedback can be less accurate than explicit feedback ratings (Nichols, 1997). The main reason for this is that the data tends to be noisy and biased (Radlinski & Joachims, 2006). The task of learning to rank from implicit feedback is an interactive process between the user and the learning algorithm. The training data is therefore influenced by the results that are presented to the user, and the method used for presenting these documents. Explicit feedback, on the other hand, has the drawback that creating datasets is a tedious and time consuming effort, which makes gathering large amounts of data impractical.

This project therefore uses a combination of implicit user feedback and a small set of explicitly assessed photo collections.

2.3.2 Learning to rank approaches

Different approaches for learning to rank have been proposed. These can be broadly categorized into three categories, which all have different approaches of modeling the process of learning to rank. More specifically, the categories are

mostly distinguished by the modeling of the input and the output space. The explanations of these approaches are adapted from the work of (Liu, 2009).

Pointwise approach

In the pointwise approach of learning to rank, the input space contains the feature vector of each single document. The output space contains the relevance prediction of each single document. A trained model for predicting the relevance of a document takes the feature vector of a document as input, and predicts the relevance degree of that document. This relevance degree can be obtained for each document, and the documents can be sorted to produce the final ranked list.

One of the major disadvantages of the pointwise approach is the fact that a single score is predicted for each document given a query, without taking into account the other documents in the collection and the context of the collection. Ranking is more concerned with predicting relative order than a global relevance degree. For photo collection summarization, this means that a ranking function of a photo should not aim to predict whether a photo should be used in *any* photo product. Instead, the ranking function should predict whether a photo is better suited than another photo in the same collection.

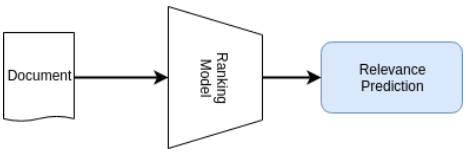


Figure 2.3: Pointwise learning to rank. The model gives a relevance prediction for each document.

Pairwise approach

The pairwise approach of learning to rank considers a pair of documents with their associated query vector as its input space. The output space then contains the pairwise preference between each pair of documents. A trained model gets two feature vectors as input, and predicts the pairwise preference degree $h(d_1, d_2)$; which of the two documents is more likely to be relevant.

The main advantage of the pairwise approach is the fact that it can model the relative order between documents, rather than a global relevance prediction. There are some problems with the pairwise approach, however. A major problem is the fact that every query (photo collection) has a quadratic number of pairs that can be evaluated. This means that the pairwise loss function which is used for training is dominated by queries with large numbers of document pairs. The consequence is that errors in smaller photo collections will be completely ignored by the system. One of the solutions is loss normalization on the photo collection level. This causes the loss of a pair from a large collection to have less impact, because the maximum loss stays consistent between all photo collections. Another possibility is to limit the amount of training data sampled from larger collections. This has the disadvantage that differences still occur, and that not all training data is leveraged.

The other problem of the pairwise comparisons, is that somehow a ranked list needs to be derived from the trained model. It is no longer possible to simply sort the input collection based on its relevance prediction. An optimal sorting algorithm can be used, which requires $\mathcal{O}(n \log n)$ comparisons. The pairwise preference can be used for sorting, since the pairwise preference produces a continuous value which represents the distance between two documents. This distance information is lost when ordering and choosing one of two documents as the 'more relevant' document. Any comparison sort algorithm will look at the pairwise distance, and just determine which of the two is the 'best' one. For example, if the differences between document d_1, d_2 and d_3 are $D(d_1, d_2) = 0.5$, $D(d_1, d_3) = 100$ and $D(d_2, d_3) = 99$, then a correct ordering would be $[d_1, d_2, d_3]$. The information that d_1 is 100 times better than d_3 is lost in this ordering.

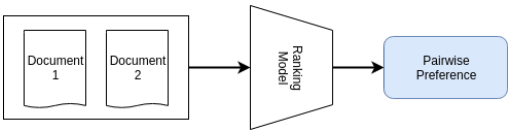
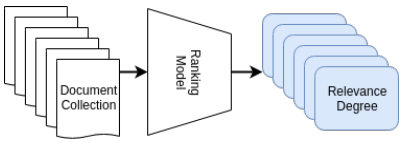


Figure 2.4: Pairwise learning to rank. The model gives a pairwise preference for each document pair.

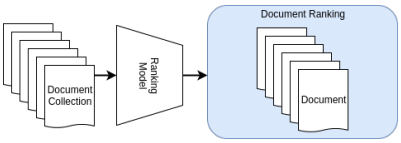
Listwise approach

The input space of the listwise approach contains the entire group of documents associated with one query. In the context of photo collection summarization, this means the whole photo collection. There are typically two types of output spaces used in the listwise approach. Some listwise ranking algorithms use individual relevance degrees of all documents associated with a query. Other algorithms use a permutation of the input space as the output space. This means that the ground truth label is not an indication for each document, but a target permutation of the collection.

The listwise approach has the great advantage that it directly fits on the problem of learning to produce rankings. The whole context of all documents in a collection is used, and there is no domination problem of larger collections. The main problem of the listwise approach is the complexity of the training procedure, as the evaluation of the loss function needs to consider the possible permutations of the result set. As Liu (2009) put it: "*a more efficient learning algorithm is needed to make the listwise approach more practical*".



(a) Model predicts individual relevance degrees for a full document collection.



(b) Model predicts a permutation of the document collection.

Figure 2.5: Two pairwise learning to rank approaches.

2.3.3 Applicability to photo collection summarization

Photo collection summarization in a broad sense is just the mapping of a photo collection to its optimal permutation, so that each summarization of length k is optimal. The theoretical advantages of the listwise approach seem to align neatly with the problem formalization. The practical implications, however, make an implementation in a production setting infeasible.

The pairwise approach is therefore pursued in this project. Chapter 4 will present an implementation of a pairwise learning-to-rank algorithm which also takes contextual features into account.

2.4 Diversification

Information retrieval techniques such as learning-to-rank rely on some modeling assumptions, which are generally overlooked. Quite some research has been dedicated to challenging these assumptions, out of which one aspect will be analyzed in this section. Diversity in information retrieval deals with resolving ambiguity in the information need and reducing redundancy in the retrieval results. The need for diversity in photo collection summarization as well as information retrieval in general is discussed first. Different strategies for increasing diversity are presented and the applicability of the methods to the photo collection summarization problem is discussed afterwards.

2.4.1 The need for diversity

Manning et al. (2008) describe information retrieval (IR) as finding material of an unstructured nature that satisfies an information need from within large collections. The most relevant part of this definition is the observation that information retrieval deals with *finding material* that *satisfies an information need*. Only a small translation is needed in order to convert this definition to one appropriate for photo collection summarization; photo collection summarization is concerned with *finding photos* that *are the most meaningful to the user* from within a large collection of photos.

The main challenge of an IR system is the determination of the *relevance* of an item given the query of the user. The goal then is to assess the relevance of a document, as happens in the learning to rank methods described before. Almost all information retrieval research is based around the probability ranking principle (Robertson, 1977). The probability ranking principle describes the relationship between the predicted relevance probabilities of documents and their related queries:

> *If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall of effectiveness of the system to its user will be the best that is obtainable on the basis of those data.* (Robertson, 1977)

Note that this principle does not prescribe how the probability of relevance should be estimated, which is the goal of all information retrieval methods.

Because of this broad definition the principle relies on two main modeling assumptions (Gordon & Lenk, 1991):

1. The probability of relevance is well-calibrated and estimated with *certainty*.

2. The probability of relevance of a document is estimated *independently* of other retrieved documents.

The first assumption only holds when there is no **ambiguity** in the information need of the user. In the context of document retrieval, non-ambiguity is defined as having a single query with exactly one meaning. There is no 'query' concept in the context of photo collection summarization. However, some idea of an information need underlying a 'query' can be determined; a photo product in some sense is the answer to a future information need. When a user manually creates a summarization for a photo product, it is acting as an information retrieval system that accurately tries to answer their own future information need. Automatic photo collection summarization needs to perform this same task, but in contrast to the user, the system knows little about the future information need. This makes the photo collection summarization request very ambiguous.

The second assumption relies on an absence of **redundancy** among the retrieved results. This is certainly not the case in the task of photo collection summarization. Digital photo collections often contain replicated photos. And even if this is not the case, photographers rarely take only one shot to capture a scene or moment.

2.4.2 Diversification strategies

Previously, the notions of ambiguity and redundancy were introduced. There are different strategies to combat these unwanted effects of relevance-oriented ranking. In general, ambiguity can be tackled by ensuring a high **coverage** of the possible information needs underlying the query among the retrieved documents. Redundancy can be tackled by ensuring that retrieved documents provide a high **novelty** with respect to their covered needs (Santos, Macdonald, & Ounis, 2015).

Figure 2.6, adapted from (Santos et al., 2015), shows that the goals of attaining maximum coverage and maximum novelty are often conflicting. In this figure, η_1, η_2 and η_3 convey the needs of the user, where A, B, and C describe the relevance of the documents to those needs. In some cases, a relevance oriented model only marks documents satisfying one specific need as relevant, as happens in the first column in the figure. In this case, it is important that all needs are **covered**. The second column shows how maximum coverage can be obtained, all needs are covered in the top six of the ranking. This however, does not automatically mean a maximum **novelty** with respect to already selected photos. The third column shows a ranking with maximum novelty, but this does not fully cover all needs. A diversity-oriented ranking, therefore needs to balance out relevance, coverage and novelty.

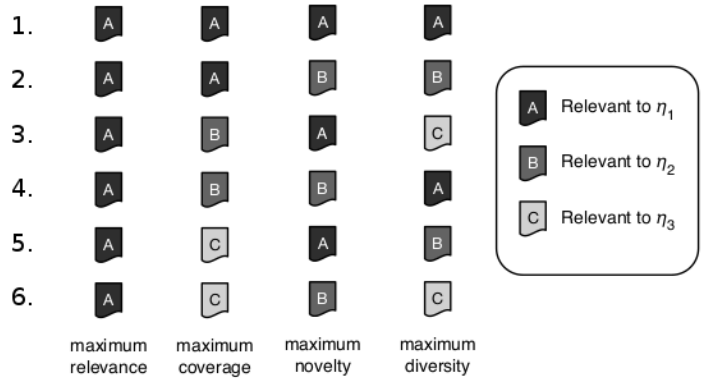


Figure 2.6: Example ranking where relevance, novelty and coverage all conflict

Diversification strategies continued

Santos et al. (2015) define a two-dimensional taxonomy of approaches for search result diversification. The first scale divides algorithms based on the way aspects are represented. Aspect representation determines how the information needs underlying a query are represented as multiple aspects. *Implicit* aspect representation relies on features belonging to each document in order to model different aspects of diversity. *Explicit* aspect representation, on the other hand, seeks to directly approximate the possible information needs underlying a query by relying on features derived from the query itself, such as query categories or reformulations.

Given any representation of aspects, the diversification strategy then determines how to optimally satisfy all aspects underlying a query. *Novelty* based approaches directly compare the documents to one another, without directly taking query aspects into account. Novelty based approaches focus on avoiding redundancy in the search results. *Coverage*-based approaches try to estimate how well each document covers the multiple aspects of the query. This does require some indication of aspects. Coverage based strategies try to resolve query ambiguity by making sure every aspect is covered in the top of the ranking. *Hybrid* approaches try to tackle both ambiguity and redundancy, by balancing them in one approach.

Since there are no queries available for photo collection summarization, explicit aspect representation does not seem applicable. Instead, if aspects were to be determined, they would have to be determined from the photos themselves, which is a form of implicit aspect representation. There are no theoretical restrictions on diversification strategies. Novelty approaches make sense for summarization, since a customer would want no (near) duplicates in their photo product. However, coverage-based or hybrid approaches are also very important. The task of photo collection summarization is defined as a ranking problem where at each sublist of length k, the summarization should be optimal. Given the example of a holiday with several days at the beach, and just one day in a city, then these days could be modeled as information needs (this will be approximated in chapter 3). At every rank k, the photos up until k should cover as many days as possible, and not just cover the beach days.

2.5 Evaluation Metrics

This project deals with the application and combination of methods that are well tested in theory, but give no guarantees about practical performance on a real world dataset. Next to this, in the selection and combination of methods, many assumptions and choices are made. It is therefore very important to evaluate the performance of the algorithms on a real world dataset, and to evaluate whether the choices, that might make sense conceptually, work in practice.

This section therefore gives an introduction to the most commonly used evaluation metrics for (binary) classification, learning to rank and diversification in information retrieval. The classification metrics are presented firstly, as they are the most intuitive and easy to understand. After this, metrics that take positions in a ranked list into account are presented. Finally, the possibility of using specific methods to assess diversification is discussed.

2.5.1 Evaluation metrics for direct relevance prediction

The first set of metrics are derived when reducing photo collection summarization from a ranking problem into a simple binary classification problem. A relevance model $f(\cdot)$ predicts whether each photo p_i is selected in the summarization or not. This relevance model predicts "true" or "false" for the statement "this photo should be selected in the summarization of the photo collection". The implicit relevance judgments lead to labels that describe whether a photo should actually be included in the result set. The relevance predictions lead to a set of **selected** photos, which is a subset of the whole collection. The true relevance judgments lead to a set of **relevant** photos.

The most basic metrics that describe the performance of a machine learning system are *precision* and *recall* (Manning et al., 2008). These metrics can be derived from the *confusion matrix* of the retrieval results. Table 2.1 describes the standard confusion matrix for information retrieval in the context of photo collection summarization. Each photo can either be "selected" or "not selected" by the relevance model. The true label of each photo can be "relevant" or "not relevant". Each photo can therefore be described by one of the cells in the confusion matrix, which gives a set of true positives, a set of false positives, a set of false negatives and a set of true negatives.

		predicted	
		Selected	**Not Selected**
actual	**Relevant**	True Positive	False Negative
	Not Relevant	False Positive	True Negative

Table 2.1: Confusion matrix for the task of photo collection summarization

Precision is defined as the ratio of selected items that are also relevant. In terms of the confusion matrix, *precision* can be calculated following equation 2.3. Precision says something about the accuracy of the selected photos and does not contain information about the photos that were not selected but should have been (the photos that are "not selected" and "relevant"). *Recall* is the metric

that describes the ratio of selected items to the relevant ones. It therefore says something about the number of photos that were not selected, instead of something about the accuracy of the selection. Recall is calculated with equation 2.4.

$$\text{precision} = \frac{|\text{True Positive}|}{|\text{True Positive} \cup \text{False Positive}|} \tag{2.3}$$

$$\text{recall} = \frac{|\text{True Positive}|}{|\text{True Positive} \cup \text{False Negative}|} \tag{2.4}$$

Precision and recall are two metrics that can be used to give a score to the quality of a photo collection summary. However, they are two different metrics that usually imply a trade-off. It is useful to capture the information of both these metrics in one metric. One such metric is the F_1 score, which is a harmonic mean between the two metrics. The metric penalizes summarization results with high precision but low recall, and vice-versa. The F_1 score can be calculated with equation 2.5 (Manning et al., 2008).

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.5}$$

Soft relevance prediction

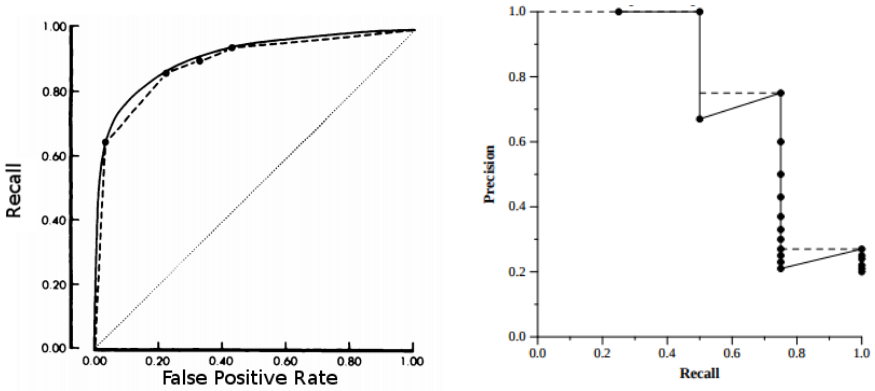
The above metrics share the assumption that the relevance model predicts either "selected" or "not selected". In reality, however, the relevance model will predict a relevance score between 0 and 1. A cutoff value c between 0 and 1 needs to be decided in order to select the photos that should be included in the summary. Any photo with a relevance score $\geq c$ should then be included in the summary. Changing this threshold often results in a trade-off between precision and recall. Lowering the threshold results in more documents being selected, which can reduce precision and increase recall. Increasing the threshold however, often results in a lower recall and a higher precision.

Because of this trade-off, it is useful to look at the retrieval results using measures that aggregate over all possible thresholds. Some of these measures provide a two-dimensional plot, which makes an overview of the precision-recall trade-off visible at a glance. The Receiver Operating Characteristic (ROC)-curve is most commonly used when inspecting the result of a binary classification task.

The Receiver Operating Characteristic curve is plotted by changing the threshold c and plotting the probability of detection and the probability of 'false alarm'. The probability of detection in this case is the *recall* of the result, while the probability of false alarm is the false positive rate, which is calculated by equation 2.6. An example ROC curve, adapted from Hanley and McNeil (1982), is displayed in figure 2.7a. The diagonal line from the bottom left to the top right depicts the ROC curve for a model that makes a random guess.

$$\text{false positive rate} = \frac{|\text{False Positive}|}{|\text{False Positive} \cup \text{True Negative}|} \tag{2.6}$$

The area under the ROC-curve is referred to as the AUC value (Area Under the Curve). This value gives a single score to the performance of an information retrieval system with soft relevance predictions. The AUC value can be interpreted as the probability that given a randomly sampled relevant photo and a non-relevant photo, the relevant photo is assigned a higher relevance prediction (Hanley & McNeil, 1982).



(a) Example ROC curve (b) Example Precision-Recall curve

Figure 2.7: Example plots of threshold based measures

2.5.2 Evaluation metrics for ranking

Photo collection summarization is formulated as a ranking problem in this work, instead of a classification problem. Metrics for binary classification typically do not give any information about the quality of the ranked list. Since the position in the ranked list determines whether a photo is included in a photo product with k photos, the evaluation metrics used should provide insight into this ranking.

The first metric that does give insight into the quality of the ranked list is the precision-recall curve. This curve-based metric is similar to the Receiver Operating Characteristic curve. The plot is created by starting at rank 1, and plotting each precision-recall value up until the whole collection is processed. An example plot is given in figure 2.7b, adapted from the Text REtrieval Conference (2015). This plot gives a simple overview of the trade-off between precision and recall. The single score that is calculated from this plot is the Average Precision (AP) metric, which is the area under the precision-recall curve. The area under the curve can easily be approximated following equation 2.7. Here k is the ranked position, $P(k)$ the precision at rank k and $\triangle r(k)$ the change in recall between position $k-1$ and k.

$$\text{AP} = \sum_{k=1}^{n} P(k) \triangle r(k) \tag{2.7}$$

The major difference between the precision-recall method and the ROC evaluation method is the fact that the ROC curve can be constructed over all photo collections, as the threshold c is independent of the collection. This means

that the ROC curve together with the AUC score is a global indication of the performance of a system. In contrast, the precision-recall curve is determined for each individual photo collection. In order to compare the results of different photo collection summarization methods, the results need to be aggregated over all collections. The traditional way of doing this is by using a simple 11-point interpolated average precision value which can be calculated with equation 2.8 (Manning et al., 2008). The interpolated precision value is calculated by $p_{\text{interp}}(r)$, which is the precision value for any recall level $\geq r$. In order to compare the AP values over all photo collections, the mean average precision (MAP) is simply calculated as the arithmetic mean of all AP values over all photo collections. Calculated MAP scores can vary widely across information needs, so a single score over all photo collections does not necessarily provide correct information. Another problem is the fact that photo collections all have a different number of photos in their collection, and that the number of relevant photos varies widely. In order to compare these different collection sizes fairly, the MAP values should be calculated for different ranges of collection sizes.

$$\text{AP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1.0\}} p_{\text{interp}}(r) \tag{2.8}$$

A simple alternative measure, called R-precision can be used when a set of known relevant photos is available (Manning et al., 2008). The measure depends on the number of relevant photos $\kappa = |\text{True Positives} \cup \text{False Negatives}|$ for a photo collection. The top κ photos in the predicted ranking are taken into account. The R-precision value is equal to the precision at κ, the precision of the top κ photos. An interesting fact of this measure is that given the top κ, the precision and recall are equivalent. Despite the fact that the R-precision metric only describes one point on the precision-recall curve, it is highly correlated with MAP empirically speaking (Manning et al., 2008, p. 161). The advantage of the R-precision metric is that by nature of the measure, the impact of different numbers of relevant photos for each photo collection is minimal.

2.5.3 Evaluation metrics for diversification

Diversification is an important part of photo collection summarization. It is therefore important that the diversification efforts described in this project are properly evaluated. Several evaluation metrics have been proposed in the search result diversification literature. Santos et al. (2015) describes the most relevant metrics for the document retrieval domain. These metrics quantify the extent to which the top κ documents in a ranking R_q cover the aspects A_q representing the information needs N_q underlying a query q, given a cutoff κ.

In contrast to document retrieval, there are no datasets available which explicitly model the aspects A_q for the task of photo collection summarization. This dataset can also not be created easily, since the information needs N_q are unknown. These are even unknown to customers, since photo products are created from a photo collection in order to answer a future, yet unknown information need. These evaluation metrics are therefore not usable in the task of photo collection summarization.

In order to cope with these difficulties, the approach taken in this work is to simply ignore all diversification evaluation metrics. Instead, the assumption

is made that the probability ranking principle is not sufficient for photo collection summarization, given the redundant and ambiguous nature of the task. Therefore, any increase in diversification should **improve** the performance on the 'standard' information retrieval metrics described above.

Chapter 3

Diversity estimation

Chapter 2 introduced the need for diversity, along with the concepts of *redundancy* and *ambiguity*. It was noted that *redundancy* deals with similarity between information retrieval results. *Ambiguity* says something about the possible information needs underlying the request of the user. This chapter describes methods for estimating the redundancy and ambiguity in photo collections.

3.1 Estimating redundancy

Photo collections contain many types of redundancy, which all have different causes. One of these redundancies is the fact that the people pictured in photo collections are often the same people. This kind of redundancy, however, is not something which needs to be solved. Two sources of redundancy that are important to resolve are estimated in this section, which are duplicates and photo series.

3.1.1 Duplicate detection

The simplest method for decreasing redundancy in the ranking is to remove duplicate content. In the case of photo collection summarization, this means that duplicate photos and slight edits of these photos would have to be filtered out. All photos with duplicate content are removed from the set as a first step in the complete pipeline of photo collection summarization.

For this purpose, the concept of perceptual hashing has been implemented. Hadmi, Puech, Said, and Ouahman (2012) describe perceptual image hashing functions that extract certain features from images and calculate a hash value based on these features. The functions should establish the "perceptual equality" of image content. Comparing images is then performed by comparing the hash values of the images. These perceptual hashes are expected to be able to stay consistent on acceptable content-preserving manipulations and change on content-changing manipulations. A list of content-preserving and content-changing manipulations is given by (Han, Chu, & Yang, 2007).

3.1.2 Photo series detection

The preprocessing step makes sure that duplicate or very near-duplicate photos are not present in the dataset. This causes an increase in the novelty of the result set.

A different level of redundancy in the photo collections of the user exists. This coincides with the observation that people often take a series of nearly duplicate pictures to capture a moment or a scene (Chang et al., 2016). These 'photo series' consist of pictures of the same object, scene or moment where there are slight differences in camera parameters and content arrangement. People often take these series in order to pick out the best ones to edit, post to social media, share with others, or include in a photo album. A good photo collection summarization method therefore does this job for the user. These photo collections should be detected and the only best photo(s) should be included in the top positions of the resulting photo ranking.

Chang et al. (2016) have created a dataset of photo series for training models to select the best photos out of a series. Photo series are identified by first discarding redundant shots, for example shots captured in a burst session. To find photo series from non-duplicate photos, SIFT descriptors are extracted and scene correspondences between neighboring photos are calculated. Photo pairs with good scene matching are grouped together into a series. Besides SIFT feature matching, color histograms are analyzed to handle special cases such as one of the two photos having severe camera motion blur. In the definition of Chang et al. (2016), each photo series should contain the same group of people. Face verification is used to split photo series with different groups. Finally, series with more than 8 photos are not allowed. A variant of k-means is used to split up photo series with more than eight photos.

The method for photo series detection in this project is conceptually simpler. Through perceptual hashing, duplicate shots are removed from the set as a whole. A method designed for assessment of visual similarity of neighboring time-ordered photos is used to create a similarity score for all consecutive image pairs. The neighboring photo similarity scores are evaluated iteratively in order to create the set of photo series. When the similarity between the current photo and the next photo is high enough, the photo is added to the current photo series. If the photos are not similar enough, a new photo series is started with the next photo in the list. In Chang et al. (2016), face verification is used to split up photo series with more than one subject group and photo series with 8 photos or more are split up. These restrictions are not applied on the photo series detection method described in this section.

3.2 Estimating ambiguity

The creation of a photo product can be interpreted as an answer to a future, unknown, information need. Generally, this future information need has to do with remembering something: an event, a person or anything else that is special for the person creating the photo product.

What the user wants to remember, however, is not clearly defined. It is therefore important that all possible information needs are estimated. These information needs are usually embedded as *aspects* in the underlying query in

traditional information retrieval. In the case of photo collection summarization, these aspects are embedded in the photo collection itself.

While there are many different perspectives of ambiguity, one specific aspect is analyzed in depth. Many photo collections are depictions of some real life event, such as a wedding or a holiday. The aspects of these photo collections are then the different temporal events in the larger life event. A wedding, for example, consists of the ceremony and the party. A method is presented that is able to segment a photo collection in its temporal events, based on temporal and visual data.

3.2.1 Temporal event clustering

Temporal event clusters are estimated using an adaptation of the method presented by Cooper, Foote, Girgensohn, and Wilcox (2005).

Similarity matrix embedding

Timestamps in minutes are extracted from the photo collection under inspection. The N photos in the collection are ordered by ascending timestamps. A family of K $N \times N$ similarity matrices S_k is computed over the whole photo collection. Each element (i, j) in S_k is calculated following equation 3.1. In this equation, t_i and t_j are the timestamps in minutes of photos i and j respectively. Parameter k controls the sensitivity of the similarity measure in minutes. Increasing the k parameter results in a coarser similarity estimate between photos i and j. The similarity function therefore only looks at the frequency of taking photos.

$$S_k(i, j) = \exp\left(-\frac{|t_i - t_j|}{k}\right) \tag{3.1}$$

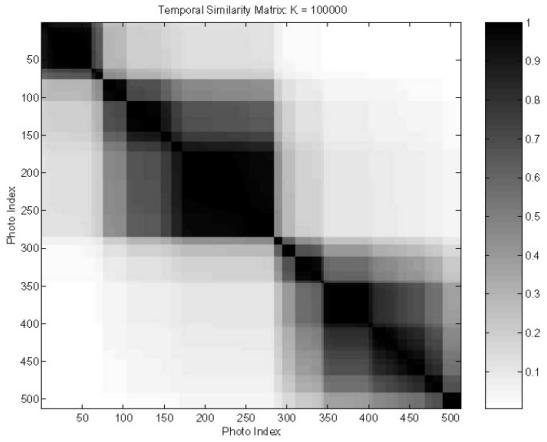


Figure 3.1: Example similarity matrix for $k = 10^5$ and 500 photos.

Computing novelty scores

As figure 3.1 shows, the event clusters are distinguishable as square blocks in the main diagonal. The boundaries between event clusters are visible as the

centers of the checkerboard patterns along the main diagonal. A photo-indexed novelty score is calculated using a matched filter approach. A Gaussian-tapered checkerboard kernel g is correlated along the main diagonal of each S_k to calculate the novelty score. Following (Cooper et al., 2005), the filter has a size of 12×12. This means that different time scales are handled through the differing values of K, instead of the filter size.

Initial event boundary selection

The resulting photo-indexed list of novelty scores $\mathcal{V}_k$ is analyzed to create a set of cluster boundary candidates for each $k \in K$. Peaks in the novelty scores are detected by calculating the first difference of each $\mathcal{V}_k$. This first difference is used to calculate the increase in novelty score between the photo and its left neighbor in the time-ordered list, and the decrease in novelty score between the photo and its right neighbor. The increase and decrease in novelty scores are summed, which gives a photo-indexed list of 'peakiness' scores for each $k \in K$. The event boundary candidates are selected by choosing all boundaries with a 'peakiness' greater than a threshold t. This threshold is a tunable fraction of the maximum possible novelty score, which is determined by the similarity measure and the checkerboard kernel correlated along the diagonal of the similarity matrix. Boundaries detected at coarse time scales (a high k') are included in all finer scales $k'' < k'$, which gives a hierarchical set of event boundaries $\mathcal{B}$. The creation of this set is depicted in figure 3.2.

BIC-based event boundary pruning

The method chosen for event boundary pruning is based on the Bayes information criterion (Schwarz, 1978). The method requires a simplifying assumption that timestamps within an event are distributed normally around the event mean. Cooper et al. (2005) remark that this is difficult to justify empirically, which in practice means that this assumption will reduce the accuracy of the event clustering. For each boundary $b_l \in \mathcal{B}$, the effect of the boundary is tested to determine if the increase in model likelihood justifies the split in events caused by b_l.

$$L(b_{l-1}, b_l) + L(b_l, b_{l+1}) \gtrless L(b_{l-1}, b_{l+1}) + \log(b_{l+1} - b_{l-1}) \tag{3.2}$$

$$L(b_l, b_{l+1}) = -\frac{b_{l+1} - b_l}{2} \left(1 + \log(2\pi\hat{\sigma}_l)\right) \tag{3.3}$$

The left hand side of equation 3.2 equals the log-likelihood of the model where b_l splits the event cluster and the right hand side is the log-likelihood of the single event cluster model with the penalty term of the additional parameters of the model with two events. If the likelihood gain associated with splitting the clusters is greater than the penalty of the additional cluster, boundary b_l is used in the final event clustering.

Visual similarity based cluster merging

Since a method for inspecting visual similarity between subsequent image has been implemented, this information can be leveraged. Cooper et al. (2005)

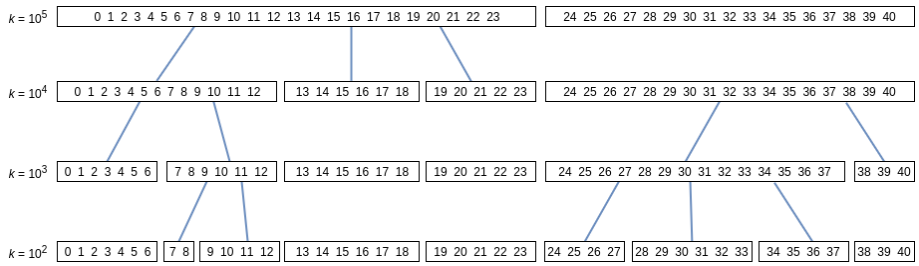


Figure 3.2: Hierarchical set of event boundaries $\mathcal{B}$. Boundaries gathered from coarse to fine time scales lead to a hierarchical set of cluster candidates.

provide a method for combining time and content-based similarity. This method is not followed for two reasons. First of all, content-based similarity is included in other parts of the photo collection summarization method. More importantly, the method provided by Cooper et al. (2005) assumes that a content-based similarity matrix with entries for all photo pairs is available. As explained in subsection 3.1.2, the visual similarity comparisons in this project are restricted to all subsequent pairs of photos. This means that the content-based similarity measure is only able to change the diagonal to the left and to the right of the main diagonal. This has minimal effect on the 12×12 kernel correlation. Furthermore, the decision whether to use the content-based similarity measure is based on a fixed amount of time (48 hours). As the photo collections in this project are very diverse, this constant needs to be determined from the photo collection under inspection.

For these reasons, the temporal event clustering method has been adapted by introducing a different approach for the inclusion of visual similarity. Instead of directly including the visual similarity in the similarity matrix, it is used in a post-processing step. The threshold in the initial boundary selection is lowered to artificially introduce more candidate cluster boundaries. Because the number of clusters is now over-estimated, there might be some cluster boundaries that should not exist.

Only clusters with sufficient time proximity are considered for merging. A data-independent approach is taken for selecting the threshold that implements 'sufficient time proximity'.

Two temporal event clusters are merged when they are allowed to be merged according to the time threshold, and when their visual similarity is 'high'. The visual similarity is considered 'high' when they belong to the same photo series according to the method presented in section 3.1.2.

Chapter 4

Relevance estimation

A ranked list of photos for photo collection summarization needs to be more than just diverse. The photos in the top ranks need to be relevant to the assumed future needs of the user. This work presents the hypothesis that this relevance can be learned from the data itself using a learning to rank mechanism.

One possible way to do this is to train a classifier on just the pixels of the photos. While advances in deep learning should make this possible, it is still a very difficult task; especially when taking into account the context of the photo album as a whole and the preferences of the user. Because of this difficulty, high level features are extracted with different machine learning models. These high level feature extractors are constructed to predict several aspects of relevance. Next to these high level features, the context of the photo is included. After all, the relevance of a photo depends on the whole photo collection and on the preferences of the user.

After describing the high level feature analysis and the context analysis, the model used for learning to rank is described in detail.

4.1 High level photo analysis

Lidon et al. (2015) explicitly model the summarization task as the query "select T images to describe the event depicted in these N frames". Ambiguity in this query is resolved by modeling this query from different perspectives. The perspectives '*Where* is the user?', '*What* activity is the user performing?' and '*With whom* is the user interacting?' are considered. In their work, the relevance of each image with respect to these questions is estimated by using high level features extracted from convolutional neural networks.

Analogously, the summarization query of photos for a photo product such as photo prints can be modeled from different perspectives. Three different query perspectives are presented below. This approach resembles the approach by Lidon et al. (2015). The major difference is the fact that the queries in this work are formulated as selection queries, rather than queries about the content of the photo. The queries of Lidon et al. (2015) try to estimate potential sources of diversity, while the queries presented shortly try to estimate potential sources of relevancy. This better aligns with the fact that the resulting features will be used to model relevance, rather than diversity. Increasing diversity is tackled in

a different part of the pipeline.

What are the most beautiful photos of this collection? Photo products need to contain photos of high aesthetic quality, as users want to remember their life events in the best possible way. Users generally tend to choose photos with high quality as key photos (Shen & Tian, 2016), and select photos that are aesthetically pleasing (Kim & Lee, 2016). High level features describing the aesthetic quality of a photo are extracted.

Which photos are the most natural? In 2007, the year smartphones were introduced to the mass consumer market, there immediately was a huge increase in consumer photography. Decreasing storage costs together with innovations in capture devices have made it exceedingly simple for people to capture photos, which they did at an ever-increasing rate (Ames & Naaman, 2007). The introduction of smartphones and cloud-based storage has brought a whole new range of possibilities, which has not slowed down this increase in photography.

The introduction of photography on mobile devices has brought its challenges, however. Many non-photograph types of images get mixed up in the image collections of the user. Images such as screenshots, memes and inspirational quotes are not a rare sight in online photo collections. The 'naturalness' of all images in the photo collection is therefore assessed with high level features describing this 'naturalness'.

Which photos contain people important to the user? Important moments in a person's live are rarely experienced alone. Photographs taken to remember these moments often contain those people important to the photographer. A good summarization of a photo collection should therefore contain these important people. The presence of import people in photos of the collection is therefore assessed, resulting in high level features describing the presence of import people in a photo.

4.2 Context analysis

The photo level analysis results in a set of features that should capture the relevance of each photo to a potential summary of the photo collection. This does not take into account the context of the photos. The context in this case consists of a set of features of the photo collection and a set of features of the user.

Photo collection features Global features of the photo collection contain a lot of information on the type of the photo album. One part of the feature set consists of simple aggregates of photo based features. These aggregates contain quite some information about the type of pictures the photographer usually takes. This makes it possible to compare the photo-specific features to an 'average' photo of the photographer.

Photo series features Besides the photo collection as a whole, the photo series the photo under inspection resides in could provide some information about the relevance of the photo. High level features describing the characteristics of the photo series in a collection are extracted.

User based features Different people have different tastes and wishes for their photo products. It therefore makes sense to include features describing the user in the feature set of the context.

Local neighborhood features Besides aggregates on photo series and photo collection level, features can be extracted on a local neighborhood. The local neighborhood is assessed by taking the photos' direct neighbors in the time-ordered photo collection. The same features as the photo under inspection are extracted and appended to the features of the photo. If diversity or feature differences with regards to time-ordered neighbors are important, then the model should be able to learn from these features.

4.3 Learning to rank photos

As explained in chapter 2, pairwise learning to rank has many theoretical advantages compared to pointwise learning to rank, as it fits more to the problem of deciding on relative importance. The method chosen is therefore a pairwise learning to rank model, with some adaptations that will be described shortly.

4.3.1 Dataset creation

Since the learning to rank method implemented is a pairwise learning to rank model, the photo collections need to be processed in such a way that pairs can be extracted.

Each photo collection n is a set of photos N_n, where a set of photos $S_n \subseteq N_n$ is selected in a final photo product, further referred to as the set of 'relevant photos'. This leaves a set of photos $T_n = N_n \setminus S_n$ that is not selected in the photo product, further referred to as the set of 'non-relevant photos'. Because any pair is plausible, each photo collection can lead to a maximum of $|S_n \times T_n|$ photo pairs.

The photo series are an important part of this dataset creation method. With the set of relevant photos S_n, all photo series that have at least one relevant photo are selected. These photo series are considered the 'relevant photo series'. From each of these relevant photo series, the relevant photo is selected. If there happens to be more than one relevant photo in the relevant photo series, then one of the relevant photos is chosen at random. These operations lead to a selection of relevant samples $\mathrm{Rel}_n \subseteq S_n$. Non-relevant photo series are those photo series that contain only non-relevant photos. A random non-relevant photo is selected from each of these non-relevant photo series, leading to a selection of non-relevant samples $\mathrm{Nrel}_n \subseteq T_n$.

From these non-relevant photos Nrel_n and relevant photos Rel_n a collection-level dataset $\Phi_n = \mathrm{Nrel}_n \times \mathrm{Rel}_n$ is created. Finally, the full dataset Φ is created by simply concatenating all Φ_n. The target label set $\mathcal{Y}$ then coincides with the first or the second photo being the relevant photo.

The result of this dataset creation method is the fact that no non-relevant photos of relevant photo series are included in the training set. The relevancy of these photos is less certain than the relevancy of the other photos that are used in the training set. A non-relevant photo of a photo series can be left out of the photo product because it is worse than the other photos in the collection, or because it is just slightly worse than the relevant photos of the photo series. In this last case, the photo could very well be relevant if the other photos in the photo series are not present. The downside of this restriction is the fact that not all cases are represented in the training set.

The creation method of a pairwise learning to rank dataset from photo collections is based on the presence of photo series. According to Chang et al. (2016), photo series triage is an important part of photo collection summarization. It is, however, a different task from the selection of photos for the summary. The pairwise ranking model trained has the task to rank photos regardless of their photo series. Because of this, it is important that it does not accidentally learn to rank photos in a photo series, instead of between photo series. In order to force this, only one photo out of each photo series is allowed in any pairwise training set. This is different from the restriction in the previous paragraph, which stated that only relevant photos of relevant photo series are allowed in the training set. The downside to this is the fact that same model is used for photo selection within photo series, thus for photo series triage.

4.3.2 RankNet

The model chosen for the learning to rank problem is the classic RankNet model (Burges, 2010). In the context of web retrieval, RankNet is defined as follows: For a given query, each pair of urls U_i and U_j with differing labels is chosen. Each url pair with feature vectors x_i and x_j is presented to the RankNet model, which computes the scores $s_i = f(x_i)$ and $s_j = f(x_j)$. Let $U_i \rhd U_j$ be the notation for the event that U_i should be ranked higher than U_j. The two outputs of the model are mapped to a learned probability P_{ij} of the fact that U_i should be ranked higher than U_j through the sigmoid function described in equation 4.1. The cost function chosen in RankNet is the well-known cross entropy cost function. Given the predicted probability P_{ij} and the true probability $\bar{P}_{ij}$, cost C can be calculated with equation 4.2.

$$P_{ij} \equiv P(U_i \rhd U_j) \equiv \frac{1}{1 + e^{-(s_i - s_j)}} = \sigma(s_i - s_j) \tag{4.1}$$

$$C = -\bar{P}_{ij} \log P_{ij} - (1 - \bar{P}_{ij}) \log(1 - P_{ij}) \tag{4.2}$$

This definition can easily be translated to the photo ranking problem. Each pair of photos $(\phi_i, \phi_j) \in \Phi$ is chosen. The dataset creation method forces that the labels differ. Each photo pair with feature vectors x_i and x_j is then presented to the model, which computes $s_i = f(x_i)$ and $s_j = f(x_j)$. Like before, $\phi_i \rhd \phi_j$ denotes the event that ϕ_i should be ranked higher than ϕ_j. The predicted probability from equation 4.1 now calculates $P(\phi_i \rhd \phi_j)$.

4.3.3 RankNet implementation

The model for RankNet can be any model of which the output is a differentiable function of the model parameters (Burges, 2010). A neural network model has therefore been constructed. The goal of the network is to predict the scores s_i and s_j, after which the sigmoid loss function can be calculated following equation 4.1.

Chapter 5

Balancing relevance and diversity

Chapter 4 describes the methods which are used to estimate the relevance of each photo in a collection. A high level photo analysis is performed to extract useful features and a context analysis is performed to collect features about the photo collection, the local neighborhood of photos and features of the user. These features are used to train a model which predicts the relevance of a photo in a photo collection, compared to the other photos in the collection.

As explained in 2.4, purely using relevance is based on two main assumptions. These assumptions hold when there is no ambiguity in the information need of the user, and there is no redundancy among the retrieved results. Both these assumptions do not hold in the task of photo collection summarization. The ambiguity in the information need of the user can be tackled by increasing coverage over all possible information needs, while the redundancy can be combated by increasing the novelty in the result set.

For this purpose, three perspectives of diversity are estimated using the methods described in chapter 3. These perspectives are duplicate detection, photo series detection and temporal event cluster detection.

5.1 Increasing novelty

Chapter 2 introduced several categories of methods for increasing diversity. One category of strategies seeks to promote novelty as the implicit diversification strategy. This means that the aim of the strategies is to infer differences between results in order to demote those with redundant contents.

The detected photo series can be used for these goals. Because photo series consist of pictures of the same object, scene or moment where there are slight differences in camera parameters and content arrangement, these photo series contain redundant contents by definition. With this interpretation of redundancy and the interpretation of relevance determined by chapter 4, implicit novelty-based diversification algorithms can be used.

The algorithms that are presented both have the shared requirement of a similarity function Sim that measures the similarity between two photos. Instead of a direct similarity function, photo series are available, out of which the

similarity function is derived.

Two algorithms are evaluated. These algorithms are the Maximum Marginal Relevance reranking algorithm (Carbonell & Goldstein, 1998) and the Quantum Probability Ranking Principle reranking algorithm (Zuccon & Azzopardi, 2010). Both algorithms follow the same structure, where the next best photo is picked in an iterative manner using an optimization criterion $\mathbb{O}$. Pseudo-code for this novelty-oriented re-ranking algorithm is given in algorithm 1. The methods that are described in the next sections differ in the choice of optimization criterion $\mathbb{O}$.

Algorithm 1: Pseudo-code for an iterative photo selection algorithm based on an optimization criterion $\mathbb{O}$.

$P = \{\text{the photo collection}\}$;
$S = \emptyset$, the subset of photos already selected;
$R = []$, the re-ranking of the photo collection;
while $|S| \neq |P|$ **do**
  $p =$ next best photo out of $P \setminus S$ according to criterion $\mathbb{O}$;
  append p to R;
  $S = S \cup \{p\}$;
end
return R

5.1.1 Maximal Marginal Relevance

The Maximal Marginal Relevance (MMR) method was introduced by Carbonell and Goldstein (1998). The method originally strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and selecting appropriate passages for text summarization. This can easily be translated into the domain of photo collection summarization. The method would then strive to reduce redundancy while maintaining relevance in re-ranking the photo collection.

In order to create the new ranked list, photos are selected according to the combined criterion of query relevance and novelty of information presented in equation 5.2, which is derived from the original formulation in equation 5.1. In the original equation, R is a ranked list of documents retrieved by an information retrieval system; S is the subset of documents already selected; $R \setminus S$ is the set of documents that is yet unselected; Sim_1 is the similarity metric used in the relevance ranking between documents and a query; Sim_2 is the similarity metric between two documents; and λ is the term describing the trade-off between a relevance and novelty oriented ranking. The next document is therefore selected by their relevance, while each document is penalized by the highest similarity between the document and the documents that were already selected.

Equation 5.2 describes the MMR criterion in the context of photo collection summarization. Using algorithm 1, the re-ranked result is created by iteratively selecting the next best photo that maximizes this criterion. In this case, P is the photo collection; S is the subset of photos already selected; $P \setminus S$ is the set of photos that is yet unselected; Rel is the relevance metric for a photo; Sim is

the similarity metric between two photos; and finally λ is the term describing the trade-off between relevance and novelty.

$$\text{MMR} \stackrel{\text{def}}{=} \arg \max_{D_i \in R \setminus S} \left[\lambda \cdot Sim_1(D_i, Q) - (1 - \lambda) \cdot \max_{D_j \in S} Sim_2(D_i, D_j) \right] \quad (5.1)$$

$$\stackrel{\text{def}}{=} \arg \max_{p_i \in P \setminus S} \left[\lambda \cdot Rel(p_i) - (1 - \lambda) \cdot \max_{p_j \in S} Sim(p_i, p_j) \right] \quad (5.2)$$

5.1.2 Quantum Probability Ranking Principle

A disadvantage of the Maximal Marginal Relevance method is the required tuning of the λ parameter. A different algorithm without any hyper-parameters was chosen for comparison. This algorithm is the method based on the Quantum Probability Ranking Principe (QPRP).

The Quantum Probability Ranking Principle was first presented by Zuccon, Azzopardi, and van Rijsbergen (2009) in response to the well-known Probability Ranking Principle. The authors reformulate the Probability Ranking Principle based on quantum theory, acknowledging the interference effects between events. This interference captures the dependency between the judgments of document relevance. In simpler terms, the QPRP directly attacks the 'independence assumption' of the Probability Ranking Principle.

A full derivation of the QPRP from quantum probability theory is given by Zuccon et al. (2009), which results in the following formulation: When ranking documents, the information retrieval system has to maximize the total satisfaction of the user given the document ranking, achievable by maximizing the total probability of the ranking. The QPRP, in contrast to the Probability Ranking Principle, assumes that underlying the relevance probability distribution there is a basic concept of a complex amplitude distribution. The maximization of the total probability of the ranking then depends on the document's probabilities and their interference.[1] The strategy imposes to select at each rank position a photo p that satisfies equation 5.3 (Zuccon & Azzopardi, 2010). In this equation, P is the photo collection; S is the subset of photos already selected; $P \setminus S$ is the set of photos that is yet unselected; *Rel* is the relevance metric for a photo; and I_{p_k,p_l} is the "quantum interference" between photos I and J. Again, algorithm 1 is used to create the re-ranked result using this optimization criterion.

$$\text{QPRP} \stackrel{\text{def}}{=} \arg \max_{p_i \in P \setminus S} \left[Rel(p_i) + \sum_{p_j \in S} I_{p_i,p_j} \right] \quad (5.3)$$

The QPRP does not differ from the PRP in the sense that the relevance function is an estimation of the probability of a photo being relevant. The main issue therefore is the estimation of the interference term. Given the underlying complex amplitude distribution, this estimation of the interference term is the estimation of the amplitude's phase. Zuccon and Azzopardi (2010) state that relationships between events (photos) can be encoded in the phase, while the square roots of the estimated relevance probabilities act as a modulation component. The relevance probabilities make sure that documents that are novel

[1] A detailed explanation is given by Zuccon and Azzopardi (2010)

compared to previously selected photos but have a low estimated relevance, are not selected as the next photo in the ranked list.

Next to the relevance probabilities, the interference depends on the phase difference between the amplitudes in the underlying complex distribution. A correct estimation of this phase difference requires the ability to generate a complex amplitude distribution from real text (or photo collection) statistics, of which the feasibility is still under discussion (Zuccon & Azzopardi, 2010). For this reason, an attempt can be made to estimate the phase difference between the amplitudes associated with the photos. An interference model must be crafted on the basis of the particular information retrieval task.

Zuccon and Azzopardi (2010) model the estimation of the interference of documents for the task of subtopic retrieval by assuming that redundant relevant documents destructively interfere. Inspired by this observation, the inference of photos in the task of photo collection summarization is formulated in a similar matter. In this case, the assumption is made that relevant photos that are redundant in their photo series destructively interfere. This results in the interference term in equation 5.4. *Rel* describes the relevance estimation function; p_i and p_j describe a photo under inspection; and Sim is a similarity function between p_i and p_j. This interference term matches the interference term described by Zuccon and Azzopardi (2010). The only difference is the fact that Zuccon and Azzopardi (2010) choose Pearson's correlation coefficient as the similarity function Sim, while the Sim function in this work relies on the presence of two photos in the same photo series.

$$I_{p_i,p_j} = -\sqrt{Rel(p_i)} \cdot \sqrt{Rel(p_j)} \cdot Sim(p_i, p_j) \tag{5.4}$$

5.2 Increasing coverage

Chapter 3 showed a perspective of diversity that is ideally suited for increasing *coverage* over the information needs. This perspective is the estimation of temporal event clusters in a photo collection. With coverage based methods it is acceptable if the top ranks contain more than one photo of the same temporal cluster, if a non-redundant selection is assumed. It is, however, important that a top-k cutoff covers as many different temporal event clusters as possible. Redundancy is therefore penalized less severely than a lack of coverage of the ambiguity in a query.

5.2.1 Diversity by proportionality

Dang and Croft (2012) note that much existing work on diversity in search results focuses on penalizing result lists with too many documents on the same aspect and thereby focus on a notion of redundancy. This closely resembles the novelty-based methods that were described in the previous section. They therefore approach the task from a different perspective. The authors describe a method for increasing the diversity in search results by promoting proportionality. In their formulation, a sublist of a result list is most diverse with respect to some set of topics related to the query when the number of documents it provides on each topic is proportional to the topic's popularity. A comparison

is made to standard democratic electoral processes, where the problem is to assign a set of seats in the parliament to members of competing political parties in a way that the number of seats each party occupies is proportional to the number of votes it has received.

A simple translation can be made to the information retrieval context, and more specifically to the photo collection summarization context. Each position in the ranked list of photos R can be seen as a "seat" $r_i \in R$, each aspect of the photo collection P as a "party" t_i, and the aspect popularity w_i as the "votes" for this party. Let $S \subseteq P$ be any selection of the photo collection. Dang and Croft (2012) define S to be proportional to P, or a proportional representation of P with respect to the set of aspects of the photo collection $T = \{t_1, t_2, \ldots, t_m\}$ if and only if the number of photos in S that is relevant to each of the aspects $t_i \in T$ is proportional to its popularity w_i

The "aspects" described by Dang and Croft (2012) are the result of an explicit modeling of the query aspects given the initial query. Since there is no query available in the context of photo collection summarization, these aspects need to be modeled in a different way. The "aspects" in this context are more in line with "facets" described by Ben Carterette (2009). These "facets" are derived from the actual items in the retrieval results and is therefore a form of implicit diversity modeling. Nevertheless, the facets can be used as a proxy for the aspects, which is used in the application of this proportionality based method.

Ensuring proportionality

Dang and Croft (2012) propose two frameworks for optimizing the proportionality of the resulting ranking R: PM-1 and PM-2. PM-2 is an adaptation of PM-1, which uses a probabilistic interpretation of the diversity aspects of the photo collection. Since there are no probabilistic aspects of diversity used in this project, the PM-1 method has been used.

Chapter 6

Results and discussion

The evaluation, results and discussion are available upon request.

References

Ames, M., & Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 971–980).

Ben Carterette, P. C. (2009). Probabilistic models of novel document rankings for faceted topic retrieval. In *In proceedings of cikm.*

Bishop, C. M. (2006). *Pattern recognition and machine learning.* springer.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems* (pp. 737–744).

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, *11*(23-581), 81.

Camargo, J. E., & González, F. A. (2016). Multimodal latent topic analysis for image collection summarization. *Information Sciences*, *328*, 270–287.

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval* (pp. 335–336).

Chang, H., Yu, F., Wang, J., Ashley, D., & Finkelstein, A. (2016). Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, *35*(4), 148.

Cooper, M., Foote, J., Girgensohn, A., & Wilcox, L. (2005). Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *1*(3), 269–288.

Dang, V., & Croft, W. B. (2012). Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (pp. 65–74).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. (`http://www.deeplearningbook.org`)

Gordon, M. D., & Lenk, P. (1991). A utility theoretic examination of the probability ranking principle in information retrieval. *Journal of the American Society for Information Science*, *42*(10), 703.

Hadmi, A., Puech, W., Said, B. A. E., & Ouahman, A. A. (2012). Perceptual image hashing. In *Watermarking-volume 2.* InTech.

Han, S., Chu, C.-H., & Yang, S. (2007). Content-based image authentication: current status, issues, and challenges. In *Semantic computing, 2007. icsc 2007. international conference on* (pp. 630–636).

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29–36.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural Networks*, *1*(Supplement-1), 445–448.

Kim, J.-H., & Lee, J.-S. (2016). Travel photo album summarization based on aesthetic quality, interestingness, and memorableness. In *Signal and information processing association annual summit and conference (apsipa), 2016 asia-pacific* (pp. 1–5).

Lidon, A., Bolaños, M., Dimiccoli, M., Radeva, P., Garolera, M., & Giró-i Nieto, X. (2015, November). Semantic summarization of egocentric photo stream events. *ArXiv e-prints*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

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

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1) (No. 1). Cambridge University Press.

Nichols, D. M. (1997). Implicit rating and filtering. In *In proceedings of the fifth delos workshop on filtering and collaborative filtering* (pp. 31–36).

Radlinski, F., & Joachims, T. (2006). Evaluating the robustness of learning from implicit feedback. *arXiv preprint cs/0605036*.

Robertson, S. E. (1977). The probability ranking principle in ir. *Journal of documentation*, *33*(4), 294–304.

Sallomi, P., & Lee, P. (2016). Technology, media & telecommunications predictions. *London: Deloitte report*.

Samani, Z. R., & Moghaddam, M. E. (2017). A knowledge-based semantic approach for image collection summarization. *Multimedia Tools and Applications*, *76*(9), 11917–11939.

Santos, R. L., Macdonald, C., & Ounis, I. (2015). Search result diversification. *Foundations and Trends® in Information Retrieval*, *9*(1), 1–90.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Shen, X., & Tian, X. (2016). Multi-modal and multi-scale photo collection summarization. *Multimedia Tools and Applications*, *75*(5), 2527–2541.

Text REtrieval Conference. (2015). *Trec15 common evaluation measures.* `http://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf`. (Accessed: 2017-07-01)

Wu, Y., Shen, X., Mei, T., Tian, X., Yu, N., & Rui, Y. (2016). Monet: A system for reliving your memories by theme-based photo storytelling. *IEEE Transactions on Multimedia*, *18*(11), 2206–2216.

Zhai, C., & Massung, S. (2016). *Text data management and analysis: a practical introduction to information retrieval and text mining.* Morgan & Claypool.

Zuccon, G., & Azzopardi, L. (2010). Using the quantum probability ranking principle to rank interdependent documents. In *Ecir* (Vol. 10, pp. 357–369).

Zuccon, G., Azzopardi, L., & van Rijsbergen, K. (2009). The quantum probability ranking principle for information retrieval. *ICTIR*, *9*, 232–240.