Université de Neuchâtel
Faculté des Sciences
Institut d'Informatique

Multimodal Information Retrieval

par

Melanie Geiger née Imhof

Thèse

présentée à la Faculté des Sciences
pour l'obtention du grade de Docteur ès Sciences

Acceptée sur proposition du jury:

Prof. Jacques Savoy, Co-directeur de thèse
Université de Neuchâtel, Suisse
Prof. Martin Braschler, Co-directeur de thèse
Zurich University of Applied Sciences, Suisse

Prof. Nicola Ferro, rapporteur
University of Padova, Italie
Prof. Peter Kropf, rapporteur
Université de Neuchâtel, Suisse

Juillet 2018

UNIVERSITÉ DE NEUCHÂTEL

Faculté des Sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel – Suisse
Tél : + 41 (0)32 718 21 00
E-mail : secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Madame Melanie GEIGER

Titre:

“Multimodal Information Retrieval”

sur le rapport des membres du jury composé comme suit:

- Prof. Jacques Savoy, Université de Neuchâtel, Suisse
- Prof. Peter Kropf, Université de Neuchâtel, Suisse
- Prof. Martin Braschler, Zürcher Hochschule für Angewandte Wissenschaften, Winterthur, Suisse
- Prof. Nicola Ferro, Università degli Studi di Padova, Italie

Neuchâtel, le 13 juillet 2018

Le Doyen, Prof. R. Bshary

ABSTRACT

Knowledge-intensive business processes, one of the essential drivers of our economy today, often rely on multimodal information retrieval systems that have to deal with increasingly complex document collections and queries. The complexity mainly evolves due to a large and diverse range of textual and non-textual modalities such as geographical coordinates, ratings and timestamps used in the collections. However, this results in a explosion of combinations of modalities, which makes it unfeasible to find new approaches for each individual modality and to obtain suitable training data. Therefore, one of the major goals of this dissertation is to develop unified models to treat modalities for document retrieval. Further, we aim to develop methods to merge the modalities with little or no training, which is essential for the methods to be applicable in a wide range of applications and application domains.

We base our approach on our experience with several multimodal information retrieval applications and thus also many different modalities. In a first step we suggest a coarse categorization of modalities into two types of modalities, which we further subdivide by their distribution. The categorization is a first attempt to reduce the number of different models. It helps to generalize methods to entire categories of modalities instead of being specific for a single modality.

Since the most popular weighting schemes for textual retrieval have generalized well to many retrieval tasks in the past, we propose to use them as a basis of the unified models for the categories of modalities. We therefore demonstrate as a second step how the three main components of the so-called BM25 weighting scheme (term frequency, document frequency and document length normalization) have to be redefined to be used with several non-textual modalities.

As a third step towards establishing clear guidance for the integration of many modalities into an information retrieval system, we demonstrate that BM25 is a suitable weighting scheme to merge modalities under the so-called raw-score merging hypothesis. We achieve this with the help of a sampling-based approach, which we use as a basis to prove that BM25 satisfies the assumptions of the raw-score merging hypothesis with respect to the average document length and the variance of document lengths.

Using our redefinition of BM25 for several non-textual modalities together with textual modalities, we finally build multimodal baselines and test them in evaluation campaigns as well as in operational information retrieval systems. We show that our untrained multimodal baselines reach a significantly better retrieval effectiveness than the textual baseline and even achieve similar performance when comparing them to a trained linear combination of the modality scores for some cases.

Keywords: Multimodal, Information Retrieval, Probabilistic Information Retrieval Model, BM25 Weighting Scheme

RÉSUMÉ

Les processus basés sur le savoir, une des composantes essentielles de notre économie, requiert souvent un système multimodal de recherche d'information. De tels systèmes doivent traiter des collections de documents et des requêtes de plus en plus complexes. Cette complexité sous-jacente se situe dans le grand nombre et la diversité des modalités textuelles ou non-textuelles comme les coordonnées géographiques, les indications temporelles, ou les cotations apparaissant dans les documents. La combinaison de toutes ces modalités rend quasi-impossible la mise au point de nouvelles approches pour chaque modalité potentielle ou d'obtenir suffisamment de données d'apprentissage. Dès lors, l'un des objectifs de ce travail de thèse est de proposer un modèle unifié afin de traiter les diverses modalités en recherche d'information. De plus, nous avons développé des méthodes permettant la fusion de modalités avec peu ou en l'absence de données d'entrainement. Une telle contrainte s'avère essentielle pour des méthodes pouvant s'appliquer à un large éventail d'applications ou de domaines.

Nous avons fondé notre approche sur notre expérience touchant de nombreux systèmes multimodaux de recherche d'information. Dans un premier temps nous présentons une approche basée sur une distinction fondée sur deux types de modalités que nous subdiviserons par la suite. Ce choix correspond à une première approche dont l'objectif est de réduire le nombre possible de modèles. Elle permet de généraliser des méthodes traitant plusieurs modalités au lieu d'être spécifiques à une unique modalité.

Comme les schémas de pondération les plus populaires pour le dépistage d'information textuelle se sont généralisés avec succès dans de nombreuses tâches de recherche, nous les avons adoptés comme fondement à nos modèles unifiés traitant diverses modalités. Dans un deuxième temps, nous démontrons comment les trois composantes principales du modèle BM25 (fréquence d'occurrence, fréquence documentaire et normalisation selon la longueur du document) peuvent être redéfinies pour pouvoir traiter des modalités non-textuelles.

Dans un troisième temps, nous définissons des lignes directrices pour l'intégration de plusieurs modalités dans un système de dépistage de l'information. Dans ce but, BM25 s'avère un système de pondération permettant la fusion de modalités sous l'hypothèse des scores bruts (raw-score). Ce but est atteint par l'usage d'une approche basée sur l'échantillonnage qui est utilisée pour démontrer que BM25 satisfait les hypothèses de la fusion par les scores bruts (la longueur moyenne des documents et la variance de celle-ci).

En se basant sur notre redéfinition du modèle BM25 pouvant traiter à la fois les modalités textuelles et non-textuelles, nous avons testé notre approche par rapport à différentes références ainsi que lors de campagnes d'évaluation internationales de même que dans des contextes de production. Nous avons démontré que notre approche sans données d'apprentissage retournait une performance significativement supérieure à des systèmes classiques. De plus notre modèle (sans apprentissage) apporte des performances similaires à des systèmes basés sur une combinaison linéaire de modalités avec entraînement.

Mots-Clés: Multimodalité, Recherche d'information, Modèle de recherche d'informations probabilistes, Schéma de pondération BM25

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my co-supervisor and mentor Prof. Martin Braschler for the continuous support of my Ph.D. study, for his insights, patience, and encouragement. His guidance helped me in all the time of research and writing of this thesis.

Prof. Jacques Savoy generously accepted to be my co-supervisor in this extraordinary collaborative dissertation between the University of Neuchatel and the Zurich University of Applied Sciences. I would like to thank him for his support, advice and time to help my understanding of the problems described in this dissertation.

Besides my supervisors, I would like to thank the rest of my thesis committee: Prof. Nicola Ferro and Prof. Peter Kropf for accepting to examine this thesis and for their insightful comments and questions.

This being an applied dissertation that is partly based on joined research projects of the Zurich University of Applied Sciences and many external partners, I want to mention my colleagues at Zurich University of Applied Sciences and my collaborators at the project partners with whom I had the opportunity to work closely on multimodal information retrieval as well as other data science problems.

Most importantly, none of this could have happened without my husband, who not only motivated me to continue my study every time I was ready to quit but also provided his unconditional love, patience and endless support.

Table of Contents

Chapter 1

Introduction

1.1 Motivation & Objectives

Starting with the eighties at the latest, we have entered the so-called *information age* in which creating, disseminating and retaining knowledge have become some of the essential drivers of our economy [18]. Since then, many new jobs and businesses have been created where the work no longer is limited to applying explicit knowledge repeatedly - such as a mason when building walls. Today's jobs in the service industry (recruiting, marketing, sales, insurances, banks, etc.) often revolve around knowledge-intensive business processes (KIBP's). According to Isik et al. [10], from a broad, conceptual point of view, KIBP's can be defined as: "Processes that require very specific process knowledge, typically expert involvement, that are hard to predict and vary in almost every instance of the process. They typically depend largely on human involvement and decisions although parts of the process could be supported by automation" (p. 3818).
From a high-level view, this dissertation describes efforts to advance the automation of the tasks encountered in KIBP's. In particular, we focus on the retrieval of documents in order to gather the necessary information for such a process. This retrieval process is primarily driven by the main business-specific aspects that describe the important information of the processes. For example, for a recruiting business these aspects might be the clients, jobs, companies, skills, salary ranges, period of notice, etc., while for other businesses such as insurance companies entirely different aspects are important, e.g. insurance holders, policies, claims, hospitals, treatment costs, etc. Notably, the processes usually include numerous aspects that we claim all need to be treated by the retrieval system.
The information age has led to an ever-increasing amount of data produced every day[1]. In order to leverage the vast amount of information contained therein, businesses as well as academics have been successfully employing information retrieval (IR) systems. IR is a rather old and well-studied academic discipline in which documents[2] that are relevant with respect to an information need are obtained from a collection of documents. An IR system in

[1] According to the 2013 IBM Annual report [6], the world has been generating more than 2.5 billion bytes per day in 2012. Thereof, about 80% is unstructured data in the form of images, videos, audio, social media, embedded sensors, distributed devices, etc.

[2] In this dissertation, we will use a wide definition of the term *document*. We will consider any kind of retrievable item a document. This means that a document can potentially consist of multiple texts, images, geographical coordinates, ratings etc.

the context of KIBP's not only has to deal with the raw processing of data, but also with increasingly complex document collections that typically contain heterogeneous documents, from different sources with many different types of information that are not only textual but also non-textual such as images, locations, timestamps and ratings. Throughout this dissertation, we use the term *modalities* to describe the different types of information in the documents that can be used for retrieval. The modalities usually represent the important aspects of the business processes. Therefore, IR systems used in the context of KIBP's are mostly multimodal IR systems.

In order to assist the business processes, it is crucial for a multimodal IR system to incorporate all the business-critical aspects with respect to all the corresponding modalities. In the past however, most IR systems did not treat the modalities individually but instead concatenated the information into a single textual modality. Another strategy was to ignore or discard some of the modalities. However, we claim that in a lot of KIBP's the exclusion of (some) modalities is not a valid option and treating them as a single modality comes with potential loss of information and retrieval effectiveness.

From a high level perspective, filtering in database systems is an alternative to deal with multimodal documents. For database systems, the documents are usually called records and the modalities in the queries correspond to the filtering criteria. However, a closer look shows that the requirements of the applications for which either database systems or multimodal IR systems are used are quite different. Filtering in database systems selects only the records that exactly match the specified criteria ("exact match"). In the worst case, the result set is empty since no records fulfill all the criteria in the query. This property is useful for applications in which exact matches are essential; e.g. the blood type in a database with blood donors. Multimodal IR systems, however, do not exclude documents that do not match a modality but simply reflect this in the score, which then may lead to a lower rank in the result list ("best match"). This is particularly important for recall-oriented applications where it is unlikely that documents will match the complete query, e.g. the search for an expert in a collection of candidate profiles.

> The goal of the research underlying this study is to develop untrained models for multimodal IR systems with the following vision: The models should be generalizable to build multimodal IR systems for a large range of different KIBP's. There should be clear guidance as to how to treat each individual modality and how to synthesize an overall result using all modalities with little or no training.

In our effort to fulfill this vision, we have tackled several research challenges in the course of this dissertation. Firstly, due to the large and diverse range of different modalities in KIBP's, we aim to treat modalities with unified models instead of finding new approaches for each new individual modality. Secondly, we will address merging the results for individual modalities into a single overall result. A main consideration in both these points was the use of no or minimal training data. This seems to run counter to published trends: in 2017, Gartner named machine learning one of the top 10 strategic technology trends [9]. At the same time machine learning and in particular deep learning significantly increased the performance of many tasks and has been used to merge modalities into a single ranked list [25]. However, the numerous different modalities used in KIBP's result in a explosion of possible combinations.

As a consequence, available academic test collections can only provide training data for a few select situations. Moreover, the long tradition of heuristic methods in IR has shown that the avoidance of training data is essential for the methods to be applicable in a wide range of applications and application domains. Both by demonstrating effective approaches to multimodal retrieval with little or no training and by reflecting on the limits with respect to approaches that can be applied in scenarios where suitable training data is available, the dissertation makes a major contribution to the field.

1.2 Definitions

In the following, we define some of the major terms and notations used in this dissertation.

Information Retrieval
Information retrieval is the activity of storing and searching large amounts of (unstructured) data [23]. Typically, a user's information need is expressed in the form of queries, while the answer of the retrieval system is given by a (ranked) set of documents relevant to the query. Hereby, the retrieval system calculates a retrieval score as an order-preserving estimation of the probability of relevance of a document with respect to the information need [20]. The retrieval scores are usually calculated by a weighting scheme using the so-called *bag of words* of the documents and the queries. A *bag of words* is the output of a pre-processing step that extracts features from the documents and the queries. For example with textual information, the features extracted are usually the analyzed and normalized words[3]. The most popular weighting schemes for retrieval can all be described in terms of how they combine three main components; the *term frequency* (tf); i.e. how often a feature appears in the *bag of words* of a given document, the *document frequency* (df); i.e. in how many documents a feature appears and the document length; i.e. the number of features in the *bag of words* of a document.
Most research efforts have been dealing with *text retrieval*, which is IR on textual documents. Since 1990, content based image retrieval, which is IR on images [24], and many other related tasks have been investigated. When moving from textual to non-textual information the individual features in the *bag of words* are no longer limited to words and thus we use the more general terminology *bag of features*. Similarly, the *term frequency* is generalized to *feature frequency*.

Modality
The documents in image retrieval often not only contain images but also the textual meta-data (e.g. image caption). Similarly, information needs might be described by queries that consist of an image and a textual description. The two constituent parts of the documents and the queries (the image and the textual meta-data respectively description) in image retrieval are usually denoted as modalities [4]. In this dissertation, we broaden the term *modality* to any type of information that is part of a document or a query. For example, a document or a query could contain multiple modalities such as the number of likes, timestamps, locations (e.g. geographical coordinates), images, prices, ratings, etc. Whenever we use the term

[3]Peters et al. [19] give a more extensive overview of the steps in this pre-processing such as tokenizing, stemming and decompounding.

non-textual modality, we want to empathize that this modality cannot per se be treated like natural language as known from classical text retrieval.

Multimodal Information Retrieval System
In a multimodal IR system, the documents $(d_1, d_2, ..., d_D)$ consist of several modalities. Hereby, d_j^m is the bag of features of modality m of document d_j and D is the number of documents. Analogously, the query q consists of several modalities and q^m is the bag of features for modality m of query q.
For simplicity, we assume that we can estimate the contribution of relevance of each modality separately and estimate the probability of relevance of a document by combining the contributions of its modalities.
During retrieval, weighting schemes define the retrieval score (retrieval status value $\text{RSV}(q^m, d_j^m)$) of modality m in document d_j w.r.t. modality m in query q, which is an estimation of the contribution of modality m to the probability of relevance of the document d_j w.r.t. query q. Hence, for each modality m the result set is a ranked list ordered by the retrieval score of document d_j and modality m. Note that most retrieval scores depend on the modality length $l(d_j^m)$ which is the number of features in the bag of features of modality m of document d_j and that the modality length might be different for each modality in a document. The ranked lists of all the modalities, similarly to image and multilingual retrieval, need to be merged into a single ranked list. Hence, a function f has to be found to compute the retrieval score for each document based on the retrieval scores of all modalities

$$\text{RSV}(q, d_j) = f(\text{RSV}(q^1, d_j^1), \text{RSV}(q^2, d_j^2), ..., \text{RSV}(q^M, d_j^M)), \tag{1.1}$$

where M is the number of modalities. In the following, we call function f that combines the scores of all modalities *merging function* [4].

Document Relevance
Generally, the term *relevant documents* denotes the set of documents that meets the information need of the user. Finding a more specific definition of document relevance is however not trivial and numerous definitions of relevance have been published so far. The most popular publications about document relevance are probably the work of Cooper [5], Borlund [1] and Mizzaro [16]. Eickhoff [7] claims that the topical overlap with the user's information need is the most frequently used criterion for relevance. We are however convinced that there are numerous other aspects of relevance that have to be considered. We claim that the additional modalities as we encounter them in multimodal IR systems help to incorporate such further noteworthy dimensions of relevance; e.g. the recency, language or popularity.
For multimodal IR systems, we need to define how the individual modalities contribute to the relevance of the whole document. In some cases, a modality might further specify an information need and therefore narrows down the set of relevant documents. For example, the timestamp, when searching for current traffic jams with a query consisting of two modalities: a timestamp and a textual part "traffic jams". In other cases, the modalities can broaden the result set, since they help to further understand the broader meaning of the information

[4] Alternatively, we refer to the merging function using the term *merging strategy* or simply *merging*.

need. This often occurs if the IR system cannot accurately extract the information need from a textual modality. For example, the textual query "popular names in Switzerland" would possibly not return documents that only contain Zurich and not Switzerland. However, if the query has an additional modality for the geographical region and the documents contain locations, probably more documents that are relevant could be retrieved; in particular such that only contain city names.
Since we defined our multimodal IR system to estimate the relevance of a document based on a combination of the contribution of relevance of each individual modality, the definition of the document relevance faces additional challenges. We not only need to define the relevance of a document with respect to each modality but also how the relevance of the individual modalities affects the relevance of the whole document. This is however not always trivial, since the observation of a ranked list of an individual modality (e.g. timestamps) may not give a lot of information about the overall relevance.

Test Collections
A test collection, as used in *evaluation campaigns*[5] such as the Text REtrieval Conference (TREC)[6] and the Cross-Language Evaluation Forum (CLEF)[7], generally consists of a collection of documents, a set of queries and the corresponding relevance assessments. As described above, the documents and the queries in multimodal IR consist of several modalities. However, not all the modalities have to be present in all the documents and queries. Ideally, the test collection covers all the possible combinations of present and missing modalities in the documents and queries. Note that this is only necessary for multimodal IR test collections and it might lead to an increased number of required queries compared to traditional IR test collections [8].
For the relevance assessments in multimodal IR test collections, it is crucial that they are conducted such that they actually include the aspects of all the modalities. This is in particular a difficult task when the relevance assessment is not performed by the original users that have the whole context of the query. Usually, the relevance assessments are published after the participants submitted their runs. However, once the relevance assessment is available the test collections can also be used as training data to learn the parameters of the retrieval models for future improvements.

1.3 Achievements

This dissertation includes the following main contributions to the development of untrained models for multimodal IR:

1. In order to treat modalities with unified models instead of finding new approaches for each new individual modality, we suggest a categorization of modalities. Therefore, models only need to be developed for each category.

[5]Evaluation campaigns support the development, evaluation and comparison of IR methods by providing shared test collections for multiple research groups [19].
[6]`http://trec.nist.gov`
[7]`http://www.clef-campaign.org`

2. We justify the use of existing weighting schemes as a basis for the unified models in multimodal IR systems. In particular, we show how to generalize the so-called BM25 weighting scheme[8] for several non-textual modalities.
3. We demonstrate that the BM25 weighting scheme is suitable for merging modalities with little or no training. For that purpose, we analyze the underlying assumptions of the BM25 formula with respect to the raw-score merging hypothesis[9].
4. We establish multimodal baselines that involve all the given modalities and merge the scores generated by unified models under the raw-score merging hypothesis. We have tested them in evaluation campaigns (CLEF Social Book Search lab[10] and GeoCLEF[11]) as well as in operational IR systems and showed that for some cases they reach a significantly better retrieval effectiveness as a baseline that only uses the textual modality. For the collections where training data was available, we compared the untrained multimodal baselines to trained linear combinations of the scores of the modalities and found similar performance in the case of the Social Book Search lab.

1.4 Structure of this Dissertation

The remainder of this dissertation is structured as follows. In Chapter 2, we present the papers published in the course of this dissertation. First, the chapter provides a high-level overview of the publications. The remainder of the chapter is divided into individual sections for each paper. Each of the sections describes the detailed context and gives a short summary of the main aspects and findings of the paper, especially with respect to the goal of this dissertation of providing clear guidance for building multimodal IR systems with little or no training. Finally, Chapter 3 summarizes our contributions with respect to the state-of-the-art and gives pointers to further open questions. Appendix A provides the reprints of the publications presented in Chapter 2.

[8] BM25 ("Best Match 25") was developed as part of the participation of the Okapi system in the early TREC evaluation conferences. The first Okapi system used a simpler version of the ranking formula and was called BM1 formula ("Best Match 1") [12]. In the subsequent participations more elaborate ideas were included and the number was used to indicate the version of the formula.

[9] The raw-score merging hypothesis postulates that similarity values are directly comparable if they are produced from similar search engines and underlying collections with similar statistics [2, 14, 21, 22].

[10] `http://social-book-search.humanities.uva.nl`

[11] `http://www.clef-initiative.eu/track/geoclef`

Chapter 2

Presentation of Publications

The core of this dissertation is based on the following five publications.

- Melanie Imhof[1], Martin Braschler.
  Are Test Collections "Real"? Mirroring Real-World Complexity in IR Test Collections
  In *Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth Jones, Eric San Juan, Linda Cappellato, Nicola Ferro (Hg.)* Experimental IR meets Multilinguality, Multimodality, and Interaction, Proceedings of the 6th International Conference of the CLEF Association, CLEF'15 Toulouse, France, pages 236-232, September 8–11, 2015, Heidelberg: Springer.
- Melanie Imhof[1], Ismail Badache, Mohand Boughanem.
  Multimodal Social Book Search
  In *Linda Capellato, Nicola Ferro, Gareth Jones, Eric San Juan (Eds.)*, CLEF 2015 Labs Working Notes, Toulouse, France, September 8-11, 2015, Aachen: CEUR.
- Melanie Imhof[1].
  BM25 for Non-Textual Modalities in Social Book Search
  In *Krisztian Balog, Linda Capellato, Nicola Ferro, Craig Macdonald (Eds.)*, CLEF 2016 Labs Working Notes, Évora, Portugal, September 5-8, 2016, Aachen: CEUR.
- Melanie Imhof[1], Martin Braschler.
  A Study of Untrained Models for Multimodal Information Retrieval
  In *Information Retrieval Journal*, 21(1):81-106, 2018
- Melanie Geiger, Martin Braschler.
  Overcoming the Long Tail Problem: A Case Study on CO2-Footprint Estimation of Recipes using Information Retrieval
  In *Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga* Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7-12, 2018, Paris: European Language Resources Association (ELRA).

[1]Legal name change to Melanie Geiger as of August 11, 2017.

These publications systematically address the challenges and necessary steps for building an untrained model for multimodal IR. The following sections give a summary of the motivation, methods and contributions for each of these publications, while the full articles, containing the related work, results and discussions, can be found in Appendix A.
In Section 2.1 we argue that multimodal test collections used in evaluation campaigns do not yet reach the complexity of operational collections in real-world applications, thus having limited value in building multimodal IR systems. Further, we propose a categorization of modalities, which allows the development of unified models for a category of modalities. Section 2.2 summarizes our participations to the Social Book Search lab at CLEF and presents a trained model using random forests and a first attempt using the weighting scheme BM25 for non-textual modalities. In Section 2.3, we discuss how to validate the suitability of the BM25 weighting scheme for multimodal IR systems and the analysis of the underlying assumptions of the BM25 formula with respect to merging modalities without training. We conclude with Section 2.4 in which we describe how we used the presented methods in an operational multimodal IR application.

2.1 Multimodal Information Retrieval Test Collections

Due to the complex notion of relevance and other factors that prohibit a well-defined result, the objective evaluation of retrieval effectiveness has a long tradition in IR. Evaluation campaigns enable the comparison of different retrieval systems and therefore help to increase the retrieval effectiveness of many IR methods. As stated before, for multimodal IR the notion of relevance of a document is a complex issue, which calls for a careful evaluation methodology and test collections to evaluate the results. Therefore, we start this synopsis of our work with an overview and position paper about the analysis of the state-of-the-art of multimodal IR evaluation, more specifically about the available test collections in evaluation campaigns such as TREC and CLEF; as well as operational applications.

The initial motivation for this dissertation dates back to 2013, when we first encountered a multimodal IR application in a research project with an industry partner[1]. In the project, we worked on a noise canceling news feed application that collects documents from various sources such as public search engines, social networks and news feeds in general. The queries in this application not only consist of a textual modality of the user's interests that the user explicitly defines, but also of the user's preferences with respect to recency, language, popularity and source-quality of the documents that is implicitly defined by the user's past behavior. The documents consist of various modalities such as timestamps, number of likes, number of re-tweets, the source of the documents, the language of the documents, etc. Due to the small scope of the project, the availability of training and test data was limited. In order to verify the developed methods in a comparable setting, we searched for similar academic collections. However, none of the collections we found fulfilled our needs.

In the course of this thesis, we analyzed the collections in the past CLEF labs and found that their complexity has increased over the years, mostly due to the INEX track introduced in 2012. In this context, we assume that more modalities lead to more complexity and thus we used the number of modalities in the documents and queries as a measure of complexity of the task. The complexity particularly increases with the number of modalities due the additional effort that is necessary to combine the contributions of the individual modalities into a single overall result. Although the complexity of the CLEF labs has increased over the years, these collections are still far less complex than what we have observed in operational collections. From this analysis, we concluded that a collection that mirrors the complexity of real-world IR applications should contain a large amount of modalities from different categories of modalities, some of the modalities should be independent from each other, and others should be inter-related.

In the paper, we further claim that the consequence of the lack of complex test collections is that it is unclear how to incorporate upcoming new modalities into IR systems leading to much effort. In addition, most methods have been developed for a single very specific modality and have not been generalized to other modalities. Further, only a few very basic approaches have been developed to combine the modalities in a meaningful way without using training. We suggest a categorization of modalities, which should help to develop methods that can be generalized to a whole category of modalities. In this categorization, we distinguish between

[1]The project was funded by the Swiss commission for technology and innovation (CTI) with the funding number "13821.1 VOUCH-ES".

ordered[2] and descriptive modalities. The values of ordered modalities such as ratings, dates and prices have a natural order and we believe that this order is important for retrieval and needs to be considered. The values of descriptive modalities such as terms and SIFT features do not have an order that contains relevant information for retrieval. We are convinced that it is easier to develop different retrieval methods for these two categories of modalities, since each of them comes with different challenges.

[2]Note that in this context, *order* denotes the order of features independent from the documents (e.g. dates ordered in a timeline) and not the order of the features in the documents such as the position of a term.

2.2 Experiments with Multimodal Collections

Our participation in the "suggestion track" of the "INEX Social Book Search (SBS) lab" at CLEF in two consecutive years gave us the opportunity to explore the multimodal test collection of this lab. The challenge is to develop methods to retrieve books, from a collection with 2.8 million book records, as requested by real users of the social cataloging web application LibraryThing. For each book not only the meta-information from Amazon (description, binding, number of pages, price, etc.) is available but also user generated information such as the books' ratings and reviews. A query consists of the user's textual description for a book recommendation as well as the user's personal catalog. In our participation in both years, we extracted additional implicit non-textual modalities from the personal catalogs, so that the query was multimodal consisting of four modalities: the textual description, the user preference w.r.t. the book length, the user preference w.r.t. the book price and an assumed general preference of books with high ratings.
The main goal of this exploration was to see if a strong textual baseline can be significantly improved using the additional modalities. Therefore, we tried different models to incorporate the modalities and we checked the correlation between the different modalities, in particular their information overlap.
Even though numerous definitions of document relevance have been proposed in the past, the document relevance is frequently limited to a topical overlap with the user's information need [7]. In the context of multimodal IR, the document relevance however needs to exceed the topicality and has to include all aspects of the modalities. We therefore carefully examined the relevance assessment provided in the SBS lab. Since one of the goals of the suggestion track of SBS is to go beyond topical relevance [13], they use the suggestions from real users as a basis for the relevance assessment. Several rules are applied to extract graded relevance values from these suggestions. Although the modalities are not explicitly considered, these rules potentially include the additional modalities implicitly. For example, one rule assigns a higher relevance value if the user that suggested a book has actually read it. This ensures that rather objective criteria are considered, such as our assumption that users prefer books with higher ratings. Further, books that have later been added to the catalog of the requesting user are considered more relevant than others. This gives us an indicator for whether the book matched the users' preferences.
In the paper "Multimodal Social Book Search" we describe our first participation in the SBS lab at CLEF 2015 for which we collaborated with Ismail Badache and Mohand Boughanem from IRIT - Paul Sabatier University, Toulouse, France. We built a strong textual baseline and combined it with a social document prior based on social signals proposed by our colleagues from Toulouse. Further, we used three modalities, the book's price, the number of pages and the book's ratings as additional non-textual modalities. We used a random forest to learn how to combine the scores of the individual modalities. Our main finding was that we were able to significantly improve the retrieval effectiveness of the textual baseline using the additional modalities. This approach resulted in the best-ranked run in the SBS competition at CLEF 2015. However, the proposed method heavily relies on the availability of a test collection and is not directly transferable to other collections and modalities.
Our second participation in the SBS lab at CLEF 2016 is described in "BM25 for Non-Textual Modalities in Social Book Search". We used the same three modalities, the book's price, the

number of pages and the book's ratings, as in 2015 but we applied a unified model based on BM25. Based on the raw-score merging hypothesis, we suspected that using the same weighting scheme for all modalities gives us the possibility to merge their scores without training. However, since we did not yet understand the underpinnings of a suitable weighting scheme to merge modalities without training, we still used a trained linear combination to merge the scores.

Similar to most weighting schemes used in IR, BM25 is defined by its three main components, the term frequency, the document frequency and the document length. Therefore, whenever using BM25 with other modalities, we need to define the three components for this particular modality or for the category of modalities. In the paper, we thus show how to redefine the three components for the three non-textual modalities.

For the ratings, we assume that users generally prefer books with higher ratings. Accordingly, we define the three components so that higher ratings will result in a higher BM25 score.

For the price and the number of pages of a book, we define the query based on the user preferences, i.e. the average price respectively the average number of pages of the books a user has already read. The price and the number of pages are continuous variables and therefore exact matches of the values in the query and the values of the books are not meaningful. Thus, we define the three components to result in a fuzzy query. This definition is only a first attempt to handle continuous variables and is probably not generalizable to all other modalities. Therefore, future work has to investigate generalizable models for continuous variables using the existing weighting schemes.

In the paper, we show that the proposed approach using BM25 for non-textual modalities and a linear combination of the scores of the modalities achieves a significantly higher retrieval effectiveness than the textual baseline.

2.3 Untrained Models for Multimodal Information Retrieval

This work is the centerpiece of this dissertation. It is based on the lessons of the other publications in this dissertation, in particular the following four:

1. In the context of KIBP's, a large variety of IR applications evolved. These applications have to deal with increasingly complex document collections and queries, primarily due to increasing number of different modalities.
2. Due to the large variety of IR applications and lack of suitable test collections, we argue that it is crucial to treat the modalities with unified models instead of finding new approaches for each individual modality.
3. Even though weighting schemes have been developed for retrieval on English text, they generalized well before to other related tasks such as multilingual retrieval, multimedia retrieval and others; and we are therefore hopeful that they can be generalized to other modalities.
4. Since weighting schemes can be described in terms of how they combine three main components (term frequency, inverse document frequency and document length normalization), we want to attempt to define these three main components for each category of modalities.

As a first step to develop generalizable models, the core of this publication discusses the underpinning of weighting schemes for textual retrieval based on the four lessons mentioned above and shows how they can be applied or adapted methodically to non-textual modalities. The main contribution is that we demonstrate that BM25 is a suitable weighting scheme for non-textual modalities and to merge them without any training.
Our early efforts in developing generalizable models focused on the definition of the term frequency and the document frequency for non-textual modalities. However, we soon realized that the role of the document length and the document length normalization in the merging process is unclear and thus needs further investigation. Therefore, we studied the document length components of the most popular weighting schemes. Note that in the context of multimodal IR system, we no longer have a single document length, but instead have a length for each modality. We conducted several experiments to find the weighting scheme that is most robust with respect to varying document lengths and therefore is most likely suitable to be used in a multimodal IR system. Furthermore, we looked into alternative weighting schemes that potentially are not dependent on document length normalization such as passage retrieval and proximity weighting. However, we found that even these weighting schemes rely on a document length normalization component[11], [3], [27], [15].
The modalities usually have vastly different lengths. For example in an application that deals with newspaper articles, the textual modality could have a length of several hundred terms while the timestamp modality usually has the length of one. The document length normalization component only focuses on the normalization of the lengths of the documents within a modality. Therefore, we looked into other application domains to find solutions to avoid the document length completely and to handle the different lengths across the modalities. We found that in multimedia retrieval, in particular image retrieval, the document length normalization as part of the weighting scheme is circumvented by sampling a fixed number

of features for each image regardless of the image size and the number of concepts in the image. Accordingly, we applied this idea to the features in multimodal IR systems. This means that we sample the same number of features for each document and each modality and then apply any weighting scheme without the necessity for a document length normalization component. This approach, however, has issues with data-loss due to downsampling and is not deterministic due to the random sampling. Therefore, we introduced a novel idealized sampling approach in which we do not randomly sample the documents but simply simulate the average resulting feature statistics. This approach solves two issues regarding merging. First, the role of the document length normalization for the merging is no longer relevant, since none is used. Second, we now have modalities that all have the same collection statistics with respect to the document length and the document length variance, which is ideal for merging under the raw-score merging hypothesis.

In the paper, we prove that applying BM25 with full document length normalization $b = 1$ is identical to our ideal sampling approach. Therefore, we show that BM25 is equally suited for merging under the raw-score merging hypothesis. Analogously, we propose a scope-aware sampling approach that deals with the fact that some documents can be more verbose than others. Therefore, we sample the documents to different lengths, similar to the concept introduced by Robertson's document length normalization in BM25. We were able to prove that BM25 with a general document length normalization parameter $b \neq 1$ is equal to the scope-aware sampling approach and therefore the raw-score merging hypothesis w.r.t. the average document length also holds for BM25 in general. However, unlike for the fully normalized BM25, only the average document length is the same for all modalities but not the variance of document lengths.

Based on these findings, we implemented a multimodal baseline using BM25 with raw-score merging for the SBS and the GeoCLEF collection. The experiments show that the approach leads to encouraging results. Not only because the textual baseline can be significantly improved, but also because for the SBS collection, our untrained multimodal baseline achieves a similar performance as a trained linear combination of the modalities.

2.4 An Application of Multimodal Information Retrieval in Industry

The initial motivation for the topic of this dissertation came from the ever-increasing complexity in operational IR applications in industry. Therefore, we conclude this presentation of our publications with the paper "Overcoming the Long Tail Problem: A Case Study on CO2-Footprint Estimation of Recipes using Information Retrieval". In this paper, we describe the solutions for an operational multimodal IR application that we proposed and implemented in a research project with our industry partner Eaternity AG[3]. The paper was co-authored by Martin Braschler and both authors have been involved in the project[4].
The goal of the project was to automatically calculate the CO_2-footprint of cooking recipes. This task is an example of the fact that the field of IR has evolved substantially beyond classical IR that has been concerned with building systems that retrieve ranked lists of documents in response to information needs formulated as queries. In recent years, new IR challenges have been addressed such as the attempts to synthesize more concise "answers" to information needs. In our case, we calculate a CO_2-footprint for a recipe. We can characterize a whole group of retrieval applications similar to this CO_2-footprint calculation. They all require the calculation of a single numerical value from a semi-structured item that consists of a list of textual elements. Besides the calculation of the CO_2-footprint using instructions lines such as "100g carrots, sliced" and "1 pizza dough" in the cooking recipe, other applications such as the calculation of the insurance value of a real estate from its facts such as "Bedrooms: 4" and "Heating: Oil-Fired Central Heating" fall into this group.
Before the project started, Eaternity calculated the CO_2-footprint of recipes by first manually matching all the ingredient descriptions to a database with food products and their corresponding CO_2-values and then calculating the estimate by the sum of the CO_2-values of each food product multiplied with the corresponding amount. This is a very time-consuming process and therefore an expensive approach. A logical first step to automate the calculation is to replicate the manual process by replacing the manual matching with an IR search on the food product database; we call this approach *ingredient matching*. Both the manual and the automatic ingredient matching approaches need to deal with the difficulties that stem from the fact that recipes are usually written in natural language and are therefore not restricted to the fixed vocabulary in the food product database. For example, the difficulties that arise with too specific ingredient description such as "Pinot Noir" as well as unspecific descriptions like "fillet of fish" need to be handled. However, the largest challenge is the long tail problem that arises since the food products used in recipes worldwide are manifold and new products are continuously introduced to the marketplace. Müller et al. [17] have shown that most of the food products appear only in very few recipes; hence they arrange themselves according to the so-called Zipf's law [26], meaning that most entries relate to entities that occur very infrequently. While this is less problematic for the manual approach where a human assessor can decide if a new food product needs to be added to the database or if it can be matched

[3]The project was funded by the Swiss commission for technology and innovation (CTI) with the funding number "16699.2 PFES-ES".

[4]Other notable contributors include Auerlian Jaggi and Manuel Klarmann from Eaternity, who not only helped us to understand the business case and the domain specific issues, but also integrated our solutions into their application.

to a very similar existing product, a fallback strategy has to be defined for the automatic matching. In our case, the fallback strategy is to assign an artificial food product that has an average CO_2-value of all food products in the database.
In order to overcome this long tail problem, we proposed a second approach *recipe matching* that is based on the idea of finding the nearest neighbor of a recipe and estimating its CO_2-value from the nearest neighbor instead of matching all the ingredient descriptions individually. The centerpiece of this paper is therefore the recipe matching approach for which we propose an adapted version of BM25, which allows us to incorporate the two available modalities; the food products and their amounts. Unlike the modalities in most other applications we dealt with, these two modalities are very tightly coupled. The amounts alone do not bear any information, as they are only an attribution of the ingredient descriptions. Therefore, our adapted version of BM25 directly incorporates the amounts of the ingredients to the weighting of the individual terms of the ingredient descriptions.
The third approach is a hybrid approach that combines the estimates of the two other approaches and therefore is able to provide a more reliable estimate.
As reported by our project partner Eaternity AG, the commercialized CO_2-calculation service based on the presented approaches experiences a decreased effort in calculation by 50-60% and a reduced overall cost by 80% compared to the previously used manual process.

Chapter 3

Conclusion

In this dissertation, we presented our findings with respect to the following vision to deal with multimodal IR systems, which often appear in the context of KIBP's. We envision clear guidance as how to treat all modalities in any complex multimodal IR system and how to synthesize an overall search result with little or no training.
In order to achieve this vision, we suggested treating modalities with unified models instead of finding new approaches for each modality in the large and diverse range of different modalities. The unified models need to cope with the different characteristics of the modalities with respect to different types of information (e.g. numeric or textual), different distributions of the feature values (e.g. Zipfian) and many others.
The numerous different modalities used in KIBP's result in a explosion of possible combinations. Hence, test collections cannot provide training data for all possible situations and we therefore need methods that use little or no training to be applicable to a wide range of applications. The untrained merging of modalities has several implications to be considered. First, the unified models used for the individual modalities possibly need to fulfill several requirements, such as the raw-score merging hypothesis. Meeting these requirements is however not sufficient, since the modalities are not completely independent. Therefore, we need methods to treat both inter-related modalities and modalities that contain overlapping information. Moreover, the methods need to address the possible differences in informativeness regarding the information need between the modalities.
In the following sections, we first describe what we achieved to get closer to this vision, followed by a summary of problems to tackle in the future in order to completely fulfill it.

3.1 Summary of Contributions

The contribution of this dissertation regarding the development of untrained models for multimodal IR systems, and towards the vision described above, can be summarized as follows.
In a first analysis of multimodal IR systems, we realized that it is unfeasible to develop models for each of the numerous modalities. Therefore, we suggested categorizing the modalities, so that models only need to be defined for each category. We proposed a first coarse categorization of modalities in which we distinguish between two types of modalities; i.e. ordered and descriptive and two distributions per type of modalities; i.e. continuous

and discrete for the ordered modalities and open and closed vocabulary for the descriptive modalities.

In the pursuit of a suitable model, we found that the most popular weighting schemes that were originally developed for textual modalities in English have been proven to generalize well to other tasks such as multilingual retrieval and multimedia retrieval. The reason is that they all consist of the same three main components, the term frequency, the document frequency and a document length normalization. These components are used to determine the characteristic terms and to make sure that verbose documents are not favored in an undue way. We have seen that the concept of *being characteristic*, embodied through the term frequency as well as the document frequency, is quite generalizable to all the modalities we encountered so far. Therefore, a major contribution of this dissertation is that we justified using the existing weighting schemes as a basis for the unified models for the categories of modalities.

Our participation in the SBS lab at CLEF gave us the opportunity to experiment with a multimodal collection in a competitive campaign. In the first participation, we employed a trained approach using three non-textual modalities (price of a book, length of a book and ratings) besides the textual modality. This approach resulted in the best-ranked run in the 2015 evaluation campaign. As the goal of this dissertation is to build untrained generalizable methods for multimodal IR systems, we used the gathered knowledge about the collection to develop a second approach that relies less on training. In this context, we demonstrated how to generalize the popular weighting scheme BM25 to the same three non-textual modalities and how to merge the produced ranked lists with the ranked list of the textual modality in order to achieve a multimodal baseline that significantly improves the textual baseline.

For the purpose of building untrained models for multimodal IR systems, we analyzed the underlying assumptions of the BM25 formula w.r.t merging modalities under the raw-score merging hypothesis. The raw-score merging hypothesis states that scores are directly comparable if they are produced by similar search engines and similar underlying collection statistics. However, one of our findings was that usually modalities have very different collection statistics. We therefore introduced a sampling-based approach for BM25 that ensures that all modalities have the same collection statistics; in particular the average document length and the variance of the document lengths of the modalities. As a major contribution, we proved that applying BM25 with full document length normalization $b = 1$ to all modalities already ensures that the raw-score merging hypothesis w.r.t. the average document lengths and the variance of document lengths is fulfilled, since it is identical to the sampling-based approach. Analogously, we proved that the raw-score merging hypothesis w.r.t. the average document length also holds for BM25 with a general document length normalization parameter $b \neq 1$, however not w.r.t. the variance of document length. In our experiments, we established a multimodal baseline that involves all the given modalities and merges the scores generated by a unified model under the raw-score merging hypothesis. The results of the multimodal baselines for the SBS and the GeoCLEF collection show that adhering to the raw-score merging hypothesis is indeed beneficial. Another important step into the direction of fulfilling the vision described above is that we found similar performance of our multimodal baseline when comparing it to a trained linear combination of the scores in case of the SBS collection.

Finally, we used the methods developed in this dissertation for an operational multimodal IR application in which the goal was to automatically calculate the CO_2-footprint of cooking recipes. The automatic calculation of CO_2-footprints is an example of a group of IR applications that deal with the calculation of a numerical value from a semi-structured item. A commonly used approach to calculate such numerical values is to individually match all the elements of the semi-structured item to a database and then compute the value of the complete item by aggregating the values of the individual elements. However, a challenge of this approach is the *long tail problem* that arises with the large diversity of possible elements. We showed that we can overcome this long tail problem using a search based approach that uses the value of the nearest neighbor as an estimate. Our main contribution is that we proposed an adapted version of BM25 for the nearest neighbor search, which allows us to incorporate the two tightly coupled modalities of this use case; the food products and their amounts.

3.2 Future Work

Developing guidelines for untrained multimodal IR systems is an ambitious goal and there are still many challenges left.

Extracting and mapping the business-critical information of KIBP's to modalities is a non-trivial task, since the information can be embedded within text, images, meta-data, etc. and the documents might come from different sources with different formats. In the scope of this dissertation, we implicitly extracted and mapped the modalities in a few examples and did not develop a generalized method. Therefore, future work would need to analyze this in more depth, potentially reusing the methods developed in the fields of knowledge extraction and data pre-processing.

The categorization of modalities we suggested is only a first step. These categories of modalities have to be further subdivided in order to make sure that we can use the same weighting scheme for all the modalities in the same category. So far, modalities like timestamps, the number of pages of a book and the ratings of a book belong to the same category of modalities, namely the category of *ordered continuous* modalities. However, the meaning and the necessary treatment of these modalities needs to differ as we showed in our work using the SBS collection. Further, inter-related and multi-dimensional modalities, such as geographical coordinates are not yet covered by the categorization of modalities. Similarly, it needs to be defined how pairs of modalities such as the amounts of ingredients in the project where we calculated CO_2-footprints from recipes can be categorized. A deeper analysis should give insights if possible categorizations should model them as multi-dimensional modalities (amount of ingredients along with the corresponding textual description), as two inter-related modalities or if another category of modalities needs to be introduced for such pairs.

Even though we did not investigate how well BM25 generalizes to modalities with a non-Zipfian distribution in this dissertation, we are convinced that this needs to be analyzed in order to broaden the applicability of the suggested methods. This question arises particularly because the heuristic definition of the inverse document frequency was originally motivated by Zipf's law. Several interpretations of the inverse document frequency give hints that BM25 potentially is still generalizable to modalities with other distributions as long as the term

frequency and the inverse document frequency can be defined in a way that the characteristic features emerge.

Once all modalities can be categorized, future work has to develop models to treat each category of modalities. Ideally, the models are based on a unified weighting scheme in order to ensure the scores of the modalities can be merged under the raw-score merging hypothesis. Moreover, the following open questions regarding the merging of scores of modalities need to be looked into: We saw that the raw-score merging did not yet work properly for inter-related modalities as we encountered them in the GeoCLEF collection. Therefore, it has to be investigated if the problem needs to be solved as part of the merging or if the model for the modalities needs to be adjusted. Moreover, we saw in the experiments that there are wildly different degrees of informativeness across the modalities that are not considered by raw-score merging. As a next step towards clear guidance for multimodal IR systems, it has to be investigated how to further extend the proposed methods in order to incorporate the informativeness of the different modalities with untrained models.

In order to tackle the remaining challenges described above an evaluation campaign on multimodal IR would be helpful. For such a campaign the modalities should be selected in a way that the desired effects and problems emerge, e.g. modalities with non-Zipfian distributions, with different degrees of informativeness as well as multi-dimensional modalities. Furthermore, not only the documents should be multimodal but also the queries so that investigations with respect to the user's context and to the effects of different combinations of modalities in the queries and the documents can be performed.

References

[1] Pia Borlund. The concept of relevance in IR. *Journal of the Association for Information Science and Technology*, 54(10):913–925, 2003.

[2] Martin Braschler. Combination approaches for multilingual text retrieval. *Information Retrieval*, 7(1):183–204, 2004. ISSN 1573-7659.

[3] James P Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, 1994.

[4] Paul Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas M Lehmann, Jeffery Jensen, and William Hersh. The clef 2005 cross–language image retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 535–557. Springer, 2005.

[5] William S Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971.

[6] IBM Corp. 2013 IBM Annual Report, 2013. URL `https://www.ibm.com/annualreport/2013/bin/assets/2013_ibm_annual.pdf`.

[7] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 871–880, 2012.

[8] Melanie Imhof and Martin Braschler. Are Test Collections "Real"? Mirroring Real-World Complexity in IR Test Collections. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 236–232. Springer International Publishing, 2015.

[9] Gartner Inc. Gartner's top 10 strategic technology trends for 2017, 2016. URL `https://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017`.

[10] Öyku Isik, Joachim Van den Bergh, and Willem Mertens. Knowledge intensive business processes: An exploratory study. In *2012 45th Hawaii International Conference on System Sciences*, pages 3817–3826, 2012.

[11] Marcin Kaszkiel and Justin Zobel. Passage retrieval revisited. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, 1997.

[12] E Michael Keen. Okapi at trec. In *The First Text REtrieval Conference (TREC-1)*, volume 500, page 21, 1993.

[13] Marijn Koolen, Toine Bogers, Maria Gäde, Mark Hall, Iris Hendrickx, Hugo Huurdeman, Jaap Kamps, Mette Skov, Suzan Verberne, and David Walsh. Overview of the CLEF 2016 social book search lab. In *International Conference of the Cross-language Evaluation Forum for European Languages*, pages 351–370, 2016.

[14] KL Kwok, L Grunfeld, and DD Lewis. TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. *NIST Special Publication SP*, pages 247–247, 1995.

[15] Yuanhua Lv and ChengXiang Zhai. Positional language models for information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2009.

[16] Stefano Mizzaro. Relevance: The whole history. *Journal of the Association for Information Science and Technology*, 48(9):810–832, 1997.

[17] Manuel Müller, Morgan Harvey, David Elsweiler, and Stefanie Mika. Ingredient matching to determine the nutritional properties of internet-sourced recipes. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 73–80, 2012.

[18] OECD, editor. *Creating Value from Intellectual Assets*, 2006. Organisation for Economic Co-operation and Development.

[19] Carol Peters, Martin Braschler, and Paul Clough. *Multilingual information retrieval: From research to practice*. Springer Science & Business Media, 2012.

[20] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[21] Jacques Savoy. *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002 Rome, Italy, September 19–20, 2002 Revised Papers*, chapter Report on CLEF 2002 Experiments: Combining Multiple Sources of Evidence, pages 66–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-45237-9.

[22] Jacques Savoy. *Data Fusion for Effective European Monolingual Information Retrieval*, pages 233–244. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-32051-7. doi: 10.1007/11519645_24. URL `http://dx.doi.org/10.1007/11519645_24`.

[23] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge University Press, 2008.

[24] Arnold W. M. Smeulders, Marcel Worring, Stefania Santini, Atul Gupta, and Ravi Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[25] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2012.

[26] George Kingsley Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

[27] Justin Zobel, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis. Efficient retrieval of partial documents. *Information Processing & Management*, 31(3):361–377, 1995.

Appendix A

Publications

A.1 Are Test Collections "Real"? Mirroring Real-World Complexity in IR Test Collections

Melanie Imhof[1], **Martin Braschler.**
In *Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth Jones, Eric San Juan, Linda Cappellato, Nicola Ferro (Hg.)* Experimental IR meets Multilinguality, Multimodality, and Interaction. Proceedings of the 6th International Conference of the CLEF Association, CLEF'15 Toulouse, France, pages 236-232, September 8–11, 2015, Heidelberg: Springer.

[1]Legal name change to Melanie Geiger as of August 11, 2017.

Are Test Collections "Real"? Mirroring Real-World Complexity in IR Test Collections

Melanie Imhof[1,2] and Martin Braschler[2]

[1] Université de Neuchâtel, Neuchâtel, Switzerland
[2] Zurich University of Applied Sciences, Winterthur, Switzerland
{imhf,bram}@zhaw.ch

Abstract. Objective evaluation of effectiveness is a major topic in the field of information retrieval (IR), as emphasized by the numerous evaluation campaigns in this area. The increasing pervasiveness of information has lead to a large variety of IR application scenarios that involve different information types (modalities), heterogeneous documents and context-enriched queries. In this paper, we argue that even though the complexity of academic test collections has increased over the years, they are still too structurally simple in comparison to operational collections in real-world applications. Furthermore, research has brought up retrieval methods for very specific modalities, such as ratings, geographical coordinates and timestamps. However, it is still unclear how to systematically incorporate new modalities in IR systems. We therefore propose a categorization of modalities that not only allows analyzing the complexity of a collection but also helps to generalize methods to entire modality categories instead of being specific for a single modality. Moreover, we discuss how such a complex collection can methodically be built for the usage in an evaluation campaign.

Keywords: Collection complexity, modality categorization, evaluation campaigns.

1 Introduction

Evaluation campaigns such as TREC[3] and CLEF[4] have been a great success in bringing objective benchmarking to many areas of IR research. A fundamental problem of the approach of those campaigns however, is their reliance on the Cranfield paradigm or IR evaluation [4, 8] and therefore the cost of producing test collections. Consequently, only a few test collections are created every year. In order to be cost-efficient and transferable to industrial applications, a common goal of those campaigns is to make the evaluations as realistic as possible. In the past years, the focus was mostly on increasing the variety of domains and tasks covered by the test collections as well as on the comprehension of the user's role [2]. However, in reality, the increasing pervasiveness of information has not only lead to an ever increasing amount of information, but also to a much larger variety of IR application scenarios that leverage this information. This leads to an increasing complexity in the document collections that underlie these applications. The

[3] http://trec.nist.gov
[4] http://www.clef-initiative.eu

complexity evolved primarily from the increasing number of different information types (modalities) used in both the collections and the queries. The collections contain heterogeneous documents, from different sources with many different modalities, such as text and images, as well as the multimodal context. Hereby, the context can include user interactions with the system, such as ratings and click-paths. Further, the information needs are represented with more complex queries that additionally contain the personal and situational context, multimedia examples and many more.

The leading evaluation campaigns have reacted to this increase in complexity and this is reflected in the test collections they produce. Figure 1 shows how the complexity of the collections used at CLEF increased over the last sixteen years. Note that the average number of modalities in the collections has increased significantly in 2012 mostly due to the INEX track. However, our experience in working with practitioners has shown that the complexity of most academic collections has still not reached the complexity level of operational collections. Collections used in practice mostly not only include more different modalities but also modalities of different importance that are sometimes highly inter-dependent but at other times are complementary to each other. As a consequence, the performance of the participants of the existing evaluation campaigns does not necessarily indicate how to approach such collections and thus the developed methods are ultimately not transferable to real use cases.

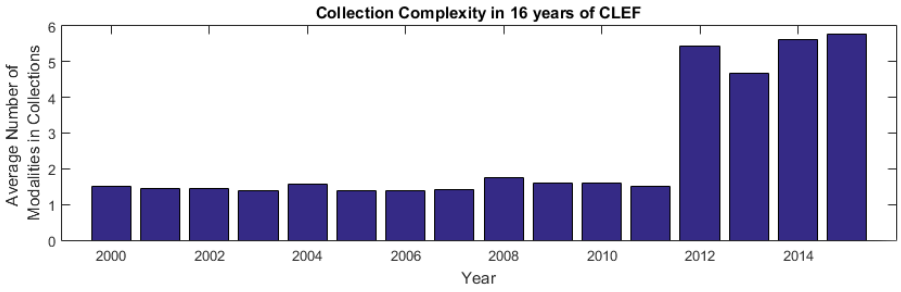


Fig. 1: Average number of modalities in the collections over the last sixteen years at CLEF.

Until now, it is often unclear how to best systematically incorporate upcoming new modalities into IR systems. Most retrieval methods have been developed for a single very specific modality, e.g. geographical coordinates, and have not been generalized to other modalities or modality categories. In practice, for complex collections, one is left with the challenging task to assemble a number of these methods and combine them in a meaningful way. As a first step to approach this problem more thoroughly, we propose a categorization of modalities which should help to generalize the methods for single modalities to the entire category. An example of how well the same methods work for different modalities is the usage of the TF-IDF and BM25 weighting schemes in both text retrieval and image retrieval.

In this paper, we compare academic collections as provided in evaluation campaigns and operational collections as found in IR applications in the industry and we propose a categorization of modalities that allows methods to be generalized to entire modality categories. Further, we show which properties a collection should fulfill to accurately mirror the complexity of real-world collections.

2 Status Quo and Related Work

In our work with practitioners we have seen that today's IR applications are unsurprisingly no longer limited to the traditional library scenario, but are used in various more complex use cases such as online shops and news streaming applications. The documents and the queries in these applications consist of a larger and more diverse set of information that has to be considered. Also, studies about the relationships of task complexity and the use of information resources have shown that the more complex a task is the more information sources are used [7][3]. IR applications designed to handle complex tasks require more complex collections, since multiple information sources need to be incorporated. In our technology transfer projects, we have been challenged to create IR systems that can handle such complex collections.

The database research community [1] has identified the problem of managing structured, semi-structured and unstructured data from various sources as one of their long-term goals. Thus, they face a similar problem to the increasing collection complexity in IR, to efficiently incorporate all aspects of this heterogeneous data. They appeal for collaboration with the IR community, for methods to query such complex collections and for creating corresponding data collections.

The lack of complexity we identified in academic test collections is not an entirely new observation, as evidenced by the following quote from Kekäläinen and Järvelin in 2002 [5]: "*The test collections, albeit nowadays large, are structurally simple (mainly unstructured text) and topically narrow (mainly news domain). The test documents mostly lack interesting internal structure that some real-life collections do have (e.g., field structure, XML, citations)*". Today, more than ten years later, this statement no longer accurately reflects the breadth of test collections available. Several new domains have been explored, e.g. patent retrieval, expert search and retrieval in the cultural heritage domain. Also, some collections with internal structure have arisen, most prominently represented by the Initiative for the Evaluation of XML Retrieval (INEX). However, we claim that even these collections, although they reflect progress in the march to more collection complexity, have not yet reached the complexity level of operational collections. In the following, will give examples of some of the most complex academic collections and describe their shortcomings with respect to operational collections.

GeoCLEF a collection from 2008 offers only two modalities - the textual description and geographical coordinates. The geographical coordinates are not available as a separate modality, but need to be extracted from the text. Thus, the main focus of the evaluation tasks using this collection lies in the extraction of the geographical coordinates rather than the combination of the two modalities.

The ImageCLEF collections mostly contain two modalities - the images and textual description thereof (captions, titles, etc.). Still, they are not ideal to study complex multimodal collections. This is not only due to the small number of modalities, but also because the two modalities mostly contain the same overall information; e.g. the caption of an image that shows a cat most likely contains the word "cat".

The Living Lab track of CLEF 2015 offers a collection that brings an online shop scenario to the academic community [2]. The live setting of an ad hoc search task in an online toy store offers for each product a limited amount of textual data together with a lot of structured modalities such as the recommended age, the brand, the availability

and the price. We came across a similar setup in our transfer projects with the shopping app "Troffy", which allows users to search for products from different retailers in their area. In this project, an important aspect of the experimentation was to include the user's context such as his location and preferences. Since in the Living Lab no user information is provided it is not possible to personalize the result lists. In this case, the academic collection is as complex as in reality, but the query is a lot simpler. For many queries, the user preferences are however an important aspect. For example, consider the search for a tractor. The results should be quite different for a model vehicle fan than for a mother searching a present for her son.

The news domain has a long history in evaluation campaigns in IR; e.g. in the ad hoc track at CLEF. However, the focus so far was on multilingual retrieval of textual modalities. In a recent project, we worked on the noise canceling news feed application "Squirro" that collects documents from various sources such as public search engines, social networks and news feeds in general. The users of this application can create topics that not only consist of a textual description of the user's interests but also the user's preferences with respect to recency, language, popularity and source-quality of the documents. Again, we are not aware of academic test collections that mirror these aspects in comparable complexity.

3 Modality Categorization

We argue that in order to build test collections that reflect a desired complexity, it is important to start with a categorization of modalities; with the goal to uncover similarities between modalities. Methods developed for very specific modalities could then be generalized to these modality categories. Such a systematic structuring of the modalities also facilitates the uncovering of modalities that are inter-related.

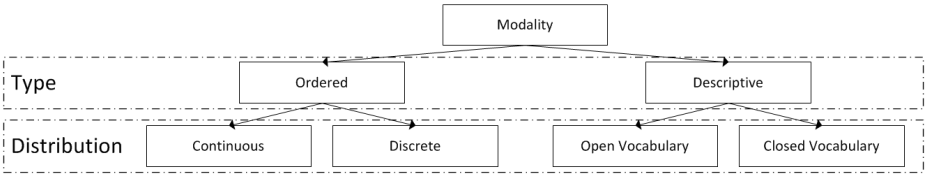


Fig. 2: Categorization of modalities into their types and distributions.

In order to come up with the categorization, we started with a huge set of different modalities that we have seen in evaluation campaigns and transfer projects of the past. The set included very specific descriptors for each modality, e.g. dates, ratings, geographical coordinates, terms, SIFT features, etc. We clustered the modalities that share similar characteristics into hierarchical modality categories. We identified the two top hierarchical levels of the categorization as shown in Figure 2. In the future, we assume that further levels will need to be introduced to handle more specific modality types.

We first distinguish between ordered and descriptive modalities. Ordered modalities such as ratings, dates, prices, number of clicks or likes have a natural order. Therefore, statements such as "which date is earlier" or "which item is more popular" can be made. We believe that the order of these modalities is important for the retrieval and needs to

be considered. In contrast, descriptive modalities such as terms in text retrieval and SIFT features in image retrieval do not have an order that contains relevant information for the retrieval. Terms usually are sorted alphabetically; however it is not important for the retrieval process if two terms start with adjacent letters.

At first, it seems that all numerical modalities are ordered modalities, while all textual modalities are descriptive. However, a modality that contains a group id may be descriptive even though the group id is numerical. The group ids are probably arbitrarily chosen without an order in mind, therefore the numerical order of the group ids is not important and hence it is a descriptive modality. On the other side, a modality describing the reading level of a book such as "Ages 4-8", "Ages 9-12" and "Young Adult" are textual, but also ordered.

We subdivide the descriptive modalities into open and closed vocabulary. An open vocabulary modality is a free text with a variable length as we know them from years of traditional text retrieval and many ad hoc retrieval collections. In a closed vocabulary modality the values that can be used are a predefined finite set; e.g. the binding of a book. For the ordered modalities we suggest a similar subdivision into discrete and continuous, since the methods need to be able to distinguish between a finite and an infinite amount of values.

Table 1: An excerpt of the modalities of the INEX Social Book Search collection with the associated type and distribution.

Name	Type	Distribution
Id	descriptive	closed vocabulary
Title	descriptive	open vocabulary
Binding	descriptive	closed vocabulary
Label	descriptive	closed vocabulary
Price	ordered	continuous
Reading Level	ordered	discrete
Release Date	ordered	continuous
No. of Pages	ordered	continuous

In Table 1, we use the INEX Social Book Search (SBS) collection [6] to show how the proposed categorization can be applied to a specific collection. The collection consists of ca. 2.8 million books from Amazon enriched with content information from Library Thing. The SBS collection is especially suited for such an assembly, since a lot of very different modalities are included.

4 Building Complex Collections

An ideal complex collection that mirrors the complexity of real-world IR applications should contain a large amount of modalities from different modality categories. Hence, it not only contains textual modalities from the category descriptive, open vocabulary as in traditional IR collections, but also non-textual and non-descriptive modalities such as images, ratings, prices and geographical coordinates. Moreover, both independent and inter-related modalities should appear in the collection. The independent modalities are important to provide preferably diverse information, while the inter-related modalities also coexist in the real-world collections and must be considered by the methods. The queries should likewise contain several modalities from different categories. Although it needs to be defined how each modality in the document contributes to the probability

of relevance, not each modality needs to have a corresponding modality in the queries. It is also possible to define their contribution based on query independent factors; e.g. a higher number of likes usually leads to a higher probability of relevance. From our experience, we saw that most operational collections contain a substantial textual part and approximately ten non-textual modalities. For some applications, the non-textual modalities must be considered in the retrieval, since they may contain key information that is required to fulfill the task requirements; e.g. the date in a recent news search. For others, they can be used in order to improve the retrieval performance.

5 Conclusions

In this paper, we argue that even though the complexity of collections increased in the last decade, academic test collections do not yet reach the level of complexity of operational collections in real-world applications. We propose a categorization of modalities, which serves two purposes: firstly, the large number of diverse modalities in operational collections makes it necessary to have unified methods for many kinds of modality types. This allows us to generalize the methods that have been developed for specific modalities to a modality category. Secondly, we suggest thinking about how to methodologically build academic test collections of higher, more realistic complexity by deriving the right mix of modalities from the categorization. This task requires an explicit reflection on the inter-dependence of the different modalities, and their characteristics. Still open is the handling of multi-dimensional modalities in the context of the presented categorization (e.g. geographical coordinates).

References

1. Agrawal, R., Ailamaki, A., Bernstein, P.A., Brewer, E.A., Carey, M.J., Chaudhuri, S., Doan, A., Florescu, D., Franklin, M.J., Garcia-Molina, H., et al.: The claremont report on database research. Communications of the ACM 52(6), 56–65 (2009)
2. Balog, K., Kelly, L., Schuth, A.: Head first: Living labs for ad-hoc search evaluation. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 1815–1818. ACM (2014)
3. Byström, K., Järvelin, K.: Task complexity affects information seeking and use. Information processing & management 31(2), 191–213 (1995)
4. Jones, K.S.: Readings in information retrieval. Morgan Kaufmann (1997)
5. Kekäläinen, J., Järvelin, K.: Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In: Proceedings of the 4th CoLIS conference. pp. 253–270 (2002)
6. Koolen, M., Kazai, G., Kamps, J., Preminger, M., Doucet, A., Landoni, M.: Overview of the inex 2012 social book search track. p. 77 (2012)
7. Saastamoinen, M., Kumpulainen, S., Järvelin, K.: Task complexity and information searching in administrative tasks revisited. In: Proceedings of the 4th Information Interaction in Context Symposium. pp. 204–213. ACM (2012)
8. Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Evaluation of cross-language information retrieval systems. pp. 355–370. Springer (2002)

A.2 Multimodal Social Book Search

Melanie Imhof[2], Ismail Badache, Mohand Boughanem
In *Linda Capellato, Nicola Ferro, Gareth Jones, Eric San Juan (Eds.)*, CLEF 2015 Labs Working Notes, Toulouse, France, September 8-11, 2015, Aachen: CEUR.

[2]Legal name change to Melanie Geiger as of August 11, 2017.

Multimodal Social Book Search

Melanie Imhof[1,2], Ismail Badache[3], and Mohand Boughanem[3]

[1] Université de Neuchâtel, Neuchâtel, Switzerland
[2] Zurich University of Applied Sciences, Winterthur, Switzerland
imhf@zhaw.ch
[3] IRIT - Paul Sabatier University, Toulouse, France
{badache, boughanem}@irit.fr

Abstract. Today's information retrieval applications have become increasingly complex. The Social Book Search (SBS) lab at CLEF 2015 allows evaluating retrieval methods on a complex search task with several textual and non-textual meta-data fields. The challenge is to incorporate the different information types (modalities) into a single ranked list. We build a strong textual baseline and combine it with a document prior based on social signals. Further, we include non-textual modalities in relation to the user preferences using random forest learning to rank. Our experiments show that both the social document prior and the learning to rank approach improve the search results.

Keywords: Relevance feedback, random forest, non-textual modalities, social signals, document prior.

1 Introduction

The suggestion track of the INEX Social Book Search (SBS) lab at CLEF 2015 challenges researchers to find methods to retrieve books as requested by real users of LibraryThing. The complex collection consists of more than 50 meta-data fields of real books from Amazon. Thus, the retrieval methods can not rely on the content of the books but only on meta-data such as product descriptions, user-generated reviews and ratings. The lab's evaluation metric nDCG@10 reflects the user behavior that in such an application only the first few "recommendations" are considered. Hence, to maximize the number of relevant books in the first few results both the textual description of the user's query and the user's profile including his personal catalog matter. For such a complex task with that many information types, methods are required to handle and fuse them into a single ranked list. Analogously to multimedia retrieval, we call these different information types "modalities". Hence, our goal in this complex task was to fuse a strong textual baseline approach with several non-textual and social modalities that respect the user preferences. Therefore, we established and refined a textual baseline using traditional information retrieval weighting schemes, blind relevance feedback, user-profile based filtering and example book based relevance feedback. We enhanced this with document priors based on social signals such

as the ratings and tags. Finally, we applied a random forest learning that further improves the results by including the non-textual modalities price and number of pages with respect to the user preferences.

2 Collection and Data

The SBS collection consists of 2.8 million book records from Amazon, extended with social meta-data from LibraryThing. Each book record is an XML file with fields like *isbn*, *title*, *review*, *summary*, *rating* and *tag*. The full list of fields is shown in Table 1.

Table 1. A list of all element names in the book descriptions.

tag name			
book	similarproducts	title	imagecategory
dimensions	tags	edition	name
reviews	isbn	dewey	role
editorialreviews	ean	creator	blurber
images	binding	review	dedication
creators	label	rating	epigraph
blurbers	listprice	authorid	firstwordsitem
dedications	manufacturer	totalvotes	lastwordsitem
epigraphs	numberofpages	helpfulvotes	quotation
firstwords	publisher	date	seriesitem
lastwords	height	summary	award
quotations	width	editorialreview	browseNode
series	length	content	character
awards	weight	source	place

There are 208 topics in the SBS 2015 lab. Each topic is a query that was posted on LibraryThing for a list of books and consists of five fields: *title, mediated_query, narrative, example* and *group*. Hereby, the *narrative* is the textual description of the query from which a hand-crafted *mediated_query* is derived. Further, the *example* field contains a list of books that the user has mentioned as positive or negative examples. Additionally, the personal LibraryThing *catalog* of each topic creator is available, which includes a list of the books the user has archived on LibraryThing along with his personal ratings.

The relevance assessments are based on the actual suggestions to the original query on the LibraryThing forum. The relevance values are weighted using a decision tree that includes reliability information such as whether the user who suggested a book has read it. The SBS 2015 topics are a subset of the topics used in 2014. However, the relevance assessments have been extended with additional book suggestions that have not been included in 2014.

3 Retrieval Models

3.1 Textual Models

As a basis for our methods we employ a textual baseline using a traditional information retrieval system. Therefore, we merge all textual fields of the document into a single textual index field. Further, we construct queries from the three topic fields *title, mediated_query* and *narrative* that are analogously merged into a single textual representation.

We extend the textual baseline with a query expansion (blind relevance feedback) based on Rocchio's method [4]. Therefore, the n most characteristic terms of the m top-ranked documents are added to the query. Hereby, the most characteristic terms of a document are chosen by the term weight determined by the weighting scheme.

As described in Section 2 the topics contain example books mentioned by the topic creators. We use the contents of the example books that are associated with a positive or neutral sentiment to expand the queries similar to the blind relevance feedback.

Additionally, we filter the books already read by the topic creator from the final ranked list, since this is a hard criterion in the relevance assessments [2]. Hereby, we determine the read books from the catalog of the topic creator as well as from the example books that are marked as read.

3.2 Social Signals-Based Model

Our approach consists of exploiting social data as a priori knowledge to take into account in the retrieval model. We combine textual relevance of a given document to a query and its social importance modeled as a prior probability.

3.2.1 Preliminaries

The social information that we exploit within the framework of our model can be represented by 3-tuple $< U, D, A >$ where *U, D* and *A* are finite sets of instances *Users*, *Documents* and *Actions*.

Documents. We consider a collection C=$\{D_1,\ D_2,...D_n\}$ of n documents, where each document D represents a book. We assume that a book can be represented by both a set of textual keywords D_w=$\{w_1,\ w_2,...w_y\}$ and a set of social actions A performed on the book, D_a=$\{a_1,\ a_2,...a_z\}$.

Actions. We consider a set A=$\{a_1,\ a_2,...a_m\}$ of m types of actions (signals) that users can perform on the documents. These actions represent the relation between users U=$\{u_1,\ u_2,...u_h\}$ and documents C.

3.2.2 Social Document Prior

We exploit textual models to estimate the relevance of a document to a query. Our approach combines the social document prior $P(D)$ and the relevance status value $RSV_{textual}(Q, D)$ between a query Q and document D as

$$RSV(D,Q) \overset{\text{rank}}{=} P(D) \cdot RSV_{textual}(Q,D) \tag{1}$$

$$\overset{\text{rank}}{=} P(D) \cdot \prod_{w_i \in Q} RSV_{textual}(w_i,D), \tag{2}$$

where w_i represents the terms in the query Q and $RSV_{textual}(w_i, D)$ can be estimated with different models such as BM25 and language model. The document prior $P(D)$ is a query-independent probability of seeing the document. It is useful for representing and incorporating other sources of evidence to the retrieval process. Our main contribution is a method to estimate $P(D)$ by exploiting social signals.

According to our previous approach [1], the priors are estimated by simply counting the number of actions performed on the documents. We assume that the signals are independent. Thus the general formula for calculating $P(D)$ is

$$P(D) = \prod_{a_i \in A} P(a_i), \tag{3}$$

where $P(a_i)$ is estimated using maximum-likelihood. It is calculated as

$$P(a_i) = \frac{\log(1 + |D_{a_i}|)}{\log(1 + |D_a|)}, \tag{4}$$

where $|D_{a_i}|$ is the number of actions of type a_i on document D and $|D_a|$ is the total number of actions on document D. Further, we use Dirichlet to smooth $P(a_i)$ by collection C to avoid zero probabilities. This leads to

$$P(D) = \prod_{a_i \in A} \left(\frac{\log(1 + |D_{a_i}|) + \mu \cdot P(a_i|C)}{\log(1 + |D_a|) + \mu} \right), \tag{5}$$

where $P(a_i|C)$, analogously to $P(a_i)$, is estimated using maximum-likelihood.

$$P(a_i|C) = \frac{\log(1 + \sum_{D \in C} |D_{a_i}|)}{\log(1 + \sum_{D \in C} |D_a|)} \tag{6}$$

In addition to considering social features separately as described above, we propose to incorporate the ratings as a measurement of the popularity and the reputation of a book. For this purpose, we use the Bayesian average (BA) of the ratings as a document prior, which takes into account how many users have rated a book. As more users rate the same book, the average becomes more reliable and less sensitive to outliers. Books that have many ratings are boosted with respect to books that have little ratings and books with high ratings are boosted more than books with low ratings. Hereby, the BA of a book is computed as

$$BA(D) = \frac{avg(D_r) \cdot |D_r| + \sum_{D' \in C} avg(D'_r) \cdot |D'_r|}{|D_r| + \sum_{D' \in C} |D'_r|}, \tag{7}$$

where avg is the average function and D_r is the set of ratings of document D. We note that considering logarithmic priors helps to compress the score range and thereby reduces the impact of the priors on the global score.

$$P_{BA}(D) = \frac{\log(1 + BA(D))}{\log(1 + \sum_{D' \in C} BA(D'))} \tag{8}$$

For books with no ratings this would result in a prior probability of zero. In order to avoid a multiplication by zero and thus ignoring the textual score, we use the Add-One smoothing method:

$$P_{BA}(D) = \frac{1 + \log(1 + BA(D))}{1 + \log(1 + \sum_{D' \in C} BA(D'))}. \tag{9}$$

3.3 Learning to Rank (Random Forests)

Besides the textual modalities, the SBS collection contains several non-textual modalities. We use random forests [3] to learn how to combine not only the different textual runs but also the non-textual modalities into a single ranked list. In particular, we use the price and number of pages of a book with respect to the user's preference as well as the book's ratings. Hereby, the user's preference is estimated by the average of the attributes in the topic creator's catalog; e.g. a user that only has short books in his catalog prefers short books. We assume that a user prefers to retrieve books that have similar attributes as the books he has read in the past. To achieve this, we add the difference between the average of the book prices in the topics creator's catalog and the price of the book to the random forest algorithm as an additional feature. Similarly, we add such a feature for the number of pages. For the ratings we assume that a general preference towards higher rated books exists for all users. Thus, we add the absolute average rating of a book as an additional feature to the random forests. To allow the algorithm to incorporate the significance of the average rating, we also add the number of ratings as a separate feature. The ratings are the ratings of the reviews of the book as well as the ratings in the catalogs of all topic creators. In order to combine these ratings, we divide the ratings in the catalogs by two, so that all ratings are in the same range.

4 Experimental Evaluation

We evaluated our approaches based on a series of experiments on the SBS 2015 task. Our goals in these experiments are to evaluate whether social signals (*tags* and *rating*) and other non-textual modalities can improve the search results.

4.1 Experimental Setup

For the textual baseline we used Lucene[4] for indexing and searching. We used the *EnglishAnalyzer*, which removes a small set of stopwords and stems terms

[4] https://lucene.apache.org/core/

using the Porter stemming algorithm. The weighting scheme used for most of the official runs is BM25 with $b = 0.75$ and $k_1 = 1.2$. We have also ran some experiments using language model with Dirichlet smoothing with $\mu = 2500$, however, we found that the BM25 achieved a better mean average precision (MAP) and nDCG@10 for the textual baseline. In order to validate the effectiveness of our approaches we used the topics and relevance assessments from SBS 2014.

For the blind relevance feedback, we experimented with the number of top-ranked documents used for the relevance feedback as well as with the number of terms extracted. However, we found that none of the combinations improve the textual baseline.

Since the topics from SBS 2015 are a subset of the topics from 2014, we were able to automatically add the example books from the 2015 topics to the corresponding topics in 2014. We found that expanding the queries with 35 terms extracted from the example books maximizes the nDCG@10 on the topics from 2014. Since we only have the example books for about 30% of the 2014 topics, the overall performance gain was not very big, however we have seen that the performance for the topics with example books has increased significantly.

Lucene does not provide a filter implementation that allows rejecting a list of documents, which is required to filter the read books. Thus we implemented our own filter with a similar concept as the Lucene's *FieldCacheTermsFilter*, which rejects all the documents that are not in the given list of documents.

As described in Section 3.2, we integrated social signals into the traditional textual model by re-ranking the results. The social signals are modeled as an a priori probability $P(D)$. We ran different experiments using all available social signals on the SBS collection (ratings, totalvotes, helpfulvotes, tags, etc.), but we found that the signals *tags* and *ratings*, estimated based on the formulas 5 and 9, achieved a better MAP and nDCG@10 compared to the other signals. We conducted our experiments in two ways: for Run3 and Run4 we multiplied $P(D)$ by the textual language model score; for Run5 and Run6, we combined the social signals score ($P(tags)$ multiplied by $P_{BA}(D)$) linearly with Run1, respectively with random forests trained with 100 trees. We set the smoothing parameter μ of formula 5 to 200, although more experiments will be necessary to get the best parameter. Experiments showed that the best combination parameter γ for the social score is 0.25 for Run5 and 0.2 for Run6.

We used RankLib[5] to train the random forests. For all the experiments, we left the default parameters unchanged except for the number of trees and the train metric which was set to nDCG@10. Unsurprisingly, increasing the number of trees results in a longer computation time, but also higher nDCG@10 values when training and testing on the SBS 2014 topics. However, with a higher number of trees the risk of over-fitting the data increases. The input for the random forests was built from the top 500 documents of six different textual runs together with the three non-textual modalities as described in Section 3.3. The textual runs were the textual baseline, the textual baseline with the read book filter, the textual baseline plus example based relevance feedback with and

[5] http://sourceforge.net/p/lemur/wiki/RankLib/

without filtering the read books and two runs using blind relevance feedback (total of 80 terms from 10 documents and total of 40 terms from 5 documents). Even though the blind relevance feedback runs on their own did not improve the textual baseline, we decided to add two runs using different parameters to the random forest in order to increase the variance of the input ranked lists. As training data we used the SBS 2014 topics and relevance assessments with the example books added from the 2015 topics. This is not an ideal situation, since the training data and the test data have an overlap. However, since we do not have example books for all the 2014 topics, we were not able to exclude the topics which are also in 2015 without losing the benefit of our example based relevance feedback.

For our participation to INEX SBS 2015 track, we built six runs by applying different configurations:

- **Run1**: Textual baseline using BM25 with example based relevance feedback using 35 terms and read book filtering.
- **Run2**: Random forests trained with 10 trees based on six textual runs and three non-textual modalities (price, number of pages and ratings).
- **Run3**: Run1 using language model combined with Bayesian average re-ranking based on *ratings*.
- **Run4**: Run1 using language model combined with re-ranking based on the *tags*.
- **Run5**: Run1 combined with re-ranking based on the *tags* and Bayesian average of *ratings*.
- **Run6**: Random forests trained with 100 trees based on six textual runs and three non-textual modalities (price, number of pages and ratings) combined with re-ranking based on the *tags* and Bayesian average of *ratings*.

In the next section we discuss the evaluation results of our official submission.

4.2 Results and Discussion

Table 2 summarizes our official results of SBS 2015 evaluated using nDCG@10 (Normalized Discounted Cumulative Gain), MRR (Mean Reciprocal Rank), MAP (Mean Average Precision) and R@1000 (Recall), whereas nDCG@10 is the official evaluation measure.

Table 2. Official results at SBS 2015. The runs are ranked according to nDCG@10.

Rank	Run	nDCG@10	MRR	MAP	R@1000	Train
1	Run6	0.186	0.394	0.105	0.374	yes
3	Run2	0.130	0.290	0.074	0.374	yes
8	Run5	0.095	0.235	0.062	0.374	no
10	Run3	0.094	0.237	0.062	0.374	no
11	Run4	0.094	0.232	0.061	0.375	no
21	Run1	0.082	0.189	0.054	0.375	no

We can see that the runs (Run2 and Run6) using random forest training far exceed the effectiveness of the runs using no training. During our experiments we saw that including the three non-textual modalities in the learning helps to increase the nDCG@10, which means that these modalities contain relevant information regarding the book suggestions.

Our textual baseline, although not submitted, achieves an nDCG@10 of 0.0768. Thus, the filtering together with the example based relevance feedback (Run1) significantly improves the nDCG@10 by 6.7% with a significance level of 58.4% calculated using the significance paired randomization test [5].

According to our experiments, Run3 and Run4 improve Run1 with language model (nDCG@10 of 0.0834) significantly (significance level $\alpha = 18.4\%$, respectively $\alpha = 15.3\%$). Using both the ratings and the tags (Run5) improves the effectiveness more than just using one of them. We note that the Run3 provides slightly better results in terms of MRR and MAP compared to Run4. One of the reasons of this is that the signal (rating) for Run3 that quantifies the reputation may be seen as expressing the engagement of a user who provides his explicit endorsement. For example, the document having more positive signals (ratings, likes, etc.) are more trustworthy than the ones that do not possess these social signals. If multiple users have found that the document is useful, then it is more likely that other users will find this document useful too. The social signals that quantify the popularity (number of reviews, tags, etc.) do not represent approval votes, as for example the reviews can be positive or negative, but they represent trend factors and a measure of information propagation. Therefore, a popular information always arouses the interest of the user.

The R@1000 is approximately the same for all runs, since they mostly are based on a re-ranking of Run1, for which we only retrieved the top 1000 documents. Since the learning based runs only used slight variations of Run1, they do not retrieve additional relevant documents beyond the top 1000 documents of Run1. For a recall-centric application, using a higher variety of runs as well as more documents per run would be beneficial.

5 Conclusions

In this paper, we described our participation to the suggestion track of the INEX SBS 2015 lab. We showed how to build a textual baseline and how to improve this using blind relevance feedback as well as example book based relevance feedback. Further, we proposed a method to include the social signals as a priori social knowledge that further enhanced the effectiveness of our system. The learning based approach using random forests, allowed us to incorporate the user preferences with respect to the book price and the number of pages as well as to combine the best aspects of the different variations of our textual methods.

So far, we did not use the anonymized user profiles from LibraryThing which would allow us to add additional ratings to the social model. Also we would like to test our learning approach with completely separated training and test datasets. Hence, we need to extract the example books for all the topics of SBS

2014. As a long term goal however, we think it is important to find methods that do not rely on learning. Although it might help to develop these by investigating the output of the random forests in order to better understand the modalities including their importance and their dependencies.

References

1. Badache, I., Boughanem, M.: Social priors to estimate relevance of a resource. In: IIiX Conference. pp. 106–114. IIiX'14, ACM, NY, USA (2014), http://doi.acm.org/10.1145/2637002.2637016
2. Bogers, T., Koolen, M., Jaap, K., Kazai, G., Preminger, M.: Overview of the inex 2014 social book search track. In: Conference and Labs of the Evaluation Forum. pp. 462–479 (2014)
3. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
4. Rocchio, J.J.: Relevance feedback in information retrieval. In: The SMART Retrieval System: Experiments in Automatic Document Processing. pp. 313–323. Prentice-Hall, Englewood Cliffs NJ (1971)
5. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 623–632. ACM, New York, NY, USA (2007)

A.3 BM25 for Non-Textual Modalities in Social Book Search

Melanie Imhof[3]

In *Krisztian Balog, Linda Capellato, Nicola Ferro, Craig Macdonald (Eds.)*, CLEF 2016 Labs Working Notes, Évora, Portugal, September 5-8, 2016, Aachen: CEUR.

[3]Legal name change to Melanie Geiger as of August 11, 2017.

BM25 for Non-Textual Modalities in Social Book Search

Melanie Imhof[1,2]

[1] Université de Neuchâtel, Neuchâtel, Switzerland
[2] Zurich University of Applied Sciences, Winterthur, Switzerland
imhf@zhaw.ch

Abstract. The Social Book Search (SBS) lab at CLEF 2016 provides a complex test collection that gives the opportunity to experiment with retrieval methods that combine various modalities in order to achieve the best possible ranked list. We show how the idea of being "characteristic", which is used as the core concept in most of the weighting schemes used for textual modalities, can be applied to non-textual modalities. Our approach re-defines BM25 for the three non-textual modalities found in the SBS collection: ratings, price and number of pages. A fuzzy query is constructed from the user preferences inferred from the user's catalog. The results are used to re-rank a textual baseline, which significantly improves the retrieval effectiveness.

Keywords: BM25, non-textual modalities, fuzzy query.

1 Introduction

The suggestion track of the INEX Social Book Search (SBS) at CLEF 2016 allows researchers to evaluate their methods on a multimodal collection with queries constructed from real LibraryThing user requests. For the books in the collection not only the book meta-information from Amazon (description, binding, number of pages, price etc.) is available but also user generated information such as book ratings. Also, the personal catalogs of the users are given and can be used to infer user preferences.

In our SBS 2015 participation [2], we found that the user preferences can be used to improve the retrieval effectiveness, by incorporating the books read by the users in a random forest based learning to rank approach. In this participation, we focus on taking the user preferences into account using a different approach. BM25 is a well known weighting scheme that has widely been used in text retrieval. It was originally developed for the English language but it has proven to be useful for other languages as well as for image retrieval [3]. We show how BM25 can be applied to the modalities ratings, price and number of pages. The BM25 scores of these non-textual modalities are then used to re-rank the textual baseline to significantly improve it.

2 Retrieval Models

2.1 Textual Models

Similar to our participation in 2015, we employ a textual baseline [2] as a basis for our methods. For the textual score, we merge all textual fields of the document into a single textual index field and construct queries from the two topic fields *title* and *request* that are analogously merged into a single textual representation. Further, we use the example books mentioned by the topic creators to expand the queries with the 35 most characteristic terms. Hereby, the most characteristic terms of the example books are computed by BM25.

Additionally, we filter the books already read by the topic creator from the final ranked list, since this is a hard criterion in the relevance assessments [1]. Hereby, we determine the read books from the catalog of the topic creator.

2.2 BM25 Model for Non-Textual Modalities

BM25 can be described in terms of how it combines three components; the feature frequency (*ff*), the document frequency (*df*) and the document length normalization component [5]. Although, it was originally developed for retrieval on English language text, it has generalized well to many related tasks, such as multilingual retrieval, multimedia retrieval and others. The *ff* and the *df* make sure that "characteristic" terms are weighed heavily. Hereby, a characteristic term is one that appears frequently in the document in consideration (*ff*) and rarely in the remainder of the collection (*df*). This concept of "being characteristic" is quite general and therefore applicable to other (non)-textual modalities [4]; i.e. bag of visual words in image retrieval, locations in geographical IR or timestamps in time-aware IR.

The retrieval status value (RSV) of document d_j w.r.t. query q when using BM25 is defined as

$$w(\varphi_k, d_j) := \frac{\mathrm{ff}(\varphi_k, d_j)}{k_1((1-b) + b\frac{l_j}{\Delta}) + \mathrm{ff}(\varphi_k, d_j)} \tag{1}$$

$$w(\varphi_k, q) := \mathrm{ff}(\varphi_k, q) \cdot \log\left(\frac{0.5 + N - \mathrm{df}(\varphi_k)}{0.5 + \mathrm{df}(\varphi_k)}\right) \tag{2}$$

$$\mathrm{RSV}_{\mathrm{BM25}}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) \cdot w(\varphi_k, q), \tag{3}$$

where k_1 is the *ff* saturation parameter and b is the document length normalization parameter. The k_1 parameter controls the amount an incremented *ff* will to contribute to the score. The notation used for the BM25 and its non-textual adaptions is described in Table 1.

For the suggestion track of the SBS lab at CLEF 2016, we adapt BM25 for three non-textual modalities the ratings, the price and the number of pages and use it to re-rank the textual baseline.

Table 1. Notation used for the BM25 for textual and non-textual modalities.

D	set of documents	$\Phi(d_j)$	set of features representing document d_j
N	number of documents	$\Phi(q)$	set of features representing query q
d_j	single document	$w(\varphi_k, d_j)$	weight of feature φ_k for document d_j
q	single query	$w(\varphi_k, q)$	weight of feature φ_k for query q
Φ	indexing vocabulary	$\text{ff}(\varphi_k, d_j)$	frequency of feature φ_k for document d_j
φ_k	single indexing feature	$\text{df}(\varphi_k)$	document frequency of feature φ_k
l_j	length of document d_j	Δ	average document length in number of tokens

Ratings For the ratings, we do not have a per-user query information, but we assume, that in general users will prefer books with higher ratings. Therefore, we define the query in the following way

$$\Phi(q) := \{1, 2, 3, 4, 5\} \tag{4}$$

$$\text{ff}(\varphi_k, q) := \varphi_k. \tag{5}$$

With this definition, each possible rating (1-5) is part of the query, however, a rating 5 is weighted 5 times heavier than a rating 1. The definition of the feature frequencies $\text{ff}(\varphi_k, d_j)$, document frequencies $\text{df}(\varphi_k)$ and document lengths l_j is analogous to the definition used for text. Hence, the *ff* is the number of times a given rating appears in a document, the *df* is the number of documents that contain a given rating and the document length is the number of ratings in a document.

Price For the price, we use the average price of the books that the user has already read $\Delta_p(q)$ as the query information. Since an exact match of the price is not meaningful, we use a fuzzy search with $\Delta_p(q)$ as the search parameter. We assume, that a user would also like books that are at most 20% cheaper and at most 30% more expensive than the average price of the books in his library. Although, we assume that generally a cheaper book is always acceptable, we still set a lower bound, because we assume that people tend to like similar kinds of books, that are usually in the same price range. The query's set of features and feature frequencies are defined as

$$\Phi(q) :=]0.8 \cdot \Delta_p(q), 1.3 \cdot \Delta_p(q)[\tag{6}$$

$$\text{ff}(\varphi_k, q) := \begin{cases} \frac{1.3 \cdot \Delta_p(q) - \varphi_k}{0.3 \cdot \Delta_p(q)} & \text{if } \varphi_k \geq \Delta_p(q) \\ \frac{\varphi_k - 1.2 \cdot \Delta_p(q)}{0.2 \cdot \Delta_p(q)} & \text{if } \varphi_k < \Delta_p(q). \end{cases} \tag{7}$$

For the definition of the *df*, we bin the prices into bins with a quadratically increasing width as shown in Figure 1. The bin index for the price p is defined as

$$\text{bin}(p) = \left\lfloor \frac{\sqrt{p}}{2} \right\rfloor. \tag{8}$$

This is based on the assumption, that with increasing prices, the tolerance for two book prices to be comparable is larger. The *df* is then defined as the number of documents with a price in a given bin. Since a book only has a single value for the price, the *ff* and the document length are always 1.

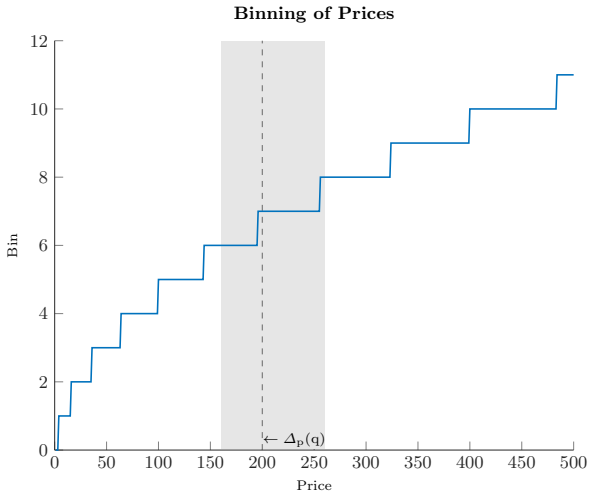


Fig. 1. Binning of prices to compute the document frequencies as well as the fuzzy query range with the average price of the books of the topic creator $\Delta_p(q)$.

Number of Pages For the number of pages of the books, we defined the *ff*, *df* and the document length as well as the query, analogous to the price.

3 Experimental Evaluation

Our goal in the experiments is to use the information present in the non-textual modalities to refine the result lists so that they reflect the users preferences.

3.1 Experimental Setup

For the textual baseline we used Lucene[3] for indexing and searching. For all modalities we used BM25 with a document length normalization parameter $b = 0.75$ and a *ff* saturation parameter $k_1 = 1.2$. For the textual modalities, we used the built-in *EnglishAnalyzer*, which removes a small set of stopwords and stems terms using the Porter stemming algorithm. For the re-ranking, we used a linear

[3] https://lucene.apache.org/core/

combination of the scores from the different modalities. Hereby, the weights of the linear combination sum up to one.

$$\mathrm{RSV}_{\mathrm{BM25}} = \alpha \cdot \mathrm{RSV}^{\mathrm{text}}_{\mathrm{BM25}} + \beta \cdot \mathrm{RSV}^{\mathrm{rating}}_{\mathrm{BM25}} + \gamma \cdot \mathrm{RSV}^{\mathrm{price}}_{\mathrm{BM25}} + \delta \cdot \mathrm{RSV}^{\mathrm{pages}}_{\mathrm{BM25}} \quad (9)$$

In order to validate the effectiveness of our approaches and to find the optimal re-ranking parameters, we used the topics and relevance assessments from SBS 2015.

For our participation to INEX SBS 2016 track, we built six runs by applying different configurations (the re-ranking parameters equal to zero are omitted):

- **Run1**: Textual baseline using BM25 with example based relevance feedback using 35 terms and read book filtering with re-ranking parameters: $\alpha = 1$.
- **Run2**: Textual baseline re-ranked with a query-independent BM25 model for ratings with re-ranking parameters: $\alpha = 0.7818, \beta = 0.2182$.
- **Run3**: Textual baseline re-ranked with a user catalog based BM25 model for the number of pages with re-ranking parameters: $\alpha = 0.3118, \delta = 0.6882$.
- **Run4**: Textual baseline re-ranked with a user catalog based BM25 model for the price with re-ranking parameters: $\alpha = 0.2332, \gamma = 0.7668$.
- **Run5**: Textual baseline re-ranked with a user catalog based BM25 model for the price and the number of pages with re-ranking parameters: $\alpha = 0.2225, \gamma = 0.3033, \delta = 0.4742$.
- **Run6**: Textual baseline re-ranked with a query-independent BM25 model for ratings and a user catalog based BM25 model for the price and the number of pages with re-ranking parameters: $\alpha = 0.265, \beta = 0.045, \gamma = 0.225, \delta = 0.465$.

In the next section we discuss the evaluation results of our official submission.

3.2 Results and Discussion

Table 2 summarizes our official results of SBS 2016 evaluated using nDCG@10 (Normalized Discounted Cumulative Gain), MRR (Mean Reciprocal Rank), MAP (Mean Average Precision) and R@1000 (Recall), with nDCG@10 being the official evaluation measure.

The submitted runs using all non-textual modalities to re-rank the textual baseline (Run6), the run using the price and the number of pages (Run5) as well as the run using the number of pages (Run3) significantly improve the nDCG@10 over the textual baseline (Run1). The significance is computed using a paired randomization test [7] with significance level $\alpha = 5\%$. Using just the number of pages (Run3) leads to the highest nDCG@10 amongst our submitted runs. Using just the ratings for the re-ranking (Run2) increases the nDCG@10 over the textual baseline, although not significantly. Our re-ranking with the scores calculated based on the price (Run4) does not help to find a better ranked list.

To further analyze the results, we also evaluated the performance of the non-textual modalities on their own. Therefore, we used the documents retrieved

Table 2. Official results at SBS 2016. The runs are ranked according to nDCG@10.[4]

Rank	Run	Features	nDCG@10	MRR	MAP	R@1000
25	Run3	text, pages	0.0674	0.1512	0.0472	0.2556
26	Run6	text, price, pages, ratings	0.0667	0.1499	0.0462	0.2556
27	Run5	text, price, pages	0.0665	0.1442	0.0461	0.2556
30	Run2	text, ratings	0.0584	0.1332	0.0419	0.2556
31	Run1	text	0.0561	0.1251	0.0396	0.2556
32	Run4	text, price	0.0542	0.1114	0.0386	0.2556

with the textual baseline and ranked them only based on the score of each non-textual modality. This will not lead to a fully textual baseline independent ranked list (e.g. the recall will not change), however it gives an indication how well they would perform on their own. Using this approach the nDCG@10 for the ratings is 0.0206, for the price it is 0.0258 and for the number of pages 0.0135. Surprisingly, we see that the price on its own results in the highest nDCG@10, although this is not reflected in the runs that combine the non-textual modalities with the textual baseline. We also trained the weights for modalities with the relevance assessments for the 2016 task, and found, that with the optimal weights, the textual baseline can also be improved by taking the price into account. Hence, the weights chosen based on the 2015 task, are not optimal. Nevertheless, the nDCG@10 for the runs using the number of pages (0.0706) and the ratings (0.0647) using optimal weights is still higher than for the run with the price (0.0596). This shows, that either there is a higher information overlap between the price and the textual modality than between the other modalities and the text, or the linear combination merging is not as effective for the price as for the others.

4 Conclusions

In this paper, we described our participation to the suggestion track of the INEX SBS 2016 lab. We investigated how the weighting scheme BM25 can be applied to non-textual, continuous modalities. Therefore, we proposed a method to discretize the continuous modalities in order to define a document frequency and a fuzzy query that takes into account that the query does not require an exact match. By using our approach on the ratings, prices and number of pages, we showed that the effectiveness of the system can be significantly increased over the textual baseline using a simple linear score combination. However, the performance of our random forest based learning to rank approach from 2015, can not be reached.

Our experiments, have shown that the merging the scores of the prices with the textual scores leads to a smaller improvements as could be expected based

[4] We have underlined any statistically significant differences in performance according to nDCG@10 to the textual baseline (Run1) resulting from a paired randomization test [7] (significance level $\alpha = 5\%$).

on the performance of the non-textual modalities individually. So far, we did not yet investigate the merging in more depth. It is possible that a different merging method could improve the merging with the price. For example, we could use a non-linear combination of the scores, or a per-query normalization strategy, like the z-score [6], to avoid that the per-query optimal weights are far apart.

Further, we would like to investigate if the function used for the fuzzy search is the best possible. We could for example use different parameters or a non-linear falloff of the weighting.

So far, we approximated the user preferences by the average price and number of pages of the books read by the user. However, it could also be possible to construct the query such that each price and number of pages is part of the query and therefore the loss of information due to the averaging is avoided.

References

1. Bogers, T., Koolen, M., Jaap, K., Kazai, G., Preminger, M.: Overview of the inex 2014 social book search track. In: Conference and Labs of the Evaluation Forum. pp. 462–479 (2014)
2. Imhof, M., Badache, I., Boughanem, M.: Multimodal social book search. In: Sixth International Conference of the CLEF Association, CLEF (2015)
3. Moulin, C., Barat, C., Ducottet, C.: Fusion of tf. idf weighted bag of visual features for image classification. In: Content-Based Multimedia Indexing (CBMI), 2010 International Workshop on. pp. 1–6. IEEE (2010)
4. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc (2009)
5. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management 24(5), 513–523 (1988)
6. Savoy, J., Berger, P.Y.: Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers, chap. Monolingual, Bilingual, and GIRT Information Retrieval at CLEF-2005, pp. 131–140. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
7. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 623–632. ACM, New York, NY, USA (2007)

A.4 A Study of Untrained Models for Multimodal Information Retrieval

Melanie Imhof[4], **Martin Braschler.**
In *Information Retrieval Journal*, 21(1), 81-106, 2018

[4]Legal name change to Melanie Geiger as of August 11, 2017.

A Study of Untrained Models for Multimodal Information Retrieval

Melanie Imhof · **Martin Braschler**

Abstract Operational multimodal information retrieval (IR) systems have to deal with increasingly complex document collections and queries that are composed of a large set of textual and non-textual modalities such as ratings, prices, timestamps, geographical coordinates, etc. The resulting combinatorial explosion of modality combinations makes it intractable to treat each modality individually and to obtain suitable training data. As a consequence, instead of finding and training new models for each individual modality or combination of modalities, it is crucial to establish unified models, and fuse their outputs in a robust way. Since the most popular weighting schemes for textual retrieval have in the past generalized well to many retrieval tasks, we demonstrate how they can be adapted to be used with non-textual modalities, which is a first step towards finding such a unified model. We demonstrate that the popular weighting scheme BM25 is suitable to be used for multimodal IR systems and analyze the underlying assumptions of the BM25 formula with respect to merging modalities under the so-called raw-score merging hypothesis, which requires no training. We establish a multimodal baseline for two multimodal test collections, show how modalities differ with respect to their contribution to relevance and the difficulty of treating modalities with overlapping information. Our experiments demonstrate that our multimodal baseline with no training achieves a significantly higher retrieval effectiveness than using just the textual modality for the social book search 2016 collection and lies in the range of a trained multimodal approach using the optimal linear combination of the modality scores.

Keywords Multimodal Information Retrieval · BM25 · Raw-Score Merging Hypothesis

Melanie Imhof
Université de Neuchâtel, Neuchâtel, Switzerland
Zurich University of Applied Sciences, Winterthur, Switzerland
E-mail: imhf@zhaw.ch

Martin Braschler
Zurich University of Applied Sciences, Winterthur, Switzerland
E-mail: bram@zhaw.ch

1 Introduction

The academic discipline that we term today "information retrieval" (IR) goes back, though opinions vary, to at least the seminal position paper by Vannevar Bush [6]. In the ensuing roughly 70 years of work, some mechanisms have been introduced early on, but have persisted and proven versatile since then; e.g. the formulae that govern the ranking of retrieved documents. Amongst these are some of the most popular weighting schemes for (textual) retrieval, which can all be described in terms of how they combine three main components; the term frequency (*tf*); i.e. how often a term appears in a given document, the document frequency (*df*); i.e. in how many documents a term appears and a document length normalization component. Originally developed for retrieval on English language text, these weighting schemes have generalized well to many related tasks, such as multilingual retrieval [31], multimedia retrieval [29] and others.

Today, we have to deal with increasingly complex document collections and queries [18] that no longer just consist of textual modalities but also of a large set of non-textual modalities such as visual words in image retrieval [43], locations in geographical IR [26] or timestamps in time-aware IR [22]. This is particularly true in enterprise search, domain-specific IR and many real IR applications, where it is not an option to simply ignore or discard entire modalities. Therefore, we claim that it becomes crucial to treat the modalities with unified methods instead of finding new approaches for each new modality or train a new model for every combination of modaltities. In this paper, we discuss the underpinnings of weighting schemes for textual retrieval and show how they can be applied or adapted methodically to non-textual modalities, such as ratings of books and geographical coordinates, which we understand as the first step into finding a unified model.

As a contribution towards establishing best practices for the integration of many modalities into an IR application, we demonstrate that BM25 is a suitable weighting scheme outperforming its alternatives to be used on non-textual modalities and to merge them under the so-called raw-score merging hypothesis by checking the assumptions underlying the BM25 formula. Being able to merge the modalities under the raw-score merging hypothesis with little or no training is particularly important due to the limited generalizability of suitable test collections and training data.

We start by considering an "ideal" robust approach, which is based on term sampling in order to correct the differences in average document length, which is one of the most obvious collection statistics. Then, we prove that there are cases, where BM25 can be interpreted as being identical to this sampling based approach. Using the sampling approach, we can further correct the difference between the variance of the document lengths. Along the investigation of the sampling approach, we further analyze the *tf* saturation parameter k_1 of BM25 and explain its significance for non-textual modalities. Finally, we present experiments on the effectiveness of merging the results of the individual modalities into a unified multimodal result. We contrast our approach, which avoids learning, with an "optimized" baseline and find encouraging results.

The remainder of this paper is structured as follows. Section 2 outlines the anatomy of multimodal IR systems and describes the challenges faced when dealing with com-

plex multimodal collections. We then demonstrate that BM25 is a suitable weighting scheme in multimodal IR systems w.r.t. document length normalization (Section 4). Section 5 describes how BM25 can be used for non-textual modalities by redefining the three main components of the weighting scheme. A sampling based BM25 approach is proposed in Section 6, which allows us to prove that BM25 fulfills the raw-score merging hypothesis w.r.t. the average document length and the variance of document lengths. In Section 7, we describe the multimodal test collections that we use for evaluation, followed by the experiments and the discussion of their results. Section 8 concludes this paper and discusses future work.

2 The Anatomy of a Multimodal IR System

2.1 Anatomy

In a multimodal IR system, both the documents as well as the queries consist of several modalities. Figure 1 shows an explanatory excerpt of four of the modalities of the documents in the social book search (SBS) collection used in the SBS lab at the CLEF evaluation forum [20]. The documents $(d_1, d_2, ..., d_D)$ consist of the modalities: book title, reviews, binding and ratings, each of which can be treated as a bag of features. Hereby, d_j^m is the bag of features of modality m of document d_j. The query both contains explicit and implicit modalities; i.e. the textual description of the request is explicit, while other information such as acceptable languages and ratings of the books are implicit. A more detailed description of the collection is given in Section 7.1.2. The queries in the SBS task are not particularly complex. In general, information needs embed several implicit and explicit modalities.

Fig. 1: Excerpt of four modalities of a sample document (denoted d_j) in the SBS collection.

1:	Title	$d_j^1 = \{\text{Skylar, in, Yankeeland}\}$
2:	Reviews	$d_j^2 = \{\text{Delightful, The, is, the, best, McDonald, has, done, in, a, decade}\}$
3:	Ratings	$d_j^3 = \{\text{5,1,3}\}$
4:	Binding	$d_j^4 = \{\text{Hardcover}\}$

During retrieval, weighting schemes define the retrieval score (retrieval status value $\text{RSV}(q, d_j^m)$) of modality m of document d_j w.r.t. query q. The retrieval scores allow producing a ranked list for each modality according to the estimated probabilities of relevance, although the retrieval scores are not necessarily probability values but are order-preserving w.r.t the probabilities of relevance [34]. These ranked lists of all the modalities, similarly to multilingual retrieval, need to be merged into a single ranked list. Hence, a function f has to be found to compute the retrieval score for each document including the retrieval scores of all modalities

$$\text{RSV}(q, d_j) = f(\text{RSV}(q, d_j^1), \text{RSV}(q, d_j^2), ..., \text{RSV}(q, d_j^M)), \tag{1}$$

where M is the number of modalities.

Evaluation has a strong tradition in IR, since information is hard to be defined in general [9]. A crucial part of an IR evaluation is the availability of a suitable test collection. However, most of the existing test collections are not representative for multimodal IR systems and it is clearly not practical to create a test collection that covers all possible modalities and their combinations [18]. We are convinced that in order to improve and broaden the applicability of multimodal IR, a generalizable method to deal with complex collections with a large amount of very different modalities is crucial. Therefore, we claim that we need a unified weighting model for all types of modalities in order to avoid a lot of effort to come up with a new model for every modality type. Further, a merging strategy that works with little or no training is necessary, both because training can become very complex for a large amount of modalities and because in practical applications training data is not always available [18].

2.2 Challenges

A multimodal IR system as described in this Section comes with several challenges that need to be solved in order to effectively use all the modalities. On the pursuit of a suitable weighting scheme for non-textual modalities, we can analyze the most popular textual weighting schemes. These can all be described in terms of how they combine three main components; the term frequency (*tf*), the document frequency (*df*) and the document length normalization component [37]. Looking at these three components, we can understand their respective roles as follows: The first two components make sure that "characteristic" terms are weighed heavily. Hereby, a characteristic term is one that appears frequently in the document in consideration (term frequency) and rarely in the remainder of the collection (document frequency). These terms are suitable to distinguish a document from other documents in the collection. The third component, the document length normalization, was introduced to ensure no documents of a particular length are favored in an undue way, offsetting the increasing probability to observe terms frequently simply due to the verbosity of the document.

The concept of "being characteristic", embodied through *tf* as well as *df*, is quite general and therefore applicable to other non-textual modalities [34]. One basically needs to check the assumption that an "unforeseen" local frequency of a feature hints at relevance. For non-textual modalities, the "term frequency" is usually referred to as "feature frequency" (*ff*). In the remainder of this paper, we will use the two expressions interchangeably. In Section 5, we show how we can define the *tf* and *df* for the two non-textual modalities ratings and geographical coordinates.

When analyzing the requirements of a weighting scheme for effective merging of ranked lists, usually the raw-score merging hypothesis is considered. The raw-score merging hypothesis describes that similarity values are directly comparable if they are produced from similar search engines and underlying collections with similar statistics [3, 21, 38, 39]. In Appendix A, we show that it is favorable to use the same weighting scheme for all modalities when using raw-score merging. However, already textual modalities often invalidate the raw-score merging hypothesis w.r.t.

to the similar collection statistics. For non-textual modalities, this is usually even more severe, since they do not follow the language statistics. Therefore, we propose a sampling-based approach in Section 6 to eliminate the differences in average and variance of document lengths and show that BM25 satisfies the derived properties, which makes it a viable weighting scheme for raw-score merging.

We can summarize the challenges of building multimodal IR systems discussed in this paper as follows.

1. Adapt BM25 to non-textual modalities
   (a) Define *tf*, *df* and document length
   (b) Validate generalizability of document length normalization
2. Evaluate merging strategies (raw-score merging hypothesis)
3. Validate suitability of BM25 for raw-score merging
4. Evaluate effectiveness of the approach

3 Related Work

Much work has been done using additional non-textual modalities in order to improve the retrieval effectiveness of textual IR systems. A famous example is the query-independent modality PageRank [4] and it is now an established practice to use modalities such as URL-type, anchor text and link indegree in retrieval of Web data [11, 15, 25]. A lot of other retrieval research sub-fields such as geographical IR [26], image retrieval [43], XML retrieval [19] and living labs [40] provide and use a large range of different modalities in order to optimize the retrieval results. Hereby, the additional modalities are often no longer query-independent, but also explicitly or implicitly (e.g. inside a user profile) part of the query. In contrast to this paper, most of these models have been developed for a specific modality and the generalization to other modalities was not a focus.

For non-textual modalities the document length normalization is particularly important, since items usually have large variances in the "length" of their content in terms of those modalities. Looking towards textual retrieval, a number of efforts investigating the role of document length in ranking textual documents exist. Generally, consensus is that including document length normalization in weighting schemes tends to improve the retrieval performance [1, 8, 23, 41]. The weighting scheme Lnu.ltn [41] is explicitly based on the idea of revisiting the cosine document length normalization of TF.IDF. Singhal *et al.* [41] estimate the likelihood of relevance and the likelihood of retrieval for all document lengths and improve the document length normalization by tilting the slope of the likelihood of retrieval in order to better match the slope of the likelihood of relevance. This tilt of the slopes then results in the new improved "pivoted document length normalization scheme". Investigations of the document length normalization of the BM25 weighting scheme have shown that it fails when documents are very long [24] and that choosing the right document length normalization parameter b in BM25 can increase the retrieval performance by 22% [8]. In XML retrieval, document length normalization is particularly important, since the retrievable items (XML elements) have a great variety in

length. Kamps *et al.* [19] revisit the role of language model document length normalization in the context of XML retrieval. Amongst others, they found that a combination of restricting the minimal size of the XML elements and length priors results in a higher effectiveness.

Oftentimes multiple intermediate result lists, one per modality, are produced when matching on multimodal collections. The problem of merging multiple ranked lists into a single ranked list is known from multilingual, multimedia and distributed retrieval. Fox and Shaw [14] propose different strategies to fuse the scores; e.g. the sum of the scores or the maximal score. However, as Callan *et al.* [7] point out, the scores might not be directly comparable, due to the different ranges of the scores.

The merging problem is very prominently studied in the multimedia IR community. Depeursinge and Müller show that 62% of the ImageCLEF working notes deal with data fusion, their detailed analysis reveals that, similar to all the other domains, the most used fusion strategy is a linear combination of the scores [13]. Mostly the weights of the linear combination are either found manually or based on training data. Wilkins *et al.* [47] however describe a method to automatically determine query-dependent modality weights using the score distribution of visual and textual modalities used in the context of video retrieval. Another unsupervised method to fuse multiple ranked lists for medical IR is presented by Mourão *et al.* [28]. Their fusion method combines the inverse rank approach of reciprocal rank fusion [10] with the number of times a document appears on a rank and achieves a high precision. The unsupervised methods proposed in this paper try to fuse the modality scores without any weights, which we claim, is possible when treating all modalities with the same model.

Robertson *et al.* [35] show the problems that arise when using a linear combination of the scores obtained from scoring multiple textual fields individually using BM25. The most important reason why this leads to poor retrieval effectiveness is the non-linear treatment of the term frequencies. This non-linearity is desirable for individual fields, since the information gain on observing a term for the first time is greater than the information gained on subsequently seeing the term. However, when using a linear combination of scores this non-linearity breaks. Therefore, Robertson *et al.* [35] propose a method that uses a linear combination of the term frequencies instead of using a linear combination of the scores, with which the problem can be solved. The term frequency is not the only point that has to be considered in a retrieval setup with multi-field documents, also the document length and the parameters of the weighting scheme have to be questioned. When computing a score for each individual field the weighting scheme parameters, in BM25 the *tf* saturation parameter k_1 and the document length normalization parameter b have to be optimized for each field individually, which results in a huge number of optimization parameters. With the method suggested by Robertson *et al.* [35] only two weighting scheme parameters have to be optimized. The suggested method also leads to substantially different term frequencies, since they replicate the content of the fields with the weight, the authors therefore suggest to use an adapted k_1 that is a scaled version of the original k_1 by the ratio between the original and the resulting average term frequency. For our methods, we use the idea of scaling k_1 when sampling all modalities to the same length.

4 Validating the Generalizability of Document Length Normalizations

Similar to traditional textual retrieval, special care needs to be taken to handle varying document lengths for non-textual modalities as well. Non-textual modalities can have large variances in document lengths. In order to find a suitable weighting scheme for non-textual modalities, we analyze four of the most known weighting schemes with respect to their document length normalization robustness.

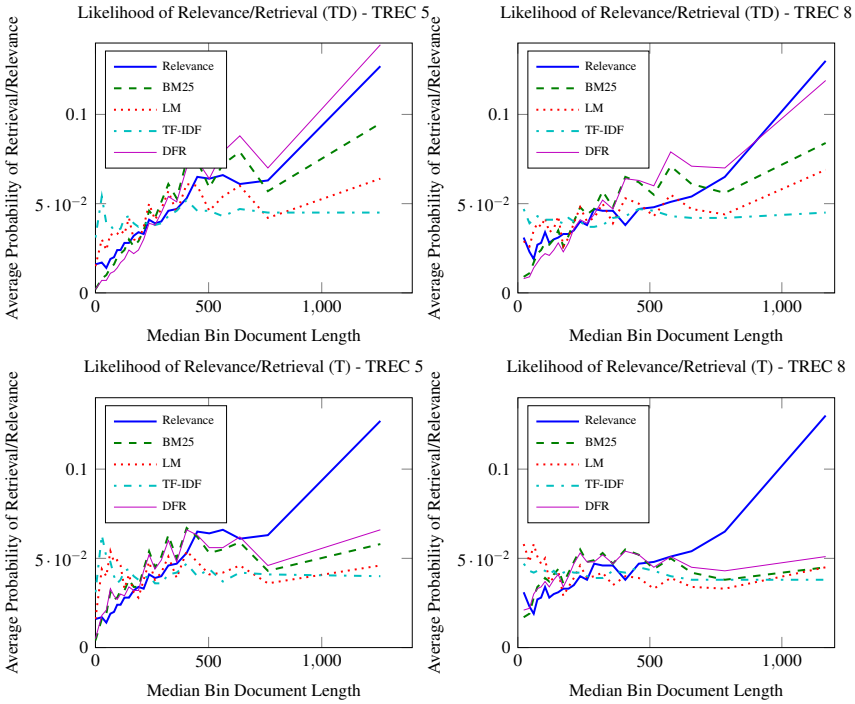


Fig. 2: Likelihood of Retrieval/Relevance for the TREC5 / TREC 8 data using 24 bins and the original weighting schemes.

The experiments are conducted using the TREC 5 ad hoc collection [45] and the TREC 8 ad hoc collection [46]. The choice of these rather classic test collections is motivated as follows: TREC 5 includes the Federal Register sub-collection that contains very lengthy documents, resulting in a high variance w.r.t. the document lengths of the collection. TREC 8 has been chosen due to its use in earlier literature about document length normalization [8,23,24], however has a smaller variance w.r.t. the document lengths than TREC 5 and we therefore expect that the effects of the document length component to be less pronounced. We used the full datasets and automatically generated queries from the topic title (T) and the description (D).

We examine the document length normalization and its impact on the retrieval effectiveness using the idea of Singhal *et al.* [41]. They calculate the likelihood of retrieval and relevance for each document length and employ these to adjust the document length normalization. We use these two likelihoods to visualize the effectiveness of the document length normalization of the four weighting schemes in study. To compute these likelihoods the documents are binned by their length. For each bin, the likelihood is defined as the ratio between the number of relevant/retrieved documents and the total number of documents in the bin. We then plot the likelihoods against the median document length in the bins.

Figure 2 shows the likelihood of relevance (bold line) and the likelihood of retrieval for all the weighting schemes for the TREC 5 and TREC 8 collections. The documents are divided in to 24 bins. As shown in this figure, longer documents have a higher probability of being relevant and retrieved. For both TREC 5 and TREC 8 as well as the long (TD) and short topics (T), BM25 and DFR match the likelihood of retrieval the best and we conclude that BM25 is able to handle large variances in document length. Since the document length normalization of BM25 is robust, it is suited to be used with non-textual modalities without any restriction regarding the variance of document lengths. Note that we did not include weighting scheme extensions, such as BM25L [24], that specifically target the robustness of the document length normalization, since they usually come with further assumptions regarding the statistics of the modalities.

5 BM25 Model for Non-Textual Modalities

5.1 BM25

Our experiments to validate the raw-score merging hypothesis and the generalizability of the document length normalization show that BM25 both works best for the raw-score merging and is amongst the most robust weighting schemes with highly varying document lengths. Therefore, we will focus our work with non-textual modalities on BM25.

Let us explore multimodal document collections such as used in GeoCLEF [26] or in the social book search lab [2]. In these collections, documents are no longer just represented by only a set of terms (textual features) but also by geographical features or by book ratings that further describe the documents.

Table 1: Notation used for the BM25 for textual and non-textual modalities.

D	set of documents	$\Phi(d_j)$	set of features representing document d_j
N	number of documents	$\Phi(q)$	set of features representing query q
d_j	single document	$w(\varphi_k, d_j)$	weight of feature φ_k for document d_j
q	single query	$w(\varphi_k, q)$	weight of feature φ_k for query q
Φ	indexing vocabulary	$\mathrm{ff}(\varphi_k, d_j)$	feature frequency of feature φ_k for document d_j
φ_k	single indexing feature	$\mathrm{df}(\varphi_k)$	document frequency of feature φ_k
l_j	length of document d_j	$\mathrm{cf}(\varphi_k)$	collection frequency of feature φ_k
Δ	average document length	$\mathrm{RSV}(q, d_j)$	retrieval status value of document d_j w.r.t. query q

In this Section, we first re-capitulate BM25 for a textual modality and then show how its idea can be adapted to geographical coordinates and to book ratings. Table 1 shows the notations used for BM25 as well as for its non-textual adaptions.

The retrieval status value (RSV) of document d_j w.r.t. query q when using BM25 can be written as an inner product

$$w(\varphi_k, d_j) := \frac{\mathrm{ff}(\varphi_k, d_j)}{k_1((1-b)+b\frac{l_j}{\Delta})+\mathrm{ff}(\varphi_k, d_j)} \tag{2}$$

$$w(\varphi_k, q) := \mathrm{ff}(\varphi_k, q) \cdot \log\left(\frac{0.5+N-\mathrm{df}(\varphi_k)}{0.5+\mathrm{df}(\varphi_k)}\right) \tag{3}$$

$$\mathrm{RSV}_{\mathrm{BM25}}(q, d_j) := \sum_{\varphi_k \in \Phi(q)\cap\Phi(d_j)} w(\varphi_k, d_j) \cdot w(\varphi_k, q), \tag{4}$$

where k_1 is the *tf* saturation parameter and b is the document length normalization parameter.

For its document length normalization, BM25 [34, 36] assumes a standard length of a document represented by the average document length. Hence, an author can decide to write a document longer or shorter than the standard length. Robertson [34,36] describes two cases why an author might decide to write a long document; either the author is more verbose than others or the author covers a larger scope. The verbosity assumption would lead to a division of the *tf* values by the document length. The scope assumption points to an opposite course of action, hence not dividing at all. Normally, the reason for a longer document is a combination of the two, thus Robertson's normalization balances the two using a tuning parameter b. Robertson proposed to use the number of tokens in a document as the document length, although he pointed out that BM25 should lead to similar results with slightly different definitions of the document length such as the number of characters. When using BM25 for non-textual modalities, it needs to be considered if this assumption holds true for those as well.

Since BM25 was originally designed for textual modalities, the question arises if its concept depends on the Zipfian distribution of the modalities as it is the case for natural language features. In particular the heuristic definition of the inverse document frequency (*idf*) can be motivated by the Zipf's law. However, over the years people have come up with several other interpretations on why the *idf* works as well as it does. For example, the theories that the *idf* corresponds to the probability of a term appearing in a document or to Shannon's information theory as described by Robertson [33]. It therefore is unclear how much the performance of BM25 depends on the Zipfian distribution of the modalities. Although we will not further investigate this question in this paper, we however assume that BM25 is generalizable to non-textual modalities with any distribution as long as the *tf* and *idf* can be defined in a way that the characteristic features still emerge.

Apart from the open question how well BM25 generalizes to modalities with a non-Zipfian distribution, it has been shown that BM25 is indeed generalizable to modalities with a Zipfian distribution such as a bag-of-visual-words in multimedia retrieval [49]. Also the distribution of the modalities we use in our experiments satisfy Zipf's law. In the case of the GeoCLEF collection, which we use for our ex-

periments with geographical coordinates, the coordinates have a Zipfian distribution, since they are extracted from the locations mentioned in the textual representation. Further, we analyzed the distribution of the ratings in the social book search collection and realized that they also have an approximate Zipfian distribution. It seems that the distribution of the ratings in this collection is not an exception, but appears to be a general phenomenon [12, 32, 48].

The *tf* saturation is parametrized by k_1 and makes sure that an increase of a high *tf* will contribute less to the score than an increase of a smaller *tf*. The higher the k_1 value, the more will an increase of a high *tf* contribute to the score, i.e. the saturation is less pronounced with high k_1 values.

The optimal choice of k_1 is not simple to make and also depends on the collection [8]. Further, k_1 needs to be adjusted if documents are replicated [35]. When replicating the content of all the documents (concatenate each document with itself; all documents have twice the length), neither the informativeness of a single document is changed nor the relevance of the documents to a particular query changes. However, if k_1 is not adjusted the BM25 weighting scheme will not lead to the same ranked list as without the replication. The BM25 weight for document d'_j that are replicated x-times is

$$w(\varphi_k, d'_j, k_1) = \frac{x \cdot \mathrm{ff}(\varphi_k, d_j)}{k_1((1-b) + b\frac{x \cdot l_j}{x \cdot \Delta}) + x \cdot \mathrm{ff}(\varphi_k, d'_j)}, \tag{5}$$

which is not order preserving. However, if we set $k'_1 = x \cdot k_1$ we get $w(\varphi_k, d'_j, k'_1) = x \cdot w(\varphi_k, d_j, k_1)$ with which we can maintain the original ordering.

5.2 Geographical Coordinates

For our BM25 model for geographical coordinates, we consider documents that are enriched with a discrete set of geographical coordinates. Let us model the three main ingredients of our weighting scheme: *ff*, *df* and document length, as follows. The *ff* of a coordinate in a document is defined as the number of occurrences of that coordinate in the document. The *df* is the number of documents that contain this coordinate and the document length is the number of locations in a document. Hereby, we assume that a document annotated with many geographical coordinates, covers a larger scope than a document with less coordinates, thus the argument of the textual BM25 document length normalization holds. Further, we assume, that the queries ask for documents in a specific geographical area, therefore a query is described by a single bounding box that encloses this area. The feature set and the feature frequency of a geographical feature φ_k for a query q is defined as

$$\Phi(q) := \mathrm{boundingbox}(q) \tag{6}$$

$$\mathrm{ff}(\varphi_k, q) := 1. \tag{7}$$

5.3 Ratings of Books

For the ratings, we consider documents, that describe books including ratings given by their readers. When searching for books with a textual query, we do not know any query specific preference for a rating. However, we assume that in general readers will prefer books with higher ratings. If the ratings are in the range between one and five, we define the query as

$$\Phi(q) := \{1,2,3,4,5\} \tag{8}$$

$$\mathrm{ff}(\varphi_k, q) := \varphi_k. \tag{9}$$

Hereby, all the possible ratings (1-5) are part of the query, while the weight of a rating is equal to the rating itself; i.e. the weight of the rating 5 is 5 times higher than the weight of the rating 1. The three main ingredients of our weighting scheme: feature frequencies $\mathrm{ff}(\varphi_k, d_j)$, document frequencies $\mathrm{df}(\varphi_k)$ and document lengths l_j, are defined analogously to their definition for textual modalities. The *ff* is the number of times a rating occurs in a given document, the *df* is the number of documents that contain a given rating and the document length is the number of ratings in a document. We assume that a document with many ratings covers a larger range of opinions, hence covering a larger scope and thus the argument of the textual BM25 document length normalization holds.

6 Sampling-Based BM25 for Modality Merging

6.1 Sampling

The proposed BM25 adaption for non-textual modalities enables us to merge modalities using the same weighting scheme, i.e a similar search engine as requested by the raw-score merging hypothesis. However, the raw-score merging hypothesis not only demands that similar search engines are used but also that the collection statistics are similar. Note, that the raw-score merging hypothesis is a rather old concept that has been introduced when merging multiple, possibly distributed textual document collections. In retrieval tasks with multiple modalities, the "collections" are no longer a set of textual documents but the different modalities. We have seen that the non-textual modalities have vastly different collection statistics, which invalidates the raw-score merging hypothesis. Therefore, we suggest a sampling based approach that allows us to adjust some properties of the collection statistics in order to reduce the difference. In particular, we adjust the average document length and the variance of the document lengths.

Our proposed sampling approach is similar to what is done in image retrieval when using dense or random feature sampling, where the same number of features for each image regardless of the pixel density and the number of concepts shown in the image is used [27]. The idea is to sample all modalities in all documents to a fixed document length as illustrated in Figure 3 for a single modality before BM25 is applied. Hereby, we use the number of tokens as the document length, although different definitions can be used. This results in the same collection statistics for

all the modalities with respect to the average document length and the variance of document lengths. Namely, the average document length is the sampling size and the variance is zero. Since all documents have the same length no BM25 document length normalization is necessary, thus we choose $b = 0$.

Fig. 3: Visualization of sampling three documents to the sampling size 5.

Original		Sampled
$d_1 : \varphi_1\varphi_2$		$d'_1 : \varphi_1\varphi_1\varphi_2\varphi_2\varphi_2$
$d_2 : \varphi_1\varphi_2\varphi_3$	$\Longrightarrow$	$d'_2 : \varphi_1\varphi_1\varphi_2\varphi_2\varphi_3$
$d_3 : \varphi_1\varphi_2\varphi_3\varphi_4\varphi_5\varphi_6$		$d'_3 : \varphi_1\varphi_2\varphi_3\varphi_4\varphi_5$

The randomized sampling, however, leads to data loss due to down sampling and non-deterministic results. Therefore, we idealize the sampling idea by not sampling the document but simply simulating the resulting term statistics. This can be done by scaling the feature frequencies by the relative change of the document length that would result from sampling. For a single document d_j and a single modality with length l_j and a token φ_k with the feature frequency $\mathrm{ff}(\varphi_k, d_j)$ the scaled term frequency $\mathrm{ff}'(\varphi_k, d_j)$ is

$$\mathrm{ff}'(\varphi_k, d_j) = \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}, \tag{10}$$

where s is the sampling size (the fixed length of all documents). For example, if s is $3l_j$, all term frequencies are multiplied by 3.

We denote our idealized sampling based BM25 adaption BM25*S, where S stands for the sampling and the asterisk shows that no traditional document length normalization is applied; i.e. $b = 0$. The resulting the BM25*S weight for document d_j with sampling size s is

$$w_{\mathrm{BM25*S}}(\varphi_k, d_j) = \frac{\mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}{k_1 + \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}. \tag{11}$$

Our sampling approach is some form of document replication, and thus the *ff* saturation parameter k_1 is not optimal anymore as described in Section 5 and by Robertson *et al.* [35]. In order to achieve the same retrieval effectiveness as without the sampling, the k_1 parameter needs to be adjusted. Since not all documents are replicated with the same factor, the optimal adjustment of the k_1 parameter cannot simply be the replication factor as in Section 5. However, we observed an approximately linear dependency of the optimal k_1 parameter to the average document length. Therefore, we set

$$k'_1 = \frac{\Delta'}{\Delta} \cdot k_1, \tag{12}$$

where Δ is the average document length of the original documents and Δ' is the average document length of the sampled documents. This adjustment is slightly different

to the adjustment Robertson *et al.* [35] suggested, who used the ratio between the average term frequencies rather than the average document lengths. However, with their setup the two ratios are equivalent. With the sampling, the two ratios are not exactly equal, although quite similar, therefore both options seem valid. Further, when sampling, calculating the ratio between the average document lengths is a lot simpler than between the average term frequencies since the average document length after the sampling is equal to the sampling size ($\Delta' = s$), while the new average term frequencies are only known after the sampling is performed.

The weight for a document d_j, when using the combination of the idealized sampling and the k_1 adjustment (BM25-sampled), is calculated as

$$w_{\text{BM25-sampled}}(\varphi_k, d_j) = \frac{\text{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}{k_1 \cdot \frac{s}{\Delta} + \text{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}. \tag{13}$$

We now have a sampling method BM25-sampled that can be applied to all modalities. We suggest using the same sampling length for all modalities, which results in the same collection statistics for all modalities with respect to the average document length and variance in document lengths. Hence, the raw-score merging hypothesis is fulfilled with respect to these two properties.

We can prove that this sampling method results in exactly the same weights as for BM25 with the normalization parameter b set to one.

Proof.

$$\begin{aligned} w_{\text{BM25-sampled}}(\varphi_k, d_j) &= \frac{\text{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}}{k_1 \cdot \frac{s}{\Delta} + \text{ff}(\varphi_k, d_j) \cdot \frac{s}{l_j}} \\ &= \frac{\text{ff}(\varphi_k, d_j)}{k_1 \cdot \frac{s}{\Delta} \cdot \frac{l_j}{s} + \text{ff}(\varphi_k, d_j)} \\ &= \frac{\text{ff}(\varphi_k, d_j)}{k_1 \cdot \frac{l_j}{\Delta} + \text{ff}(\varphi_k, d_j)} \\ &= w_{\text{BM25}(b=1)}(\varphi_k, d_j). \end{aligned}$$

□

This proof shows, that BM25 with full document length normalization ($b = 1$) already guarantees that the raw-score merging hypothesis is fulfilled with respect to the average document length and variance in document lengths. Therefore, BM25 seems to be suited to be used in a multimodal retrieval task. It however has been shown, that using $b = 1$ for BM25 tends to underestimate the relevance of long documents and therefore usually a smaller b is used; e.g. $b = 0.75$. In the following, we show how the sampling idea can be extended to allow arbitrary document length normalization parameters b.

6.2 Scope-aware Sampling

Sampling all documents to the same length, which is equal to using BM25 with full document length normalization ($b = 1$), assumes that all documents have the same

scope. However, some documents might discuss more topics than other documents and thus indeed should be represented with more tokens as described in Section 5. Similarly to BM25, we assume that the original document lengths of the documents give an indication about their scope. Thus, we can account for different document scopes by sampling the documents to different lengths based on their original length.

Many different definitions of a scope-aware sampling length using a document length normalization parameter *bs* are possible. We can however choose a definition so that the sampling based approach is identical to the traditional BM25 with parameter *b*=*bs*. We therefore define the adjusted number of sampled tokens s' for a document d_j as

$$s'(d_j) = l_j \cdot \frac{s}{(1 - bs + bs \cdot \frac{l_j}{\Delta}) \cdot \Delta}. \tag{14}$$

All documents are now sampled to their corresponding sampling size $s'(d_j)$ rather than the same sampling size s for all documents. The adjusted feature frequencies therefore are

$$\begin{aligned} \mathrm{ff}'(\varphi_k, d_j) &= \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s'(d_j)}{l_j} \\ &= \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{(1 - bs + bs \cdot \frac{l_j}{\Delta}) \cdot \Delta}. \end{aligned} \tag{15}$$

Unfortunately, this non-linear transformation of the document lengths does not exactly result in the same average document length for each modality, which would be necessary to fulfill the raw-score merging hypothesis. However, we found that the new sampled average document lengths of the modalities are close to each other and it is in practice a valid assumption that they are equal.

Further, we have found, that the optimal k_1 has no longer a linear dependency on the new average document length Δ' as we found for the sampling with a fixed sampling size s (BM25-sampled) as described in Section 6. It rather has a linear dependency to the sampling length s. Thus, for the scope-aware sampling we adjust the k_1 parameter as

$$k_1' = \frac{s}{\Delta} \cdot k_1. \tag{16}$$

We denote this scope-aware sampling with the k_1 adjustment and the non-normalized BM25 as BM25-scope. Its weight for a document d_j is calculated as

$$w_{\text{BM25-scope}} = \frac{\mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{(1 - bs + bs \cdot \frac{l_j}{\Delta}) \cdot \Delta}}{k_1 \cdot \frac{s}{\Delta} + \mathrm{ff}(\varphi_k, d_j) \cdot \frac{s}{(1 - bs + bs \cdot \frac{l_j}{\Delta}) \cdot \Delta}}. \tag{17}$$

With the scope-aware sampling it is possible to achieve approximately the same average document length for all modalities in all documents, while documents with a large scope are still represented by more tokens, by using the same sampling size parameter s for all modalities.

We can show that this scope-aware sampling is identical to the traditional BM25 for any document length parameter *bs*.

Proof.

$$
\begin{aligned}
w_{\text{BM25-scope}} &= \frac{\text{ff}(\varphi_k, d_j) \cdot \frac{s}{(1-bs+bs\cdot\frac{l_j}{\Delta})\cdot\Delta}}{k_1 \cdot \frac{s}{\Delta} + \text{ff}(\varphi_k, d_j) \cdot \frac{s}{(1-bs+bs\cdot\frac{l_j}{\Delta})\cdot\Delta}} \\
&= \frac{\text{ff}(\varphi_k, d_j)}{k_1 \cdot \frac{s}{\Delta} \cdot \frac{(1-bs+bs\cdot\frac{l_j}{\Delta})\cdot\Delta}{s} + \text{ff}(\varphi_k, d_j)} \\
&= \frac{\text{ff}(\varphi_k, d_j)}{k_1 \cdot (1-bs+bs\cdot\frac{l_j}{\Delta}) + \text{ff}(\varphi_k, d_j)} \\
&= w_{\text{BM25}(b=bs)}(\varphi_k, d_j). \qquad (18)
\end{aligned}
$$

□

Since BM25 is identical to our sampling approach BM25-scope, also BM25 is fulfilling the raw-score merging hypothesis with respect to the average document length with any document length normalization parameter. We can therefore conclude, that differences between average document lengths can be ignored when using raw-score merging with BM25. Hence, we can use BM25 with the same document length normalization parameter b for all modalities. The sampling approach is not needed in practice, since we have shown that it is identical to BM25.

Unlike BM25 with full document length normalization ($b = 1$), the variances of the document lengths are however not necessarily the same. Using our sampling idea, we can further adjust the definition of the sampled number of tokens in order to compensate the different variances of document lengths. We first apply a transformation to the document lengths to adjust the variance and then adjust the average document lengths as in equation 14 using the transformed document lengths. Thus, we do not ensure that all variances in document length are the same, but we ensure that the ratio between the standard deviation and the average document length is the same for all modalities. The adjusted number of tokens s'' with the adjustment for the variance of document length is

$$l'_j = (l_j - \Delta) \cdot rs \cdot \frac{\Delta}{\sigma} + \Delta \qquad (19)$$

$$s''(d_j) = l'_j \cdot \frac{s}{(1-bs+bs\cdot\frac{l'_j}{\Delta})\cdot\Delta}, \qquad (20)$$

where σ is the standard deviation of the document lengths and rs is the variance parameter that defines the target ratio between the standard deviation and the mean. We denote this sampling variation as BM25-var.

7 Experiments

The focus of our evaluation lies on measuring the effectiveness of a multimodal IR system built according to our guidelines (consistent treatment of the modalities, little or no training). In the scenarios we are interested in, the system needs to incorporate *all* modalities; ignoring modalities is not an option.

Our test system is built on top of Lucene[1] and is using the built-in weighting schemes wherever possible. For the scaled feature frequency and the k_1 adjustment, we adapted the built-in BM25 implementation. The merging of the modalities is performed using a raw-score merging (".raw") or a linear combination of the scores (".opt"). By using the latter, we violate our goal of using no training phase. Indeed, we use the opt-variant only for comparison purposes as a benchmark. In line with this role as a sort of "upper bound" on performance, we train the optimal weights using the same collection as used for testing. In essence for the opt-variant, we are only interested in showing that the effectiveness can be improved using BM25 on multiple textual as well as non-textual modalities.

Our experiments use two multimodal test collections, GeoCLEF and SBS.

7.1 Test Collections

7.1.1 GeoCLEF

For the experiments with the geographical modality, we use the topics and collection of the GeoCLEF 2008 [26] monolingual English search task. The collection is composed of the news articles from the British newspaper *The Glasgow Herald* (1995) and the American newspaper *The Los Angeles Times* (1994). In this task, 24 geographically challenging topics have been defined; e.g. "*Nobel prize winners from Northern European countries*". Here, we can differentiate between the textual information "Nobel price winners" and the geographical information "from Northern European countries". One of the challenges of geographical IR is that relevant documents not only contain the textual representation of geographical information "Northern European countries", but also concepts such as unions, countries or cities inside the geographical region.

Overell *et al.* [30] and Buscaldi *et al.* [5] proposed to separate the geographical information from the textual information, so that the two modalities (geographical and textual) can be treated differently. This allows that the additional information about geographical regions can be considered. Buscaldi *et al.* [5] extracted location names from the documents and topics and mapped them to their geographical coordinates (longitude, latitude) using GeoWordNet. D. Buscaldi provided us a preprocessed geotagged version of the GeoCLEF 2008 collection. Further, we preprocessed the title fields of the topics by manually extracting a geographical bounding box for each topic. This could also be done automatically using the convex hull of the locations found with GeoWordNet [5].

An important characteristic of the collection and task described above is the overlap of the textual and geographical modalities, since the geographical modality is extracted from the text. Therefore, we also created a second modified version of the GeoCLEF 2008 test collection, which separates the geographical and textual information. For this, we removed the textual description of the geographical region from the queries; e.g. the query "*Nobel prize winners from Northern European countries*" becomes "*Nobel prize winners*" with the geographical bounding box that includes all

[1] https://lucene.apache.org/core/

Northern European countries. In the experiments, we refer to this task as "geoCLEF-mod".

7.1.2 Social Book Search

For the experiments using the ratings as an additional modality, we use the Social Book Search (SBS) 2016 lab task [20]. The collection consists of 2.8 million books from Amazon, extended with social meta-data from LibraryThing. For each book the fields ISBN, title, review, summary, ratings and tags are given. Each query is constructed from a real user request on LibraryThing. The query not only includes the title of the request and the description of the request itself but also example books mentioned by the user. Additionally, the personal catalog of each topic creator is available, which includes a list of the books the user has archived on LibraryThing along with his personal ratings. The relevance assessments are based on the actual suggestions to the original query on the LibraryThing forum. Forum suggestions normally get a relevance value of 1, however if the suggested book is already in the personal catalog of the topic creator the relevance value is 0. When the topic creator actually adds a suggested book to his library it is considered highly relevant and receives a relevance value of 4.

For the textual modality, we use the textual baseline established in our SBS participation [16,17]. We combine all textual fields of the documents into a single textual index field. The queries are constructed from the two textual topic fields title and request that are analogously combined into a single textual representation. Further, we expand the query text with the 35 most characteristic terms (determined by BM25) from the textual representation of the content of the example books given by the topic creator. All books already read by the topic creator are filtered from the result list.

7.2 Results

Following our own guidelines on how to build a multimodal IR system, we sample the non-textual modalities to the same length as the textual modality. For the GeoCLEF 2008 collection, we therefore sample the geographical modality from an average document length of 7.4 to the sampling length of 357.7. Analogously, the ratings in the SBS collection with an average document length of 5.05 are sampled to the sampling length of 674.7. The target standard deviation ratio parameter *rs* is chosen based on the textual modality as well. For GeoCLEF 2008 this is 1.01 and 2.75 for SBS. This results in a reduction of the standard deviation for the non-textual modalities to 83% respectively 93%. For the runs using the scope-aware sampling (BM25-scope and BM25-var) the normalization parameter *bs* is 0.75. Note that the scope-aware sampling BM25-scope is identical to BM25 and BM25-sampled is identical to BM25 with document length normalization parameter $b = 1$.

As mentioned, the goal of this paper is to establish a baseline for a multimodal IR system that involves all the given modalities and merges the scores generated by a unified model under the raw-score merging hypothesis. Hereby, we require all the modalities to be considered in the result list. We argue that in practice, it is not

possible, for many reasons, including e.g. regulatory ones, to simply ignore or discard entire modalities, or parts of the document collection. For example, a book selling company might find that good ratings of books positively influences the purchase behavior of their customers and thus the ratings have to be included in the search engine.

Building an effective multimodal IR system that integrates all modalities with little or no training remains a hard challenge. Wildly different characteristics, and wildly different degrees of informativeness across the modalities means that the average retrieval effectiveness may *drop* when integrating all modalities, such as evaluated through popular measures like MAP. We advise caution in overinterpreting such a result. Firstly, the average hides many meaningful changes in system behavior and secondly, user perception will likely be different from the measured average improvement if a user realizes that parts of his query or of the documents are ignored. For the time being, a lower retrieval effectiveness of an experiment integrating all modalities versus an experiment discarding some modalities thus mainly serves to highlight how far we still are from finding the perfect recipe for multimodal retrieval, but not to point to a reduced system as a viable, practical alternative.

In the following experiments, we show the effectiveness of our multimodal baseline using the three derived versions of BM25 as the unified weighting scheme for all the modalities merged under the raw-score merging hypothesis. In each of the following Tables 2, 3, 4 and 5, we compare two runs with the same collection. We underline any statistically significant differences in performance according to the MAP to the first run resulting from a paired randomization test [42] (significance level $\alpha = 5\%$). For the GeoCLEF 2008 collection, we removed the outlier query 79-GC to calculate the significance. In Appendix B we additionally show the same runs evaluated using the nDCG@10 measure. The following conclusions drawn from the results using the MAP are all supported by the results using the nDCG@10.

Base performance of systems integrating non-overlapping modalities

We start our experiments by establishing the base performance of multimodal systems that integrate all non-overlapping modalities as built according to our guidelines.

Table 2: Retrieval results (MAP) for the runs with the textual modalities and the raw-score merging of both modalities for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
SBS.text	0.0320	0.0396	0.0396
SBS.text+ratings.raw	0.0390	0.0448	0.0447
geoCLEFmod.text	0.1310	0.1419	0.1419
geoCLEFmod.text+geo.raw	0.1226	0.0688	0.0678

To this end, Table 2 shows the MAP for the SBS 2016 and the GeoCLEFmod 2008 collection both for the multimodal baseline (denoted as ".raw") and the runs

with the textual modalities alone (denoted as ".text"). As a consequence of our discussion above, the ".text"-run can only serve as a yardstick: it violates the rule that we want to integrate all modalities. Effectively, it gives us a "lower bound" of performance to compare to. For the SBS collection, the multimodal baseline achieves a significantly higher MAP than the textual run. For the GeoCLEFmod 2008 collection the run with BM25 with no document length normalization (BM25 (b=1)), which is identical to BM25-sampled, achieves a MAP in the range of the textual run. The BM25-scope and BM25-var runs with raw-score merging achieve a lower MAP than the run with text only.

Analysis of individual modalities

It is helpful to further look into the contributions of individual modalities to the overall result. Table 3 shows the retrieval effectiveness of each modality individually. Both the geographical modality and the ratings do not achieve the same retrieval effectiveness as the textual modality. This was expected for both, since intuitively the textual description of a book is more important than its ratings and the textual content of a newspaper article is more important than the mentioned geographical locations.

Merging under the raw-score hypothesis suggests adding the scores of the different modalities into a single score without any weights. However, as shown in Table 2 even though we proved that the raw-score merging hypothesis is fulfilled w.r.t. the average document length as well as for the variance of the document lengths (for BM25-var) the merged result list is only better than the textual run for the SBS task and not for the GeoCLEF task. We claim that this is since the method so far cannot properly capture the difference in informativeness of the modalities.

Table 3: Retrieval results (MAP) for the runs with the textual modalities and the non-textual modalities (geographical coordinates and ratings) for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
SBS.text	0.0320	0.0396	0.0396
SBS.ratings	0.0089	0.0121	0.0121
geoCLEFmod.text	0.1310	0.1419	0.1419
geoCLEFmod.geo	0.0540	0.0589	0.0588

Dealing with overlapping modalities

We next want to explore to what extent the overlapping of content in modalities has an impact on the overall effectiveness. Table 4 shows the MAP of the textual run and the multimodal baseline using the GeoCLEFmod 2008 task as well as GeoCLEF 2008 task.

As expected the textual modality in the GeoCLEFmod task achieves a lower MAP than the textual modality in the original GeoCLEF task. This is due to the deletion

of the geographical information in the textual modality as described in Section 7.1.1. The modalities in the GeoCLEFmod 2008 task therefore do not have an information overlap, while the modalities in the GeoCLEF 2008 task do contain overlapping information, namely all the information present in the geographical modality is also present in the textual modality. The experiments that merge the two modalities under the raw-score merging hypothesis show that without the information overlap between the modalities the MAP of the merged run ("geoCLEFmod.text+geo.raw") is within the range of the textual modality alone. However, when merging modalities with an information overlap ("geoCLEF.text+geo.raw") the MAP drops significantly - it is much harder to merge the modalities so that only the "additional" contribution makes a beneficial impact.

Table 4: Retrieval results (MAP) for the runs with the textual modalities and the raw-score merging of both modalities for the GeoCLEFmod 2008 and the GeoCLEF 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
geoCLEFmod.text	0.1310	0.1419	0.1419
geoCLEFmod.text+geo.raw	0.1226	0.0688	0.0678
geoCLEF.text	0.2509	0.2566	0.2566
geoCLEF.text+geo.raw	0.1548	0.0705	0.0703

Optimal merging potential due to training

We argue that a lot of the drop in retrieval effectiveness from the ".text" to the ".text+geo.raw" experiment is due to the inherent difficulty of appropriately merging the contributions of the individual modalities into the overall result. The closest method to raw-score merging that allows us to weight the contributions of the individual modalities is a linear combination of the scores. Therefore, we try to verify this assumption through comparing the multimodal baseline (".raw") with an approximate upper bound using a linear combination of the scores with trained weights (".opt") (see Table 5). The optimal weights are trained on the information available in the relevance assessments of the test collection. Clearly, this information is not available in practice. Furthermore, training the optimal weights on the same queries as were tested turns this in a retrospective evaluation. As the obtained result is merely a data point to compare our results to, we accept these limitations. For SBS there is no significant difference between merging the modality scores under the raw-score hypothesis and merging using the optimal linear combination. However, for the GeoCLEFmod 2008 collection merging the scores of the textual and the non-textual modalities using optimal linear combination has a significantly higher MAP then the merging under the raw-score merging hypothesis. Consider, however, that the opt-variants only serve as a yardstick: They can only be used when training data is available which is often missing in practical applications and which was not the goal of this paper. The optimal run also shows that the usage of BM5 for the non-textual modalities not only

leads to good results when merging under the raw-score merging hypothesis but also when training optimal weights. The traditional BM25, which is identical to BM25-scope, already seems to be a good choice, since the variance adjustment does not lead to a significantly better result neither for raw-score merging nor for the optimal linear combination of the scores.

To get more context in order to judge the performance of our ".raw" runs, we have also explored the use of reciprocal rank fusion [10], another well known unsupervised fusion method. These runs are denoted with ".rcpr" in Table 5, where we underline the runs that are significantly different to the ".raw" runs. For the SBS collection, reciprocal rank fusion leads to a significantly lower MAP for all BM25 variants. However, for the GeoCLEFmod 2008 collection the MAP is in the same range as the raw-score merging run with BM25-sampled but significantly better with BM25-scope and BM25-var, although still significantly lower than the optimal linear combination (".opt").

Table 5: Retrieval results (MAP) for the runs with the raw-score merging of the modalities and the optimized linear combination of the modality scores for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
SBS.text+ratings.raw	0.0390	0.0448	0.0447
SBS.text+ratings.opt	0.0398	0.0450	0.0450
SBS.text+ratings.rcpr	0.0104	0.0139	0.0139
geoCLEFmod.text+geo.raw	0.1226	0.0688	0.0678
geoCLEFmod.text+geo.opt	0.2351	0.2442	0.2446
geoCLEFmod.text+geo.rcpr	0.1292	0.1393	0.1393

Summary of results

We can summarize the results of our experiments with the following questions.

1. Can we produce a multimodal baseline with an effectiveness in the range of the textual run? **Yes**, we find better retrieval effectiveness for the SBS collection and retrieval effectiveness in the same range (within statistical significance) for the GeoCLEF collection without overlapping modalities.
2. Do modalities differ with respect to their contribution to relevance? **Yes**, in both collections the contribution by the textual modality is by far the greatest, thus turning the ".text" yardstick into a challenging lower bound.
3. Does it matter that modalities have overlapping information? **Yes**, it is much harder to merge individual contributions by modalities in case they are overlapping.
4. Is it possible to get competitive performance without training? **Yes and no**. We have found competitive performance in the case of the SBS collection, where we have no overlapping modalities. We are still a long way from matching the performance of the opt-variant on the GeoCLEF collection, however.

8 Conclusions

In this paper, we demonstrate best practices for the integration of many modalities into an IR application without the use of training data. We claimed that in complex multimodal collections with a large number of diverse modalities, it becomes crucial to treat the modalities with a unified model, due to the quickly increasing complexity. We started by analyzing the requirements for such a unified model and showed that BM25 is a suitable weighting scheme to be used and to merge the modalities under the raw-score merging hypothesis. We proposed an adaptation of the BM25 weighting scheme for the two non-textual modalities ratings and geographical coordinates and established a multimodal baseline that uses all the modalities and merges them under the raw-score merging hypothesis without any training.

In order to show the suitability of BM25 scores to be merged under the raw-score merging hypothesis, a sampling based approach for BM25 was introduced to deal with the different collection statistics, in particular the average document length and the variance of the document lengths of the modalities. We proved that applying BM25 with full document length normalization $b = 1$ to all modalities already ensures that the raw-score merging hypothesis w.r.t. the average document lengths and the variance of document lengths is fulfilled, since it is identical to the sampling approach. Analogously, we proved that the raw-score merging hypothesis w.r.t. the average document length also holds for BM25 with a general document length normalization parameter $b \neq 1$, however not w.r.t. the variance of document length. Our experiments show that adhering to the raw-score merging hypothesis is indeed beneficial.

In our experiments, we established a multimodal baseline that involves all the given modalities and merges the scores generated by a unified model under the raw-score merging hypothesis. We showed that by following our approach the multimodal baseline reaches a significantly better retrieval effectiveness than the textual run for the SBS collection and lies within the same range (within statistical significance) for the GeoCLEF 2008 collection without overlapping modalities. Further, we analyzed the contribution of the individual modalities to relevance and found that the contribution of the textual modalities is the greatest. Also, we saw in the experiments that dealing with modalities with overlapping information is a hard problem. Finally, we found similar performance of our multimodal baseline when comparing it to a trained linear combination of the scores in case of the SBS collection, which we consider to be very encouraging.

The multimodal baseline presented in this paper merges the modality scores under the raw-score merging hypothesis and therefore assumes that each modality is equally important for the overall relevance of a document. However, in the experiments we saw that there are wildly different degrees of informativeness across the modalities. As a next step towards best practices for multimodal IR systems, we will investigate to further extend the proposed methods but incorporate the informativeness of the different modalities without the usage of any training data.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) **20**(4), 357–389 (2002)
2. Bogers, T., Koolen, M., Jaap, K., Kazai, G., Preminger, M.: Overview of the INEX 2014 Social Book Search Track. In: Conference and Labs of the Evaluation Forum, pp. 462–479 (2014)
3. Braschler, M.: Combination approaches for multilingual text retrieval. Information Retrieval **7**(1), 183–204 (2004)
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1-7), 107–117 (1998). DOI 10.1016/S0169-7552(98)00110-X. URL http://dx.doi.org/10.1016/S0169-7552(98)00110-X
5. Buscaldi, D., Rosso, P.: The UPV at GeoCLEF 2008: The GeoWorSE system. In: Working Notes from the Cross Language Evaluation Forum (2008)
6. Bush, V., et al.: As we may think. The Atlantic Monthly **176**(1), 101–108 (1945)
7. Callan, J.P., Lu, Z., Croft, W.B.: Searching distributed collections with inference networks. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 21–28. ACM (1995)
8. Chowdhury, A., McCabe, M.C., Grossman, D., Frieder, O.: Document normalization revisited. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, pp. 381–382. ACM (2002)
9. Cleverdon, C.: The cranfield tests on index language devices. In: Aslib proceedings, vol. 19, pp. 173–194. MCB UP Ltd (1967)
10. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 758–759. ACM (2009)
11. Craswell, N., Robertson, S., Zaragoza, H., Taylor, M.: Relevance weighting for query independent evidence. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 416–423. ACM (2005)
12. Dalvi, N.N., Kumar, R., Pang, B.: Para'normal'activity: On the distribution of average ratings. In: ICWSM (2013)
13. Depeursinge, A., Müller, H.: Fusion techniques for combining textual and visual information retrieval. In: ImageCLEF, pp. 95–114. Springer (2010)
14. Fox, E.A., Shaw, J.A.: Combination of multiple searches. NIST Special Publication SP pp. 243–243 (1994)
15. Hashemi, S.H., Kamps, J.: Venue recommendation and web search based on anchor text. Tech. rep., DTIC Document (2014)
16. Imhof, M.: BM25 for Non Textual Modalities in Social Book Search. In: Seventh International Conference of the CLEF Association, CLEF (2016)
17. Imhof, M., Badache, I., Boughanem, M.: Multimodal social book search. In: Sixth International Conference of the CLEF Association, CLEF (2015)
18. Imhof, M., Braschler, M.: Are test collections real? Mirroring Real-World Complexity in IR Test Collections. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 241–247. Springer (2015)
19. Kamps, J., De Rijke, M., Sigurbjörnsson, B.: Length normalization in XML retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, pp. 80–87. ACM (2004)
20. Koolen, M., Bogers, T., Gäde, M., Hall, M., Hendrickx, I., Huurdeman, H., Kamps, J., Skov, M., Verberne, S., Walsh, D.: Overview of the CLEF 2016 social book search lab. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 351–370. Springer (2016)
21. Kwok, K., Grunfeld, L., Lewis, D.: TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. NIST Special Publication SP pp. 247–247 (1995)
22. Li, X., Croft, W.B.: Time-based language models. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM '03, pp. 469–475. ACM, New York, NY, USA (2003). DOI 10.1145/956863.956951. URL http://doi.acm.org/10.1145/956863.956951
23. Losada, D.E., Azzopardi, L.: An analysis on document length retrieval trends in language modeling smoothing. Information Retrieval **11**(2), 109–138 (2008)

24. Lv, Y., Zhai, C.: When documents are very long, BM25 fails! In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1103–1104. ACM (2011)
25. Macdonald, C., Dinçer, B.T., Ounis, I.: Transferring learning to rank models for web search. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, pp. 41–50. ACM (2015)
26. Mandl, T., Carvalho, P., Di Nunzio, G.M., Gey, F., Larson, R.R., Santos, D., Womser-Hacker, C.: GeoCLEF 2008: the CLEF 2008 cross-language geographic information retrieval track overview. In: Evaluating Systems for Multilingual and Multimodal Information Access, pp. 808–821. Springer (2009)
27. Moulin, C., Barat, C., Ducottet, C.: Fusion of tf. idf weighted bag of visual features for image classification. In: Content-Based Multimedia Indexing (CBMI), 2010 International Workshop on, pp. 1–6. IEEE (2010)
28. Mourão, A., Martins, F., Magalhães, J.: Multimodal medical information retrieval with unsupervised rank fusion. Computerized Medical Imaging and Graphics **39**, 35–45 (2015)
29. Müller, H., Clough, P., Deselaers, T., Caputo, B.: Image-CLEF: Experimental evaluation in visual information retrieval series. The information retrieval series, Springer (2010)
30. Overell, S., Rae, A., Rüger, S.: MMIS at GeoCLEF 2008: Experiments in GIR. In: Working Notes from the Cross Language Evaluation Forum (2008)
31. Peters, C., Braschler, M., Clough, P.: Multilingual information retrieval: From research to practice. Springer Science & Business Media (2012)
32. Rajaraman, S.: Five stars dominate ratings (2009). URL https://youtube.googleblog.com/2009/09/five-stars-dominate-ratings.html
33. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation **60**(5), 503–520 (2004)
34. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc (2009)
35. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 42–49. ACM (2004)
36. Robertson, S.E., Van Rijsbergen, C., Porter, M.F.: Probabilistic models of indexing and searching. In: Proceedings of the 3rd annual ACM conference on research and development in information retrieval, pp. 35–56. Butterworth & Co. (1980)
37. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
38. Savoy, J.: Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002 Rome, Italy, September 19–20, 2002 Revised Papers, chap. Report on CLEF 2002 Experiments: Combining Multiple Sources of Evidence, pp. 66–90. Springer Berlin Heidelberg, Berlin, Heidelberg (2003)
39. Savoy, J.: Data Fusion for Effective European Monolingual Information Retrieval, pp. 233–244. Springer Berlin Heidelberg, Berlin, Heidelberg (2005). DOI 10.1007/11519645_24. URL http://dx.doi.org/10.1007/11519645_24
40. Schuth, A., Balog, K., Kelly, L.: Overview of the living labs for information retrieval evaluation (LL4IR) CLEF lab 2015. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 484–496. Springer (2015)
41. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, pp. 21–29. ACM (1996)
42. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledge management, pp. 623–632. ACM, New York, NY, USA (2007). DOI http://doi.acm.org/10.1145/1321440.1321528
43. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., et al.: General overview of imageCLEF at the CLEF 2015 labs. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 444–461. Springer (2015)
44. Voorhees, E., Gupta, N.K., Johnson-Laird, B.: The collection fusion problem. NIST Special Publication SP pp. 95–95 (1995)

45. Voorhees, E.M., Harman, D.: Overview of the fifth text retrieval conference (TREC-5). In: TREC, vol. 97, pp. 1–28 (1996)
46. Voorhees, E.M., Harman, D.: Overview of the eighth text retrieval conference (TREC-8). In: TREC (1999)
47. Wilkins, P., Ferguson, P., Smeaton, A.F.: Using score distributions for query-time fusion in multimediaretrieval. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, pp. 51–60. ACM (2006)
48. Woolf, M.: A statistical analysis of 1.2 million amazon reviews (2014). URL http://minimaxir.com/2014/06/reviewing-reviews/
49. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proceedings of the international workshop on Workshop on multimedia information retrieval, pp. 197–206. ACM (2007)

A Validating the Raw-Score Hypothesis

As shown in Section 2, we propose to handle each modality separately. This fundamental approach models that a unified "merged" result list needs to be synthesized. As pointed out in Section 3, the raw-score merging hypothesis states that merging scores from multiple ranked lists is more effective when the scores are produced from the same underlying weighting scheme with the same collection statistics. As the hypothesis is an important stepping-stone to the definition of a consistent, "best-practice" way of treating each modality, we present an attempt to verify it experimentally. Similar to Savoy [39], we investigate this hypothesis on the multilingual document collection used in the CLEF 2004 AdHoc-News task. It consists of four document collections in four languages: English, Finnish, French and Russian. The queries are also provided in all the four languages; however the goal is to present a single ranked list with all the relevant documents from all languages. Hence, the result lists resulting from the monolingual retrieval have to be merged.

In contrast to the work by Savoy, we are interested in how different commonly used weighting schemes behave in respect to the raw-score merging hypothesis. In the following experiments, we therefore use the four weighting schemes: BM25, divergence from randomness (DFR), language models (LM) and TF-IDF for the retrieval and show the resulting retrieval effectiveness when merging them into a single ranked list. Since the goal of the experiments is not to get the highest effectiveness possible but to show the validity of the raw-score merging hypothesis, we did not optimize any parameters of the weighting schemes and used the default analyzers for each language provided by Lucene. The saturation parameter k_1 of BM25 is set to 1.2 and the document length normalization parameter b = 0.75. The basic model of the DFR weighting scheme is the limiting form of Bose-Einstein, for the first normalization the Laplaces law of succession is used and the second normalization is based on the second hypothesis. For the LM we use Dirichlet smoothing with a smoothing parameter μ of 2000. We constructed short queries using the title and the description given for each topic. Table 6 shows the mean average precision (MAP) for each language individually using the four different weighting schemes. For English the MAP is highest when using BM25, for Finnish the highest effectiveness is reached using LM and for French and Russian DFR leads to the highest MAP, when merging the scores resulting from these four runs we call it "Best".

We underline any statistically significant differences with respect to the run with the highest per-language MAP, which is printed in bold letters. Hereby, the significance is calculated using a paired randomization test [42] (significance level $\alpha = 5\%$). Looking closely at the difference in MAP between the BM25 and LM for the English collection, we can observe that for 10 queries over 50, LM offers a higher performance while for 25 requests BM25 performs better than LM. For the remaining 15 queries, the MAP difference between the two runs is smaller than 0.02. Thus, in average, BM25 depicts a higher MAP than LM. From a statistical point of view however, the difference cannot be viewed as significant because for several queries, LM presents a higher performance. A similar reasoning applies to the LM and the TF-IDF run for the Russian collection.

Table 6: Monolingual retrieval results (MAP) for CLEF 2004 using short queries (TD).

Run	English	Finnish	French	Russian
BM25	**0.4320**	0.3728	0.1618	0.2686
DFR	0.4228	0.3748	**0.1642**	**0.2760**
LM	0.4075	**0.3809**	0.1606	0.2360
TFIDF	0.4121	0.3580	0.1583	0.2577

We merge the ranked lists produced from four languages into a single ranked list using four different well-known merging strategies. In general when using raw-score merging all the scores of a document in all ranked lists are added to a single score, which is then used to produce the merged ranked list. However, in this multilingual setup each document is only available in a single language and therefore only gets a single score. In this case, the raw-score merging just results in ordering the documents from all languages with respect to the scores in the per language ranked list. In the round robin merging approach, we take one document in turn from each individual ranked list [44]. The third merging strategy is "Norm(max)" were we normalize the scores before merging by dividing them by the maximal document score of the corresponding query. The last strategy is a linear combination of the scores of the ranked lists. This strategy requires a training set to find the optimal weights for each language. We used the same collection for the training and testing since we are not interested in optimizing the effectiveness but to show the difference of the individual weighting schemes. Table 7 shows the MAP of four runs in which ranked lists produced by a single weighting scheme are merged into a ranked list (BM25.all, DFR.all, LM.all, TFIDF.all) and the MAP of the run were the ranked lists that produced the best MAP for each language individually are merged into a ranked list ("Best"). We underlined any statistically significant differences in performance according to the MAP of the runs using a single weighting scheme with respect to the "Best" run. Hereby, the significance is calculated using a paired randomization test [42] (significance level $\alpha = 5\%$).

As expected from the results by Savoy [39], the runs using the same weighting scheme for all languages perform significantly better than the "Best" run using the raw-score merging, while BM25 performs slightly better than the other weighting

Table 7: Multilingual retrieval results (MAP) for CLEF 2004 using different merging strategies.

Run	Raw-Score	Round Robin	Norm(max)	lin.comb.
BM25.all	**0.2494**	0.1874	0.2018	0.2606
DFR.all	0.2471	0.1892	0.2018	0.2589
LM.all	0.2400	0.1825	0.1940	0.2430
TFIDF.all	0.2412	0.1807	0.1866	0.2494
Best	0.1812	**0.1926**	**0.2046**	**0.2656**

schemes. Using the other merging strategies, the "Best" run performs the best, although not significantly. Also, the "Best" run requires that the best weighting scheme per-language is known, which usually is not the case in practical applications.

B Experimental Results with nDCG@10

The following tables show the results of the experiments described in Section 7.2 using the nDCG@10 measure.

Table 8: Retrieval results (nDCG@10) for the runs with the textual modalities and the raw-score merging of both modalities for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
SBS.text	0.0467	0.0561	0.0561
SBS.text+ratings.raw	0.0561	0.0634	0.0633
geoCLEFmod.text	0.1709	0.1826	0.1826
geoCLEFmod.text+geo.raw	0.1500	0.0728	0.0646

Table 9: Retrieval results (nDCG@10) for the runs with the textual modalities and the non-textual modalities (geographical coordinates and ratings) for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
SBS.text	0.0467	0.0561	0.0561
SBS.ratings	0.0122	0.0205	0.0207
geoCLEFmod.text	0.3573	0.3932	0.3932
geoCLEFmod.geo	0.2038	0.0733	0.0654

Table 10: Retrieval results (nDCG@10) for the runs with the textual modalities and the raw-score merging of both modalities for the GeoCLEFmod 2008 and the GeoCLEF 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
geoCLEFmod.text	0.0467	0.0561	0.0561
geoCLEFmod.text+geo.raw	0.0122	0.0205	0.0207
geoCLEF.text	0.1709	0.1826	0.1826
geoCLEF.text+geo.raw	0.0591	0.0630	0.0630

Table 11: Retrieval results (nDCG@10) for the runs with the raw-score merging of the modalities and the optimized linear combination of the modality scores for the SBS 2016 and the GeoCLEFmod 2008 collection using the three BM25 versions.

Run	BM25 (b=1) BM25-sampled	BM25 (b=0.75) BM25-scope	BM25-var
SBS.text+ratings.raw	0.0561	0.0634	0.0633
SBS.text+ratings.opt	0.0584	0.0648	0.0648
SBS.text+ratings.rcpr	0.0145	0.0228	0.0230
geoCLEFmod.text+geo.raw	0.1500	0.0728	0.0646
geoCLEFmod.text+geo.opt	0.3425	0.3869	0.3894
geoCLEFmod.text+geo.rcpr	0.1614	0.1773	0.1773

A.5 Overcoming the Long Tail Problem: A Case Study on CO2-Footprint Estimation of Recipes using Information Retrieval

Melanie Geiger, Martin Braschler.
In *Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga* Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 7-12, 2018, Paris: European Language Resources Association (ELRA).

Overcoming the Long Tail Problem: A Case Study on CO2-Footprint Estimation of Recipes using Information Retrieval

Melanie Geiger[1,2], **Martin Braschler**[2]
[1] Université de Neuchâtel, Neuchâtel, Switzerland
[2] Zurich University of Applied Sciences, Winterthur, Switzerland
{imhf, bram}@zhaw.ch

Abstract

We propose approaches that use information retrieval methods for the automatic calculation of CO2-footprints of cooking recipes. A particular challenge is the "long tail problem" that arises with the large diversity of possible ingredients. The proposed approaches are generalizable to other use cases in which a numerical value for semi-structured items has to be calculated, for example, the calculation of the insurance value of a property based on a real estate listing. Our first approach, ingredient matching, calculates the CO2-footprint based on the ingredient descriptions that are matched to food products in a language resource and therefore suffers from the long tail problem. On the other hand, our second approach directly uses the recipe to estimate the CO2-value based on its closest neighbor using an adapted version of the BM25 weighting scheme. Furthermore, we combine these two approaches in order to achieve a more reliable estimate. Our experiments show that the automatically calculated CO2-value estimates lie within an acceptable range compared to the manually calculated values. Therefore, the costs of the calculation of the CO2-footprints can be reduced dramatically by using the automatic approaches. This helps to make the information available to a large audience in order to increase the awareness and transparency of the environmental impact of food consumption.

Keywords: BM25 weighting scheme adaptation, cooking recipe retrieval, CO2-footprint estimation

1. Introduction

One easily measurable quantitative quality criterion of a language resource (LR) is its coverage. However, achieving a high coverage usually requires a lot of human effort. One of the reasons is that often the frequencies of potential LR entries; e.g. words in human language or food products in cooking recipes (Müller et al., 2012), arrange themselves according to the so-called Zipf's law (Zipf, 1949), meaning that most entries relate to entities that occur very infrequently. Therefore, LRs are most likely never complete. This long tail problem is relevant for most applications that rely on LRs, however, it is particularly severe for information retrieval (IR) applications that not only use the LR to enhance their effectiveness (e.g. expanding queries with synonyms) but directly use the LR entries to compile their output.

To illustrate, consider a new class of retrieval applications that require the calculation of a single numerical value from a semi-structured item that consists of a list of textual elements. For example, such a semi-structured item may be a cooking recipe in which the elements are the instruction lines, such as "100g carrots, sliced" and "1 pizza dough", or a real estate listing in which the elements are the components, such as "Bedrooms: 4" and "Heating: Oil-Fired Central Heating". In those examples the numerical values to be calculated can be the nutrition value or the CO2-footprint of a recipe or for the real estate example the insurance value of a property.

An *element-wise* approach to calculating such values splits the problem into sub-problems by first calculating the value for each element individually and then computing the value of the complete item by aggregating the values of the individual elements. For most use cases, this means that the individual elements are matched to an LR, which then helps to estimate their values. In the case of recipes, the LR (Eaternity AG, 2017) we use contains the nutrition value and the CO2-value for each food product. For the estimation of real estate insurance values, a suitable LR contains the costs of the corresponding components; e.g. the average costs of a bathroom with a shower and a double washbasin. This *element-wise* approach, however, heavily relies on the completeness of the LR and has to use a fallback strategy if elements are not found in the LR. In practice, the fallback usually means that additional entries need to be added manually, an excessively costly option.

In the real estate example, an alternative human line of action is often to estimate the value of a property based on the values of other similar properties for which the value is already known. Hence, the value is estimated based on the whole item rather than the individual elements and thus the problem of the incompleteness of the LR can be circumvented. Gonzalez and Laureano-Ortiz (1992) replicate this process for automatic property appraisal. We propose an *item-based* approach using IR technology. We claim that this approach is applicable to many scenarios that include the calculation of a value for a semi-structured item whenever a similarity between the items can be defined.

In this paper, we focus on the use case of the automatic calculation of CO2-footprints of cooking recipes. The motivation for such a use case is that about one-third of CO2-emissions produced by the final household demand in Europe is caused by the consumption of food (Tukker and Jansen, 2006) and that the calculation of CO2-footprints for cooking recipes helps to increase the awareness and transparency of the environmental impact of food consumption. However, so far the footprint of a recipe was calculated with

a manual process (O'Connor et al., 2018) which is time-consuming and therefore too costly to be applied to a wide range of cooking recipes.

We describe and evaluate an *element-wise*, an *item-based*, and a *hybrid* approach, combining the two, to automatically calculate the CO2-footprints of recipes. In the context of our CO2 use case, we call the *element-wise* approach "ingredient matching" and the *item-based* approach "recipe matching". The ingredient matching approach uses an IR pipeline to match the instruction lines to the corresponding entries in the LR through retrieval from an index. The recipe matching approach finds the most similar recipe in a corpus of indexed recipes for which the CO2-footprints are already assessed. A novelty is our proposal of an adapted version of the BM25 weighting scheme which also considers the amounts of the individual ingredients in the recipes. Finally, the hybrid approach combines the two other approaches so that a higher accuracy and stability of the CO2-value estimates can be achieved.

In our experiments, we compare the automatic approaches to the manual process as well as to each other. Both the ingredient as well as the recipe-based approaches perform similarly, while our hybrid approach outperforms the individual approaches. We show that the automatic approaches lie within an acceptable range to the CO2-values calculated manually and therefore are serious alternatives. Using the approaches suggested in this paper, the cost of calculating CO2-footprints of recipes can be reduced dramatically, which makes it possible to make this information available to a large audience. The company Eaternity, which has commercialized a CO2-calculation service based on the approaches we describe, reports that it realizes a reduction in the calculation effort of 50-60% and an overall cost reduction of 80% compared to their old, manual process.

2. Related Work

Processing and more specifically choosing, designing, adapting and comparing cooking recipes has proven popular with case-based reasoning (CBR) researchers ever since the two automated meal recommendation systems CHEF (Hammond, 1986) and Julia (Hinrichs, 1989) have been presented. Many efforts are related to the Computer Cooking Contest, which runs since 2007. We distinguish between work to automatically process the ingredients of cooking recipes and work that deals with the similarity of recipes.

Several publications deal with automatically constructing a process flow graph of a given recipe (Hamada et al., 2000), (Walter et al., 2011). Hamada et al. (2000) create domain-specific dictionaries and match the keywords in the recipe to the words in the dictionaries. Based on the structure of the sentences they then construct the process flow graph. Walter et al. (2011) preprocess and annotate the recipes with GATE, a natural language processing (NLP) framework. Based on rules created from a domain expert the ingredients, as well as the actions, are linked to a workflow. Moreover, Müller et al. (2012) automatically match the ingredients of a recipe to a nutrition database in order to estimate the nutritional value of the recipe. The similarity of recipes is mostly investigated for content-based recommender systems (Teng et al., 2012), (van Pinxteren et al., 2011).

The CO2-database that we use in our experiments as well as the whole CO2-application is described by O'Connor et al. (2018), while other CO2-reduction experiments that are conducted using the automatic ingredient matching approach are described by Itten et al. (2018).

Gonzalez and Laureano-Ortiz (1992) propose a CBR system that automatically estimates the value of a property based on similar real estates handled in past experiences. If the markets for particular properties are too sparse, they use heuristic knowledge.

The K-nearest neighbor (kNN) approach is usually applied to solve classification problems where the only prerequisite is the definition of a similarity of feature vectors. It was first mentioned in a technical report in 1951 (Fix and Hodges Jr, 1951). Since then, kNN is also used for text classification amongst others by Yang (1999) and Sebastiani (2002). In this paper, we do not classify the recipe but only use the idea of nearest neighbors in order to estimate the CO2-value based on them.

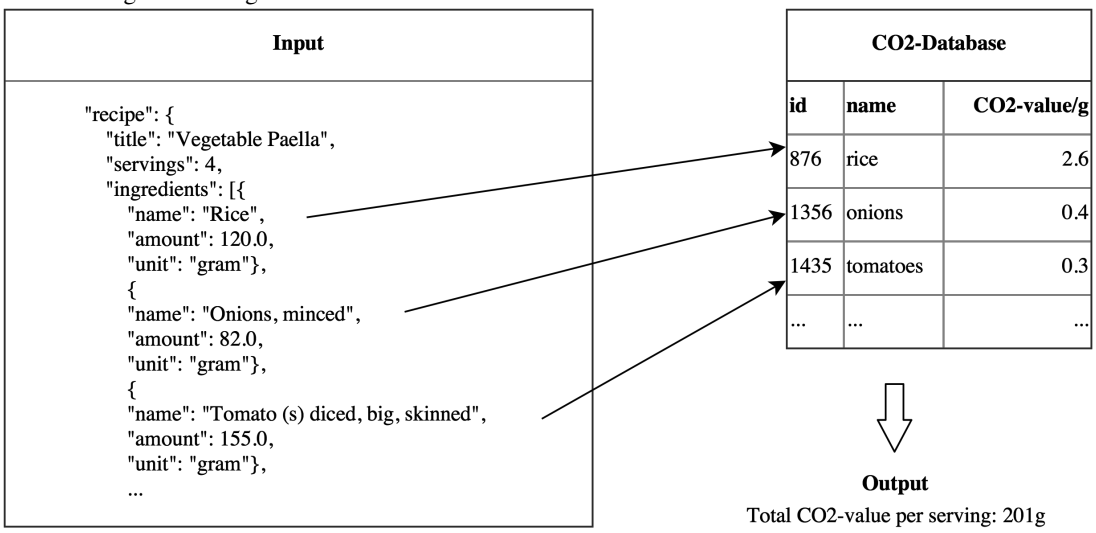
3. Methods

The case of calculating CO2-footprints is interesting for IR research on multiple fronts. As described above, in an *element-wise* approach the value is retrieved by either manually or automatically matching all the ingredient descriptions in the recipe to the appropriate food products in an LR. However, this matching is more challenging than it may appear at first glance. The difficulty stems from the fact that recipes are usually written in natural language and are therefore not restricted to use the fixed vocabulary used in the food product database. The following challenges are all very well known in NLP and IR.

The first challenge is that the authors might use synonyms in order to describe the ingredient; e.g. the ingredient description for "chard" in German might be either "Mangold" or "Krautstiel". In the cooking domain, synonyms are frequently used due to regional differences. The second challenge is the specificity of the ingredient description; both cases, very unspecific descriptions and overly specific descriptions are hard to handle correctly. For example, the unspecific description "fillet of fish" has a lot of different options to be interpreted by the cook and therefore the correct assignment to a food product in the database is a non-trivial choice. On the other hand, the description "Pinot Noir" may be too specific and in order to correctly match it to a food product in the database, the fact that this is a red wine has to be known. The third challenge is the handling of combined products, such as a pizza dough, which themselves consist of several other products.

In order to match the combined products to the food products in the database, the database either needs to contain them as well or a process to recursively split them into their base food products has to be defined. In addition to the three challenges described, we have to handle different word forms, word compounds, special characters, etc. Moreover, the food products used in recipes worldwide are manifold and it has been shown that most of them only appear in very few recipes (Müller et al., 2012). Also, new

Figure 1: Visualization of the manual as well as the automatic ingredient matching approach on an explanatory excerpt of a recipe. Note that the CO2-values are simplified for this example. In the real application, they differ depending on the season and the origin of the ingredients.



products are continuously introduced to the marketplace. According to these facts, a food product database is basically never complete.

In the following sections, we propose three approaches to automatically calculate the CO2-value of cooking recipes using NLP and IR methods. Section 3.1. describes our first approach, which we call *ingredient matching*. It reproduces an automatic version of the traditional manual process of assigning CO2-values to ingredient descriptions. Section 3.2. describes the *recipe matching* approach, which estimates the CO2-value of a recipe based on similar recipes rather than the individual ingredients. Therefore, it does not depend on the completeness of the food product database. Our third approach, the *hybrid approach*, combines the ingredient and the recipe matching approaches and therefore benefits from the advantages of both. It is described in Section 3.3.

3.1. Ingredient Matching

The ingredient matching approach matches all ingredient descriptions individually to the corresponding food product in the CO2-database (Eaternity AG, 2017) and the estimate is computed by the sum of the CO2-values of each food product multiplied with the corresponding amount. As shown in Figure 1, the input is a semi-structured recipe and the output is a mapping of the ingredients to the food products in the CO2-database as well as the total CO2-value of the recipe per serving.

For each ingredient description in the German input recipes, we find the best matching food product in the database. Therefore, we create an index of food products using a traditional IR pipeline including stemming, stopword removal, decompounding and synonym handling. Apart from the commonly used stopwords, we also use several domain-specific stopwords such as pasteurized, portion and minced. To ensure a high precision for the matching, we use a light stemmer (Savoy, 2002). Since the experiments are based on German recipes, we employ a decompounding component that splits compound words such as "Zitronensaft" (lemon juice) into their components "Zitrone" (lemon) and "Saft" (juice) using a dictionary-based n-gram decompounder with a minimum word size of 10 and a minimum and maximum word constituent size of 4 respectively 12. Along with the original food products, we also index their synonyms with the same CO2-information. Moreover, we also add combined products to the index, for which the CO2-values are manually pre-calculated.

Table 1: Notation used for BM25 and our adaptations thereof.

d_j	single document
q	single query
Φ	indexing vocabulary
φ_k	single indexing feature
l_j	length of document
Δ	average document length
$\Phi(d_j)$	set of features representing document d_j
$\Phi(q)$	set of features representing query q
$w(\varphi_k, d_j)$	weight of feature φ_k for document d_j
$w(\varphi_k, q, d_j)$	weight of feature φ_k for query q and d_j
$\text{ff}(\varphi_k, d_j)$	feature frequency of feature φ_k for d_j
$\text{df}(\varphi_k)$	document frequency of feature φ_k

The search in the index is performed using an adaptation of the BM25 weighting scheme (Robertson and Zaragoza, 2009) that ignores the inverse document frequency. Unlike in most other IR applications, the fact that a term appears often in the collection does *not* mean that it is less important. For example, the database might contain several products containing the term "apple", such as apple, apple juice, and apple puree. However, the terms juice and puree should not be weighted heavier than apple, since a match to one of the three food products containing apple is already a much

better fit than a match to for example orange juice. On the other hand, the term frequency is needed since some ingredient descriptions contain the same stemmed term multiple times and thus we assume that it is indeed more important than others. Since the number of terms in the ingredient description varies, we apply a document length normalization. Hence, we use the retrieval status value (RSV) of document d_j w.r.t. query q according to BM25

$$w(\varphi_k, d_j) := \frac{\text{ff}(\varphi_k, d_j)}{k_1((1-b) + b\frac{l_j}{\Delta}) + \text{ff}(\varphi_k, d_j)} \cdot \log\left(\frac{0.5 + N - \text{df}(\varphi_k)}{0.5 + \text{df}(\varphi_k)}\right) \quad (1)$$

$$w(\varphi_k, q) := \text{ff}(\varphi_k, q) \quad (2)$$

$$\text{RSV}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) \cdot w(\varphi_k, q), \quad (3)$$

where we set $\text{df}(\varphi_k) = 1$ for all features φ_k. Table 1 shows an overview of the notation used. Apart from the ignored inverse document frequency, we employ BM25 with the commonly used term frequency saturation parameter $k_1 = 1.2$ and document length normalization parameter $b = 0.75$. The default parameters are used due to the lack of suitable training data and to avoid overfitting. For the ingredient descriptions for which no food product can be retrieved from the index, we assign an artificial food product that has an average CO2-value.

3.2. Recipe Matching

The goal of our recipe matching approach is to estimate the CO2-footprint of an arbitrary recipe from the most similar recipe in a database of recipes for which the CO2-footprints are already known. Hence, we exploit the knowledge we already gathered with either a manual or a semi-automatic process that allocates the CO2-values. Unlike the ingredient matching, the recipe matching does not rely on assigning the individual ingredients to a database entry. Therefore, this nearest neighbor approach overcomes the long tail problem introduced by the incompleteness of the ingredient database.

An approach using IR techniques to find the most similar recipe is to run a textual search with the description of the ingredients in the query recipe against an index in which the recipes are indexed with all their ingredient descriptions. However, the similarity used in this approach does not reflect the amounts of the ingredients in the recipes, e.g. a recipe with 500g flour and 3g salt would be similar to a recipe with 500g salt and 3g flour.

Therefore, we suggest a method that also considers the amounts as an additional information, so that recipes have a higher similarity if the difference between the amounts of their respective ingredient descriptions is small. Our approach is based on an adjustment of the BM25 weighting scheme although other popular weighting schemes such as language models or divergences from randomness could be used. We adapt the weight of the query terms so that the difference between amounts of the ingredients in the query recipe and the document recipes is considered. A query term, i.e. an ingredient description, is weighted with the reciprocal difference between the amounts of the two the ingredients. Hereby, we choose the formula so that a difference of zero leads to a weight of one. Therefore, we also store the amounts of each term in each recipe, so that we can quickly retrieve the amount of a term in a given recipe when comparing recipes. In case an ingredient description consists of several terms, the amount of the ingredient will be assigned to each of its terms.

The retrieval status value (RSV) of document d_j w.r.t. query q of the adjusted BM25 that considers the amounts of the ingredients is therefore defined as:

$$w(\varphi_k, d_j) := \frac{\text{ff}(\varphi_k, d_j)}{k_1((1-b) + b\frac{l_j}{\Delta}) + \text{ff}(\varphi_k, d_j)} \cdot \log\left(\frac{0.5 + N - \text{df}(\varphi_k)}{0.5 + \text{df}(\varphi_k)}\right) \quad (4)$$

$$w(\varphi_k, q, d_j) := \frac{\text{ff}(\varphi_k, q)}{|\text{a}(\varphi_k, q) - \text{a}(\varphi_k, d_j)| \cdot \alpha + 1} \quad (5)$$

$$\text{RSV}_{\text{BM25}}(q, d_j) := \sum_{\varphi_k \in \Phi(q) \cap \Phi(d_j)} w(\varphi_k, d_j) \cdot w(\varphi_k, q, d_j), \quad (6)$$

where $\text{a}(\varphi_k, r)$ is the amount of the term k in the recipe r and α is a tuning parameter to weight the difference between the amounts. The *tf* saturation parameter k_1 and the document length normalization parameter b are used as in the original definition of BM25. Note, that only the definition of $w(\varphi_k, q, d_j)$ is different to the one in the original BM25, where it is equal to $\text{ff}(\varphi_k, q)$.

Once the most similar recipe is known, we can use its CO2-value as an approximation of the CO2-value of the input recipe.

3.3. Hybrid Matching

The use of a hybrid approach is motivated by the failure analysis of the two individual approaches. Our goal is to obtain a more robust estimate that reduces the number of outliers, where the automatically generated value is far from the correct, manual assessment. Table 2 summarizes the reasons why the ingredient matching and recipe matching approaches produce inaccurate estimates which are outside of an acceptable range with respect to the manually computed value. The ingredient matching results in a bad estimate when one or several ingredient descriptions can not be matched to the correct food product in the database. The reason is either that the correct food product does not exist in the database (long tail problem) or that the IR pipeline fails to retrieve the correct food product. Generally, the estimates do not lie within an acceptable range either if many ingredient descriptions are not correctly matched; i.e. the error accumulates; or if a few ingredient descriptions with a high CO2-impact are matched to food products with a low CO2-impact or vice versa.

There are three main reasons for the recipe matching to produce an estimate that does not lie within an acceptable range. The first reason is that the search space in which the recipe matching approach finds the nearest neighbor often has regions in the vector space in which it is not dense. In these regions, the distance between the recipe and its nearest neighbor is bigger than in other regions where the search

Table 2: Summarized reasons for estimation errors.

Ingredient Matching	Recipe Matching
Long Tail Problem	Sparse Space Problem
IR Pipeline Problem	IR Pipeline Problem
	Granularity Problem

space is less sparse. In the recipe domain, the different regions in the vector space might also correspond to cultural differences. For example, our test collection contains a lot of Swiss menus and not so many Asian recipes; therefore, in general, the estimates for Asian menus are less accurate than for Swiss menus. The second reason for bad estimates is that the true nearest neighbor can not be retrieved since it uses a different vocabulary. The third reason for estimates that are far from the manually calculated value can be summarized as granularity problems. This means that the recipe matching, which operates on the whole recipe rather than on the individual elements, fails to produce a good estimate if the nearest neighbor recipe is similar to the input recipe, although there are small but decisive differences in the ingredients that lead to a completely different CO2-footprint. The different kinds of failures of the two approaches lead to situations where either only one of the approaches produces an estimate that is rather far from the ground truth or that one overestimates and the other underestimates the CO2-value. Therefore, we propose a hybrid matching approach that uses the average of the two CO2-estimates of the ingredient matching and the recipe matching as a new estimate, as shown in equation 7.

$$\text{estimate}_{\text{hybrid}} = \frac{\text{estimate}_{\text{ingredient}} + \text{estimate}_{\text{recipe}}}{2} \quad (7)$$

This flattens the outliers produced by the individual approaches and makes sure the system can provide a CO2-estimate for more recipes.

4. Test Collections and Language Resources

For the experiments, we use two collections of recipes that were created specifically for this task. The first collection, the so-called *hobby collection*, consists of 243 vegetarian and vegan recipes from chefkoch[1], an online platform for recipes. The second collection, the *catering collection*, contains 600 recipes from the catering company "Compass Group (Schweiz) AG", a subsidiary of Compass Group PLC, the largest caterer worldwide. The recipes in both collections are in German and are in a semi-structured form given by either chefkoch or the catering company's enterprise resource planning system. This means, each instruction line is provided with separate fields for amount, unit and ingredient description, hence no information extraction is needed. Different units, such as the number of teaspoons, are converted to grams using a simple set of rules.

The ingredient matching approach matches the ingredient descriptions to a product index that contains 3,121 food products that was generated from the LR (Eaternity AG, 2017). The LR contains base food products with their CO2-values as well as synonyms that are linked to the base food products. The LR contains a lot of very region specific food products, such as "Cervelat" and "Roesti" which are frequently used in cooking recipes in Switzerland. The recipe matching approach searches for the nearest neighbor recipes in a recipe index that contains approximately 50,000 recipes. Most of the recipes are from catering companies others are from chefkoch and various other sources. Both the product index and the recipe index are primarily in German.

We manually built a ground truth for both the hobby and the catering collections. That means that for each ingredient in each recipe we manually assigned the best matching food product in the product index. Based on this ground truth it is possible to calculate the CO2-footprint of each recipe in the collection. For example, "spaghetti carbonara" has an expected CO2-value of 774g. There are also recipes with a much larger CO2-value such as "schnitzel with french fries" which has a CO2-value of 2,366g. Table 3 shows the range of the CO2-values of the recipes in these collections. The hobby collection has a significantly smaller average CO2-value per recipe (1,100g) than the catering collection (1,700g) since the hobby collection only contains vegan and vegetarian recipes.

Table 3: Statistics of the test collections.

Collection	Catering	Hobby
Number of recipes	600	243
Minimum CO2-value	32g	113g
Maximum CO2-value	13,513g	1,732g
Average CO2-value	1,700g	1,100g

5. Experiments

5.1. Ingredient Matching

The ingredient matching approach matches the ingredient descriptions to the food products in the database. Table 4 shows the precision, the fraction of correctly matched ingredient descriptions, as well as the mean absolute error (MAE) and the Pearson correlation between the CO2-value estimate from the ingredient matching and the CO2-values from the manual matching. Hereby, correctly matched means that the automatic matching is strictly equal to the manual matching. The mean absolute error is the average of all the absolute errors in the test collection, where the absolute error of a recipe is the difference between the expected CO2-value and our estimate. For example, "spaghetti carbonara" has an expected CO2-value of 774g and an estimate of 684g which results in an absolute error of 90g. The Pearson correlation measures the linear dependence between two variables, in our case the manually assessed CO2-values and the estimates from the automatic process. The possible values are between 1 and -1, where 1 is the maximal positive correlation, 0 means no correlation and -1 is the maximal negative correlation.

The precision for the catering collection is slightly higher than for the hobby collection, mostly since the food product database was mainly designed for catering recipes. At

[1] http://www.chefkoch.de/

Table 4: Matching results using the ingredient matching approach on the two test collections *catering* and *hobby*.

Collection	Catering	Hobby
Precision	0.72	0.68
Mean absolute error	336g	163g
Pearson correlation	0.81	0.73

first glance, the achieved precisions of 0.72 respectively 0.68 are not that encouraging. However, given that other studies show that even the consensus of human assessors is smaller than 75% for 23% of the recipes (Müller et al., 2012), the achieved precision is at least acceptable. Having a closer look at some of the wrongly matched ingredients descriptions, we indeed find many examples which are within the margin of human disagreement. For example, "celery large" is wrongly matched to "celery root" instead of "celery stalks" as denoted in the ground truth, although both seem to be valid options. There are however also some IR specific issues. For example, "red trout fillet (breed)" is wrongly matched to "salmon trout (breed, fillet)" rather than "trout".

The significantly smaller average CO2-value per recipe in the hobby collection, as shown in Table 3, is the main reason why the MAE of the hobby collection is smaller than the MAE of the catering collection.

5.2. Recipe Matching

The recipe matching approach, in which we estimate the CO2-value of an input recipe by its most similar recipe, heavily relies on the size of the recipe corpus from which the similar recipes are retrieved. Our retrieval system is built on top of Lucene and is using the built-in BM25 weighting scheme with the default saturation parameter $k_1 = 1.2$ and the document length normalization parameter $b = 0.75$.

Table 5 shows the MAE and the correlation between the CO2-value estimate from the recipe matching and the CO2-values from the manual matching. For the experiments, we use $\alpha = 0.02$ as the tuning parameter of the weight of the difference between the amounts. Note that our ground truth does not include the closest neighbor of the recipes, but only the manually assigned food products and the total CO2-value of each recipe, thus we do not specify the precision for the recipe matching approach.

Table 5: Matching results using the recipe matching approach on the two test collections *catering* and *hobby*.

Collection	Catering	Hobby
Mean absolute error	310g	360g
Pearson correlation	0.83	0.14

In order to explain the different performances of the algorithm on the two datasets, we first have a look at the two collections. As already stated previously the hobby collection only contains vegan and vegetarian recipes from a hobby cooking platform, while the catering collection contains recipes from several canteens in Switzerland. Having a closer look shows that the two collections are quite different regarding the number of ingredients used in each recipe. An average recipe in the hobby collection consists of 12.7 ingredients and an average recipe in the catering collection has 20.5 ingredients. However, not only the number of ingredients is different but also the ingredients themselves. Therefore the most similar recipe from which the CO2-value is used as an estimate is most likely a recipe from the same category (hobby or catering) as the input recipe. The corpus used to retrieve the recipes with already allocated CO2-values consists of approximately 50,000 recipes from which only around 1% are recipes from the hobby domain, while all the others stem from the catering domain. The lack of close neighbors; i.e. too few recipes from the hobby domain, therefore explains the small correlation (0.14) of estimates in the hobby collection. Even though the performance of the recipe matching for the hobby collection is not as good as for the catering collection, the MAE for the hobby collection (360g) is still in the same range as for the catering collection (310g) due to the smaller average CO2-value of the vegan and vegetarian hobby recipes.

5.3. Hybrid Matching

The hybrid matching approach combines the ingredient matching and recipe matching by averaging the two estimates and therefore is able to account for their individual shortcomings. Table 6 shows the MAE and the correlation between the CO2-value estimate from the hybrid matching and the CO2-values from the manual matching.

Table 6: Matching results using the three matching approaches on the two test collections *catering* and *hobby*.

Method	Measure	Catering	Hobby
Ingredient	Precision	0.72	0.68
	Mean absolute error	336g	163g
	Pearson correlation	0.81	0.73
Recipe	Mean absolute error	310g	360g
	Pearson correlation	0.83	0.14
Hybrid	Mean absolute error	279g	206g
	Pearson correlation	0.90	0.55

For the catering collection the hybrid matching approach achieves a better result for both measures (MAE and correlation) than the other approaches individually. In spite of the significantly worse performance of the recipe matching for the hobby collection the hybrid matching only achieves a slightly worse result as the ingredient matching. These results show that in the case in which both individual approaches achieve an acceptable performance the hybrid matching results in more reliable estimates.

6. Conclusions

We proposed three approaches using IR to automatically compute a single numerical value of a semi-structured item that consists of a list of textual elements based on the use case of calculating CO2-footprints of cooking recipes,

The first approach, ingredient matching, calculates the CO2-footprint on an element-basis; i.e. the ingredients.

Our experiments show that the CO2-value estimates of the ingredient matching lie within an acceptable range compared to the estimate of the manual calculation. The second approach estimates the CO2-values based on similar recipes rather than individual ingredients. Since the estimate is no longer based on the individual ingredients, this recipe matching approach overcomes the long tail problem of the ingredient matching, i.e. that the food product LR is most likely not complete.

For the similarity of recipes, we proposed an adaptation of the BM25 weighting scheme that takes the different amounts of the ingredients into account. We showed that the recipe matching slightly outperforms the ingredient matching, if the recipe corpus is large enough. We have reason to believe, that the effectiveness of matching would increase as the size of the collection of recipes that is searched against increases.

Combining both the ingredient matching and the recipe matching with our hybrid approach allows us to estimate the CO2-value even more accurately. It is therefore able to balance the shortcomings of the individual approaches. The achieved correlation of 0.9 between the CO2-value estimates of the hybrid matching and the CO2-value estimates of the manual matching shows that the automatic calculation is a serious alternative to the manual calculation and therefore the costs of a manual calculation can be reduced dramatically by instantiating the automatic calculation. Indeed, first experiences from using the approaches in the commercial CO2-calculation service of our partner Eaternity indicate a reduction in effort for the calculations in the range of 50-60%, with an even higher overall cost reduction of 80%.

As a next step, the accuracy of the estimates of the hybrid matching approach could possibly be further improved by weighting the estimates of the ingredient and the recipe matching based on an estimate of their reliability. The reliability of the CO2-estimates of the recipe matching could for example be predicted using the distance between the input recipe and its nearest neighbor.

7. Acknowledgments

We thank Compass Group (Schweiz) for the recipe dataset and Eaternity AG for the project work in which most of the approaches presented in this paper were developed.

8. Bibliographical References

Fix, E. and Hodges Jr, J. L. (1951). Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, DTIC Document.

Gonzalez, A. J. and Laureano-Ortiz, R. (1992). A case-based reasoning approach to real estate property appraisal. *Expert Systems with Applications*, 4(2):229 – 246.

Hamada, R., Ide, I., Sakai, S., and Tanaka, H. (2000). Structural analysis of cooking preparation steps in japanese. In *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 157–164. ACM.

Hammond, K. J. (1986). Chef: A model of case-based planning. In *AAAI*, pages 267–271.

Hinrichs, T. R. (1989). Strategies for adaptation and recovery in a design problem solver. In *Proc. of the 2 nd Workshop on Case-Based Reasoning*, pages 115–118.

Itten, R., Imhof, M., Jaggi, A., and Stucki, M. (2018). *Combining Natural Language Processing and Life Cycle Assessment for Computer-Aided Optimisation of Greenhouse Gas Emissions in System Catering*. unpublished manuscript.

Müller, M., Harvey, M., Elsweiler, D., and Mika, S. (2012). Ingredient matching to determine the nutritional properties of internet-sourced recipes. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 73–80. IEEE.

O'Connor, I., Aleksandrowicz, A., Scheuss, A., Jaggi, A., Klarmann, M., and Ellens, J. (2018). *Description of the Eaternity Database*. unpublished manuscript.

Robertson, S. and Zaragoza, H. (2009). *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Savoy, J. (2002). *Morphologie et recherche d'information*. Université de Neuchâtel, Faculté de droit et des sciences économiques, Division économique et sociale.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Teng, C.-Y., Lin, Y.-R., and Adamic, L. A. (2012). Recipe recommendation using ingredient networks. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 298–307. ACM.

Tukker, A. and Jansen, B. (2006). Environmental impacts of products: A detailed review of studies. *Journal of Industrial Ecology*, 10(3):159–182.

van Pinxteren, Y., Geleijnse, G., and Kamsteeg, P. (2011). Deriving a recipe similarity measure for recommending healthful meals. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 105–114. ACM.

Walter, K., Minor, M., and Bergmann, R. (2011). Workflow extraction from cooking recipes. In *Proceedings of the ICCBR 2011 Workshops*, pages 207–216.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69–90.

Zipf, G. K. (1949). Human behavior and the principle of least effort.

9. Language Resource References

Eaternity AG. (2017). *Eaternity Food Product Database*. Eaternity AG, not publicly available, 1.0.