University of Glasgow | School of Computing Science

Microblog Retrieval Challenges and Opportunities

Jesus Alberto Rodriguez Perez
School of Computing Science
University of Glasgow

A thesis submitted for the degree of
Doctor of Philosophy (Ph.D)

12 January 2018

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University.

This dissertation is the result of my own work, under the supervision of Professor Joemon M. Jose and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Permission to copy without fee all or part of this thesis is granted provided that the copies are not made or distributed for commercial purposes, and that the name of the author, the title of the thesis and date of submission are clearly visible on the copy.

Jesus Alberto Rodriguez Perez
January, 2018

I dedicate this doctoral work to my family, who have always and unconditionally supported me during the best and worst times of this adventure. Particularly:

To Victoria Perez Rodriguez
To Jose Remedios Rodriguez Jimenez
To Jose Diego Rodriguez Perez
To Felix Daniel Rodriguez Perez
To Manuel Angel Rodriguez Perez
To John Sebastien Bartholomew
And Elisa Vignaga

I would have not managed without your help, love and positive reinforcement.

Abstract

In recent years microblogging services have changed the way we communicate. Microblogs are a reduced version of web-blogs which are characterised by being just a few characters long. In the case of Twitter, messages known as *tweets* are only 140 characters long, and are broadcasted from followees to followers organised as a social network. Microblogs such as tweets, are used to communicate up to the second information about any topic. Traffic updates, natural disaster reports, self-promotion, or product marketing are only a small portion of the type of information we can find across microblogging services. Most importantly, it has become a platform that has democratised the communication channels and empowered people into voicing their opinions. In fact, it is a very well known fact that the use Twitter amongst other social media services tilted the balance in favour of ex-president Obama when he was elected president of the USA in 2012. However, whilst the widespread use of microblogs has undoubtedly changed and shaped our current society, it is still very hard to effectively perform simple searches on such datasets due to the particular morphology of its documents. The limited character count and the ineffectiveness of state of the art retrieval models in producing relevant documents for queries, thus prompted TREC organisers to unite the research community into addressing these issues in 2011 during the first Microblog 2011 Track.

This doctoral work is one of such efforts, and its focused on improving the access to microblog documents through ad-hoc searches. The first part of our work individually studies the behaviour of the state of the art retrieval models when utilised for microblog ad-hoc retrieval. First we contribute with the best configurations for each of the models studied. But more importantly, we discover how query term frequency and document length relates to the relevance of microblogs. As a result, we propose a microblog specific retrieval model, namely MBRM, which significantly outperforms the state of the art retrieval models described in this work.

Furthermore we define an informativeness hypothesis in order to better understand the relevance of microblogs in terms of the presence of their inherent features or dimensions. We significantly improve the behaviour of a state of the art retrieval model by taking into consideration these dimensions as features into a linear combination re-ranking approach. Additionally we investigate the role that structure plays in determining the relevance of a microblog, by encoding the structure of relevant and non-relevant documents into two separate state machines. We then devise an approach to measure the similarity of an unobserved document towards each of these state machines, to then produce a score which is utilised for ranking. Our evaluation results demonstrate how the structure of microblogs plays a role in further differentiating relevant and non-relevant documents when ranking, by showing significantly improved results over a state of the art baseline.

Subsequently we study the query performance prediction (QPP) task in terms of microblog ad-hoc retrieval. QPP represents the prediction of how well a query will be satisfied by a particular retrieval system. We study the performance of predictors in the context of microblogs and propose a number of microblog specific predictors. Finally our experimental evaluation demonstrates how our predictors outperform those in the literature in the microblog context.

Finally, we address the "vocabulary mismatch" problem by studying the effect of utilising scores produced retrieval models as an ingredient in automatic query expansion (AQE) approaches based on pseudo relevance feedback . To this end we propose alternative approaches which do not rely directly on such scores and demonstrate higher stability when determining the most optimal terms for query expansion. In addition we propose an approach to estimate the quality of a term for query expansion. To this end we employ a classifier to determine whether a prospective query expansion term falls into a low, medium or high value category. The predictions performed by the classifier are then utilised to determine a boosting factor for such terms within an AQE approach. Then we conclude by proving that it is possible to predict the quality of terms by providing statistically enhanced results over an AQE baseline.

Acknowledgements

I want to dedicate this thesis to my family. Their unconditional support and encouragement has been the key ingredient of this work. They supported me when I decided to study abroad initially particularly in an emotional level, giving me the energy to keep fighting when hard times were very present. My mother Victoria, father Jose, and siblings Diego, Felix and Manuel have always been a source of loving words, and never-ending energy which I can never thank enough. I would like to extend this gratitude to my "family abroad" Sebastien Bartholomew and Elisa Vignaga, who have equally been there for me at every step of the way.

Writing this thesis has been a long and hard journey. It as been full of adventures, challenges, but also full of learning experiences. Not only about Information Retrieval, but about myself and my own capabilities and the interaction with others. Thus I want to thank Joemon M. Jose for giving me this opportunity.

My colleagues have been also an important part of this intellectual adventure. Teerapong Leelanupab and Yashar Moshfeghi helped me to "learn the trade" when I first started my PhD. Whilst Ke Zhou (Adam), Stewart Whiting, Philip McParlane, James McMinn, Rami Alkhawaldeh and Stefan Raue with whom I shared all the highs and lows of this long academic endeavour. Also I am deeply grateful to Frank Hopfgartner for playing such an important and inspirational role along the way.

I would also like to thank Marie Jespersen for encouraging me to keep fighting and for telling me "finish your Phd!", even when our paths diverged. These words still resonate in my head, and they will finally be satisfied.

Moreover, I would like to thank Stephanie Malzer for her patience, understanding and relentless encouragement during the hardest times of my doctorial adventure.

Finally, I am very grateful to Natalia Lukaszewicz for giving me that "last push" I needed towards completing my thesis and helping me throughout that final stage.

This doctorial work has been an endeavour in which we have all participated as a team. And for this I will be eternally grateful to all of you.

Contents

List of Figures

Part I

Introduction and Background

Chapter 1

Introduction

1.1 Introduction

The way people communicate and access information has been in a permanent evolution since the beginning of the World Wide Web. However in recent years, we have experienced a "boom" in the uptake of microblogging services spearheaded by Twitter. In fact many mayor political events have been said to be influenced by the utilisation of such platforms. A very recurrent example was the USA elections of 2012 where Obama was praised by his effective use of social media to reach the electorate, thus significantly helping him to win the election [1]. Similarly in later years, a whole political party in Spain attempting to change the current crisis situation within the country - namely Podemos - has been largely self-organising and communicating through the use of social media, including Twitter and Facebook in order to provide voice to those unpopular to the current government[2]. Microblogs have also been useful for the reporting of events such as natural disasters[3] or terrorist attacks, which usually reach the population much faster than traditional communication channels. However the most important aspect of social media is that it provides a unique insight into events, such as first hand reports as events unfold, along with the public opinion of those discussing in real-time.

Above all, social media and particularly microblogging services such as Twitter have been very useful to shorten the distances between people, and to allow people to publicly and freely speak about important issues which affect humanity as a whole. Therefore it is our responsibility to improve the access to microblog services since it will become - or it already is - an essential and unavoidable part of our everyday lives.

Twitter[4] represents the biggest microblogging service in the world. As of 2016, Twitter users generated about 6000 tweets a second, which adds up to around 500 million tweets a day[5]. Twitter is used in a variety of ways, from self-promotion to advertising or real-time news broadcasting and open public discussions. This type of information cannot be found on more traditional sources, as they are more mediated and closed in terms of their content. Also the dynamics produced by the character

[1] https://www.theatlantic.com/technology/archive/2017/01/did-america-need-a-social-media-president/512405/

[2] https://beta.theglobeandmail.com/news/world/digital-innovation-propels-political-success-story-in-spain/article23542220/

[3] https://blog.twitter.com/official/en_in/a/2016/twitter-for-crisis-and-disaster-relief-in.html

[4] https://twitter.com/

[5] https://www.dsayce.com/social-media/tweets-day/

limitations imposed, mixes very well with the current “right here and right now” communication culture of our society. Another important characteristic is the inclusion of socially agreed tags to identify a particular topic, namely hashtags (I.e. #2012Elections) and mentions which refer to an intended audience (I.e. @ObamaThePresident).

1.2 Ad-Hoc Retrieval Task In Microblogs

The predominant methodology to access information in information retrieval (IR) is represented by the ad-hoc retrieval task. The goal of this task is *to return documents that are relevant to an immediate information need expressed as a query posed by a user*. In the context of microblogs, users need to find previously published tweets, or to expand their knowledge on a particular topic by issuing textual queries to a search engine.

However, ad-hoc search in microblogs can be extremely challenging due to the morphology and limited content of the microblog documents in comparison with longer formats such as websites. Microblog messages posted to Twitter (known as *Tweets*) are limited to 140 characters in length. Additionally, tweets present a varied linguistic quality, as they often contain spelling mistakes or slang and abbreviations to overcome the length restrictions. Thus, it is often the case that relevant tweets for a topic are not expressed with the same terminology utilised in the textual query posed by the searching user. This discrepancy is known as the “vocabulary mismatch problem”, and it has been studied in IR as early as in 1987 by Furnas et al. (1987).

Thus given the increasing importance of microblog services to the public, it is no surprise that ad-hoc retrieval has been very actively studied in the context of Twitter since 2011 with the first iteration of the TREC microblog track (Ounis et al., 2011).

1.3 Thesis Statement

Overall, this doctorate work can be organised into five areas under the umbrella of ad-hoc retrieval for microblogs. Firstly we investigate the reasons behind state of the art retrieval models not behaving effectively. We hypothesise that state of the art retrieval models do not appropriately capture the relevance of microblog documents due to their design. We then challenge the previous agreement about the effect of the morphology of microblogs over search performance and confirm how longer documents should be promoted over short ones. As a result we contribute a novel retrieval model - namely MBRM - which significantly better captures a microblog's relevance than other state of the art retrieval models.

Secondly we extend our work by studying what makes a tweet relevant in terms of its dimensions. We define as dimensions the four intrinsic elements that make up almost every tweet, namely text, urls, hashtags and mentions. We design an approach that assigns a score to a tweet by linearly combining statistical evidence from these four dimensions, based on knowledge from a training set. Finally, we contribute a technique based on state-machines to measure the similarity of any given document to known relevant and non-relevant tweets in terms of their structure. We demonstrate how structure similarity can be leveraged to enhance a retrieval model and to improve ad-hoc searches, thus confirming that structure matters in estimating the relevance of a microblog document.

Thirdly, we study the applicability of query performance prediction (QPP) approaches to the context of microblogs. QPP is a task by which the level of success in retrieving the right documents in response to a query is measured, in the absence of any human-annotated relevance judgements. The utility of such approaches is undeniable, as accurate techniques would allow for selective techniques applied to the topics where they are most likely to be successful. On the other hand, it could relieve the use of human-annotated relevance judgements, in retrieval evaluations. In this part we demonstrate the working performance of existing QPP approaches and propose a number of microblog specific predictors which significantly outperform those in the literature.

Finally, we address the “vocabulary mismatch” problem through the application of automatic query expansion (AQE) approaches. We challenge the use of scores produced by retrieval models which is often utilised by AQE approaches based on pseudo revence feedback. Such scores are used by state of the art AQE approaches - such as RM3 - in order to estimate how appropriate are the terms for query expansion, under the assumption *that terms are as good as the documents containing them.* However we believe that those scores can be unreliable, and propose to utilise the discrete rank number for a document in the pseudo relevant set. We then propose two different normalisations which provide a linear and logarithmic discount of the score assigned to terms within a given document with respect to its rank position. We demonstrate that our approach produces statistically improved results over the baseline more often than RM3. In addition, we show how RM3 and our approaches improve the baseline results for different types of topics, thus demonstrating the possibility to utilise selective AQE approaches in the future. Finally, we propose a technique to estimate the quality of terms to be used for AQE. We do so by building a classifier to assign a class to each term found in a pseudo relevant set, and applying a boosting factor to their value, based on their class, thus enhancing the behaviour of the state of the art RM3.

1.4 Research Questions

In this doctoral work we organised our research around a set of research questions. Chapter 3 is driven by the following research questions:

- **RQ1:** How are state of the art retrieval models affected by the morphology of microblog documents in an ad-hoc retrieval scenario?
  - **RQ1.A:** Why do certain models perform better than others in the Microblog domain?
  - **RQ1.B:** What are the best parameters for each state of the art retrieval model in the Microblog domain?
  - **RQ1.C:** Can we build a custom retrieval model to better capture the relevance of documents?

Subsequently, Chapter 4 studies what makes microblog documents relevance, and introduces the following research questions:

- **RQ2:** Can we define informativeness for microblogs in terms of their inherent features?
  - **RQ2.A:** Can we exploit microblog specific features in order to improve ad-hoc retrieval searches?
  - **RQ2.B:** Are there differences between relevant and non-relevant microblogs in terms of their structure? Can we leverage their structure to produce better rankings in ad-hoc searches?

Additionally, Chapter 5 provides an explorative study on query performance predictors within the context of Microblog ad-hoc retrieval, which is driven by the following question:

- **RQ3:** To what extent can we predict query performance during ad-Hoc retrieval of microblog documents?

Moreover, Chapter 6 explores the area of automatic query expansion by answering the following research questions:

- **RQ4:** Are retrieval model scores unreliable when determining the importance of terms in a pseudo relevant set, when utilised by automatic query expansion techniques?
- **RQ5:** Is it possible to predict the importance of a term within a pseudo relevant set before it is used for query expansion? Can this evidence improve AQE approaches?

Research questions **RQ1** and **RQ2** deal with more fundamental issues than the rest. We therefore investigate them in more depth than other questions as evidenced by the posed sub-questions.

1.5 Contributions

In this Section we summarise the contributions resulting from our work, and map them to the related research question.

C1 An investigation into the performance issues of state of the art retrieval models when applied to microblog ad-hoc retrieval (Related to RQ1.A).

C2 A study to determine the best configurations for each state of the art retrieval model considered in this work (Related to RQ1.B).

C3 A novel retrieval model that better adapts to the morphology of microblog documents and significantly outperforms the best state of the art models in the context of microblog ad-hoc retrieval (Related to RQ1.C).

C4 A study into what makes a microblog document relevant from the perspective of its structure. We explore the structure of microblog documents and demonstrate how it can be utilised to significantly improve ad-hoc retrieval (Related to RQ2.A).

C5 A re-ranking approach based on state machine models of the structure of relevant and non-relevant documents. The similarity of an unobserved document to both models is measured and then utilised to re-rank results accordingly (Related to RQ2.B).

C6 A study into applying query performance prediction techniques to the context of microblog retrieval. And the introduction of novel microblog specific predictors which outperform those in the literature (Related to RQ3).

C7 The introduction of Discounting Automatic Query Expansion (AQE) approaches which increment the independence from scores produced by retrieval models in the pseudo relevant set by relying in the rank number of a document in the result list instead. Discounting AQE approaches achieve significantly better results than a given baseline more often than the state of the art RM3, due to the reduced sensitivity to the document scores (Related to RQ4).

C8 A term quality classification approach for AQE. Terms in pseudo relevant set are classified utilising a machine learned classification model and a boosting factor assigned accordingly to them when determining their importance towards being used as expansion terms (Related to RQ5).

1.6 Thesis Roadmap

This doctoral work is divided in five parts, and structured as follows:

- **Part I: Introduction and Background**

  This part is made up of two chapters. Firstly it introduces the importance of microblog retrieval in the context of our current society, and highlights the retrieval challenges of microblogs. Then the objectives and structure of this doctoral work is described in Chapter 1. Secondly, the fundamental information retrieval concepts and background used across this doctoral work is introduced in Chapter 2.

- **Part II: Relevance and Informativeness of Microblogs**

  This part divided into two different chapters. In Chapter 3 we investigate the performance issues experimented by the state of the art retrieval models when applied in microblog ad-hoc retrieval conditions. Additionally we explore the best configurations for each model, and we finalise the chapter by proposing a novel retrieval model. Our microblog retrieval model - namely MBRM - adapts significantly better to the morphology of microblog documents as demonstrated by significantly outperforming the state of the art models in the context of microblog ad-hoc retrieval. Chapter 4 presents a study into what makes a microblog document relevant, from the perspective of its structure. We explore the structure of microblog documents and show how it can be utilised to significantly improve ad-hoc retrieval. Finally, we modelled the transitions between elements of known relevant and non-relevant documents as state machines, and utilised an algorithm to compute a similarity score for an unobserved document which is then utilised as a re-ranking feature.

- **Part III: Query Performance Prediction**

  This part comprises Chapter 5 which introduces a study into applying known query performance prediction techniques in the context of microblog retrieval. A number of microblog specific predictors are also introduced and their performance benchmarked against the state of the art predictors from the literature.

- **Part IV: Automatic Query Expansion**

  This part is composed of two chapters. Chapter 6 challenges the use of scores produced by retrieval models in the pseudo relevant set by automatic query expansion techniques. Additionally it presents a number of Discounting Automatic Query Expansion (AQE) approaches which rely instead on the rank number of documents in the pseudo relevant set. The experimental results show how Discounting AQE approaches achieve significantly better results than a given baseline more frequently than the state of the art RM3. The rest of the chapter proposes a term quality classification approach for AQE. Terms in a pseudo relevant set are classified utilising a machine learned model to determine their quality. These classes are in turn used to provide a boosting factor for each of the terms when applying an AQE approach such as RM3, in agreement with their estimated quality.

- **Part V: Conclusions**

  The whole thesis is summarised and concluded in Chapter 7, where we also highlight our contributions.

1.7 Publications

The following are research publications produced as a result of this doctoral work:

- [Rodriguez Perez and Jose (2015)] Jesus A. Rodriguez Perez, and Joemon M. Jose. "On Microblog Dimensionality and Informativeness: Exploiting Microblogs' Structure and Dimensions for Ad-Hoc Retrieval" Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR). 2015.

- [Rodriguez Perez and Jose (2014)] Jesus A. Rodriguez Perez, and Joemon M. Jose. "Predicting Query Performance in Microblog Retrieval" Proceedings of the 37th Annual ACM SIGIR conference. 2014.

- [Rodriguez Perez et al. (2013b)] Jesus A. Rodriguez Perez, Yashar Moshfeghi, and Joemon M. Jose. "On using inter-document relations in microblog retrieval." Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

- [Rodriguez Perez et al. (2013a)] Jesus A. Rodriguez Perez, Andrew J. McMinn, and Joemon M. Jose. "University of Glasgow (uog_twTeam) at TREC Microblog 2013."

- [Rodriguez Perez et al. (2012a)] Jesus A. Rodriguez Perez, Teerapong Leelanupab, and Joemon M. Jose. "CoFox: A Synchronous Collaborative Browser." Proceedings of the 8th Asia Information Retrieval Societies conference (AIRS 2012). Springer Berlin Heidelberg, 2012. 262-274.

- [Rodriguez Perez et al. (2012b)] Jesus A. Rodriguez Perez, Andrew J. McMinn, and Joemon M. Jose. ”University of Glasgow (uog_tw) at TREC Microblog 2012.”
- [Rodriguez Perez et al. (2011)] Jesus A. Rodriguez Perez, Stewart Whiting, and Joemon M. Jose. ”CoFox: A visual collaborative browser.” Proceedings of the 3rd international workshop on Collaborative information retrieval. ACM, 2011.

Chapter 2

Background

2.1 General Background

This section of the background covers definitions and common literature which will be used throughout this work. Particularly, the following content is essential for the understanding of Part II, as we will explore retrieval problems connection to core concepts regarding search and retrieval models.

2.1.1 Ad-hoc Retrieval

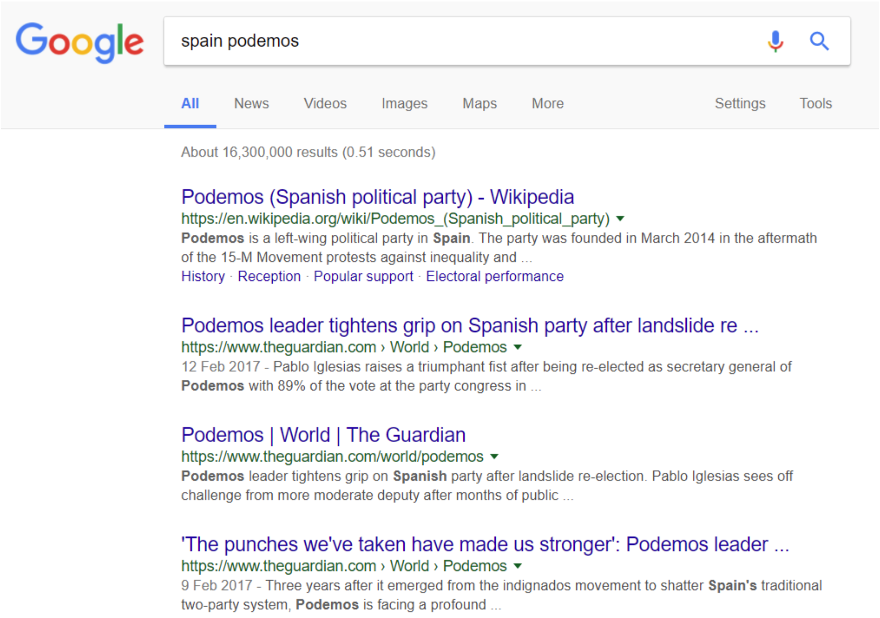
The ad-hoc retrieval task is the most commonly studied task in information retrieval (IR). The main goal is to retrieve documents from a collection that most closely match an information need expressed as a query. The set of retrieved documents should be presented to the searching user in decreasing order of probability to satisfying his information need, also known as "relevance probability". Moreover the relevance probability of each document is considered independent, thus it is not affected by any other retrieved document. This probability is computed by means of a retrieval model which in turn relies on a set of document and collection statistics. The user behaviour is modelled after the assumption that he/she sequentially evaluates the documents in a result set starting from the top of the list, which greatly simplifies the evaluation of retrieval systems (Voorhees and Harman, 2005). An example of an ad-hoc search is that provided by any search engine such as Google[1]. Figure 2.1 shows search results produced by Google for the textual query "spain podemos" as part of an ad-hoc retrieval task. In this particular case, any given user is assumed to assess the results starting from the "Wikipedia" article, and progress downwards on the list.

2.1.2 IR Evaluation

In information retrieval (IR), systems are statistically evaluated and compared in order to measure the level of success of certain approaches over others, and thus measure the progress of novel techniques. Perhaps, the most common evaluation methodology in IR is the Cranfield paradigm (Voorhees and Harman, 2005), and can be summarised as follows:

- Setting up of a collection of documents, to provide a common test set to allow fair comparisons between different approaches or systems

[1] www.google.com



spain podemos

About 16,300,000 results (0.51 seconds)

Podemos (Spanish political party) - Wikipedia
https://en.wikipedia.org/wiki/Podemos_(Spanish_political_party)
Podemos is a left-wing political party in **Spain**. The party was founded in March 2014 in the aftermath of the 15-M Movement protests against inequality and ...
History · Reception · Popular support · Electoral performance

Podemos leader tightens grip on Spanish party after landslide re ...
https://www.theguardian.com › World › Podemos
12 Feb 2017 - Pablo Iglesias raises a triumphant fist after being re-elected as secretary general of **Podemos** with 89% of the vote at the party congress in ...

Podemos | World | The Guardian
https://www.theguardian.com/world/podemos
Podemos leader tightens grip on **Spanish** party after landslide re-election. Pablo Iglesias sees off challenge from more moderate deputy after months of public ...

'The punches we've taken have made us stronger': Podemos leader ...
https://www.theguardian.com › World › Podemos
9 Feb 2017 - Three years after it emerged from the indignados movement to shatter **Spain's** traditional two-party system, **Podemos** is facing a profound ...

Figure 2.1: Ad-hoc search results provided by Google.

- Creation of a set of information needs, commonly known as queries or topics and usually expressed textually. These queries are used as inputs for the systems being evaluated, so they are assessed within the same conditions.
- Gathering of relevance judgements for a particular task. Documents associated with the above-mentioned queries are human-annotated as relevant or non-relevant, and compiled into relevance judgement files.
- Computation of evaluation metrics utilising the relevance judgements for each topic in order to assess the performance of a retrieval system, and allow statistical comparisons with respect to others.

2.1.3 Retrieval Models

Probability Ranking Principle (PRP) is a core concept in Information Retrieval (IR) first introduced by Robertson (1977). In the context of Ad-hoc retrieval, PRP states that documents should be ranked and presented to the user, based on a document's estimated probability of being relevant given a query. Consequently the scores produced by retrieval models with respect to a query are utilised to organise the result list of documents in decreasing order of relevance probability. Thus PRP forms the basis for any retrieval model or retrieval system in which numerous results are to be shown in order of relevance to a given user.

Probability of Relevance Framework. For many years researchers have developed their understanding on estimating the relevance of documents, thus leading to many models and definitions of relevance. One of the most representative works in this area of research is the Probability of Relevance Framework (PRF) (Roelleke, 2013). PRF is formulated by $P(r|\hat{d}, q)$, where r refers to relevance, q a given query and $\hat{d}$ represents a document as a vector of features $\hat{d} = (f_1, ...f_n)$. Note that vector features can be any imaginable data. The main importance of this framework is the formalisation of relevance as a function of a given query and document vectors. This can be utilised as a framework for any probabilistic retrieval model, thus becoming the basis of numerous research works. It is worth mentioning that the relevance probability of a document may depend on other previously observed documents. However in most IR evaluations the relevance of documents are assumed to be independent from each other, as we also do in this work.

Document length normalization. The work of Singhal et al. (1996) has been employed by retrieval models to counterbalance the effects of longer documents, which may not necessarily add any new information to a topic, but are prone to contain higher term frequencies. In line with this effort, the design of BM25 by Robertson and Zaragoza (2009) involved the study of document characteristics, resulting in the definition of the **scope** and **verbosity** hypotheses. The **verbosity** hypotheses supports that some documents are more verbose than others, thus applying length normalization by dividing by the length of the document is beneficial to better capture relevance, as repetition of terms is superfluous. On the other hand, the **scope** hypotheses states that some authors simply have more to say, thus adding more relevant information to the topic and occupying more space. BM25 applies a soft normalisation that takes into account both cases. As we will evidence in Chapter 3, the study of document length normalization is of particular interest in the context of microblog documents due to their substantially limited length in comparison to other documents.

Inverse Document Frequency (*IDF*) is an estimation of the discriminatory power of a query term. That is, a term q_i is more discriminatory of a particular document d than another term q_j if, is less likely to appear in other documents than q_j. Therefore the highest IDF score for any given collection, belong to those terms appearing in a single document.

$$\text{IDF}_{t,D} = \log_2 \left(\frac{|D|}{|d \in D : t \in d|} \right) \tag{2.1}$$

where t is the current term, D is the set of all documents in the collection and $|d \in D : t \in d|$ is the number of documents in which term t occurs. *IDF* is commonly used as a component of retrieval models such as TF*IDF (Salton and Buckley, 1988). In this model, IDF is used in conjunction with term frequency (TF).

$$\text{TF*IDF}_{t,d,D} = \text{TF}_{t,d} * \text{IDF}_{t,D} \tag{2.2}$$

In Chapter 3 we study the behaviour of the state of the art retrieval models in the context of microblog ad-hoc retrieval. We simulate their behaviour in order to draw conclusions as to why they fail to properly capture the relevance of microblogs, and as an outcome produce a microblog specific retrieval model.

2.1.4 Automatic Query expansion

Automatic query expansion approaches (AQE) approaches have been the focus of research effort for many years, as it has been shown to be effective in alleviating the vocabulary mismatch problem. This problem arises from the difference in the textual representation of documents and queries. Given a textual query, the relevant documents may not include the set of terms defined by the searching user, thus the required documents may not be retrieved. Work by Carpineto and Romano (2012) produced a comprehensive study about these approaches, giving insight on the challenges that AQE approaches face. Most importantly it introduces critical issues such as parameter setting, efficiency and usability of the approaches which has greatly contributed to the design of our own query expansion approaches based on Pseudo Relevance Feedback.

Pseudo Relevance Feedback (PRF) (Xu and Croft, 1996) is a technique used in automatic query expansion which - given an initial query - assumes the top N retrieved documents to be relevant. Since there is no certain knowledge about their relevance, they are called pseudo-relevant. Terms are then extracted from the top N terms, then ranked by a scoring function and utilised to expand the original query.

Chapter 6 will expand the background on AQE and explore a set of novel approaches based on pseudo relevance feedback and applied to the context of microblog document retrieval.

2.1.5 Query performance prediction

Query performance prediction (QPP), refers to the study of predictors which can give a performance estimate for a given retrieval model during a retrieval task. Effective QPP can be very useful in many applications such as "selective query expansion". It is well known that AQE approaches based on PRF may worsen the initial retrieval results if the PRF set of documents is not representative of the topic. Thus effective QPP could provide a selective mechanism to prevent worsening the results when the initial retrieval was not good enough.

Consequently QPP has been actively studied in the context of web document retrieval, especially in TREC's Robust track (Voorhees and Harman, 2005). During TREC robust tracks, participants were to come up with systems that better satisfied the highest number of queries. The motivation behind this track was the realisation that many systems returned excellent results for a set of queries yet fared badly for another significant set of queries. In this case it could be comparable in terms of most averaging evaluation metrics, to another system achieving mediocre performance for a wider range of queries. The latter system would be considered more robust than the earlier one, as it is able to satisfy a higher number of queries, even if its performance is not excellent. QPP is an interesting approach towards building more robust systems, as performance predictors can provide an estimation of the success of a system in retrieving relevant documents for a given query. These estimations can in turn be leveraged to apply specific techniques to those badly performing topics.

Examples of QPP works include Zhao et al. (2008) where they defined predictors in terms of pre-retrieval features, or the work by Cronen-Townsend et al. (2002) which proposes “Clarity and Ambiguity” post-retrieval predictors to estimate the retrieval success of a system given a query.

Effective query performance prediction for microblog documents would be an invaluable technique towards enhancing the behaviour of microblog retrieval systems. This is due to the high variability of success experimented for different topics, as we will introduce later (Chapter 5). In Chapter 5 we expand and review existing QPP techniques in the context of microblogs and propose a number of QPP approaches specific to microblogs corpora.

2.2 Microblog Retrieval

Microblog retrieval is very different from other retrieval tasks given the structural differences of microblog documents. Microblogs are generally very short (140 characters in the case of Tweets) compared to other documents and thus they introduce new retrieval challenges. In this section we will cover the literature relevant to microblog retrieval which will serve as a starting point for the rest of this doctoral work.

2.2.1 TREC Microblog Retrieval Tracks

The “Text REtrieval Conference” (TREC[1]) is an internationally recognised conference which is dedicated to the advancement of information retrieval technologies in a diverse number of ways. Sponsored by the National Institute of Technology (NIST) and the U.S. Department of Defense, TREC has run for over 20 years striving to increase the communication between industry, academia and other stakeholders, as well as facilitating large scale system evaluations. Consequently and following the rising importance of microblog documents, TREC organised a number of “Microblog tracks” over four consecutive years 2011-2014 in order to organise the research community and jointly address this retrieval problem.

In order to evaluate the performance of the prospective solutions and allow for comparability they agreed on a collection of documents and a set of topics, as well as relevance judgements on those topics provided by NIST obtained through pooling. To this end they sampled two collections of documents from a Twitter stream over two different periods of time. The first collection was gathered in 2011 but was used for during both the 2011 and 2012 microblog tracks. Similarly, the second collection was

[1] http://trec.nist.gov/

gathered in 2013 and was used for both the 2013 and 2014 microblog tracks. Finally the number of topics varied between 50 and 60 for each of the tracks, totalling 225 topics.

Relevance judgements were gathered by NIST[1] assessors in all iterations of the microblog track. Moreover, the assessors utilised the following set of rules when evaluating the relevance of a tweet (Ounis et al., 2011):

- **Not Relevant**. The content of the tweet does not provide any useful information on the topic, or is either written in a language other than English, or is a retweet.
- **Relevant**. The tweet mentions or provides some minimally useful information on the topic.
- **Highly Relevant**. A highly relevant tweet will either contain highly informative content, or link to highly informative content.

All participants of the microblog track submitted runs containing ranked documents for the agreed topics. The set of relevance judgements were compiled by "pooling" at a depth of 90 (Ounis et al., 2011) and all documents were subsequently evaluated by the assessors as stated above.

The summary results for each of the tracks are presented in Table 2.1 for reference. Amongst the top performing participants we can find Amati et al. (2011); Li et al. (2011); Metzler and Cai (2011) for microblog 2011 and Aboulnaga et al. (2012); Han et al. (2012); Kim et al. (2012) for 2012, which mostly employed query and document expansion techniques as well as learning to rank (L2R) approaches. Additionally, the 2013 track followed a similar trend producing works in the same categories L2R (Gao et al., 2013; Zhu et al., 2013), query expansion (Rodriguez Perez et al., 2013a; Yang et al., 2013) and document expansion (Jabeur et al., 2013). Moreover, the work by Damak et al. (2013) produced a comprehensive summary of the features used by different approaches, and demonstrated how to successfully combine them using Naive-Bayes as an L2R approach combining a number of features including hashtags, mentions, url presence, recency, etc.

Additionally, work by Thomas (2012) studied the effects that preprocessing had on retrieval performance. Their findings showed that the best performance was achieved

[1] National Institute of Standards and Technology

Table 2.1: TREC Tracks results in terms of precision@30

2011		2012		2013		2014	
Best	Median	Best	Median	Best	Median	Best	Median
0.502	0.298	0.470	0.362	0.560	0.370	0.722	0.629

when applying all preprocessing steps, which include (i) language detection, (ii) Emotion removal, (iii) Lexical normalization, (iv) Mention Removal and (v) Link Removal. Additionally, works by Ferguson et al. (2012); Naveed et al. (2011) have identified that problems affecting retrieval models in microblogs are related to *term frequency* and *document length normalization.*

Finally, whilst all the works undertaken by the participants of the Microblog tracks attempted to improve performance retrieval performance by applying their particular set of retrieval techniques, there was no significant attempt to provide an in-depth study on the behaviour of current state of the art retrieval models. Consequently we address this literature gap in Chapter 3, which led us to develop our own microblog specific retrieval model.

2.2.2 Temporal Features In Microblog Retrieval

The work by Efron (2010b) estimated the term weight with respect to its temporal behaviour up until a point in time. In their approach they assign weights depending on how well a term's frequency distribution over time fits a linear model. Their argument follows that more discriminatory terms exhibit a more erratic behaviour in terms of changes of frequency compared to more common terms.

Burst Detection. One of the earliest works to integrate the temporal dimension into a retrieval model is presented by Li and Croft (2003). They identified a set of queries that need to favour recent documents, such as news articles. As their approach they proposed to utilize a recency component along with a language model, in order to offer the most temporally relevant information.

Kleinberg (2003) introduced the use of burst detection in the context information retrieval. His approach models a stream of data as an infinite-state automaton, in which bursts represent state transitions. A burst, caused by a term's frequency surpassing a threshold, causes a transition to another state. His approach scores documents according to the burstiness exhibited by the terms contained within. This approach was evaluated on a collection of short emails and research paper titles spanning over 20 years. More recent approaches build upon Kleinberg's approach to burst detection such as Lee et al. (2011), Shan et al. (2012), Kifer et al. (2004) and Song et al. (2012).

Retrieving events. Metzler et al. (2012) worked in ways to structure and link Twitter documents as retrieval units. To this end they proposed the use of a query expansion approach coupled with a burstiness estimation algorithm, which helped them discover temporal similarities between terms within tweets. Moreover tweets are combined into their "event" retrieval units, which group topically related tweets together to be retrieved as a "structured document".

Other features The use of other features such as temporal evidences in conjunction with geographical locations has been studied by Lappas et al. (2012) and Ishikawa et al. (2012). Finally Weng et al. (2011) proposed that clusters of features which show bursty behaviour in close temporal proximity suggest an event. Their system builds signals for individual features using wavelet analysis for each of the terms. Events are then formed by clustering terms with similar behaviour over time.

2.2.3 Automatic Query And Document Expansion in microblogs

Automatic query expansion (AQE) approaches have been proven to effectively work in the context of web search. Likewise they have also been successfully deployed in the context of Twitter retrieval by a numerous authors such as Whiting et al. (2011) and Lau et al. (2011). Moreover, it was repeatedly utilised amongst the top performing participants during TREC Microblog tracks in their proposed systems for ad-hoc retrieval, often reporting significant improvements on retrieval effectiveness [2011 Amati et al. (2011); Li et al. (2011); Metzler and Cai (2011) and 2012 Aboulnaga et al. (2012); Han et al. (2012); Kim et al. (2012)]. However it was also reported how these AQE approaches can also decrease the performance for some topics, whilst boosting retrieval performance in average. Whilst all these approaches performed well, many of them directly rely on the scores produced by a retrieval model for the promotion of query expansion terms found in pseudo relevant documents. In Chapter 6 we propose that the use of these scores can be misleading due to the unreliability of retrieval models under microblog conditions, and thus we introduce a number of more independent alternatives.

On the other hand, another commonly used technique in the context of microblogs is document expansion. Since microblogs are very short in length, it often means that the information contained within is insufficient to make an informed retrieval decision. Document expansion, attempts to add content to the documents from external sources. The most common approach is to follow the links already published within the documents themselves. Other approaches use the document itself as a query to search for related terms on a commercial search engine such as Google (Bandyopadhyay et al., 2012) or Bing. These approaches benefit from the information contained in external sources by adding information to the tweets, however their reliance on external sources can also be problematic in terms of availability of the external sources.

2.2.4 Social and Semantic features for ad-hoc retrieval

Social features such as hashtags, and mentions have also been utilised in retrieval. An example is the work by Efron (2010a) which focuses on the finding of related hashtags, related to the initial interests reported by a user. In their work the assume that users interested in a particular topic will also be interested in a particular set of hashtags, thus they propose a relevance feedback approach for query expansion based on this relation.

Other authors have also explored semantic features in the context of microblog retrieval. The work by Zingla et al. (2016) proposed to expands queries by leveraging semantic sources such Wikipedia or DBpedia, reporting significant improvement in retrieval performance on the TREC 2011 microblog collection. Similar work was carried out by Zhang et al. (2012), as they provided an automatic query expansion mechanism which utilised the WordNet ontology as a source of semantic evidence in their approach.

2.3 Conclusion

In this chapter we have introduced the relevant background that will be utilised throughout the remainder of this work. We have covered the basics of information retrieval, ranging from IR evaluations to retrieval models or automatic query expansion. Furthermore, we have introduced microblog specific literature as an overview of the most common retrieval approaches and features utilised by other works. The following chapters on this doctoral work contain their own background sections which will provide more specific and contextualised information where required.

The objective of this work is two-fold. We will firstly investigate the problems faced by retrieval models when utilised in the context of microblogs through an in-depth study of their characteristics. Secondly, based on our findings we will explore different approaches to enhance the ad-hoc retrieval of microblog documents.

Part II

Relevance and Informativeness of Microblogs

Chapter 3

Microblog Ad-Hoc Retrieval Problems and MBRM

3.1 Introduction

From the start the main objective of information retrieval has been the understanding and promotion of documents that contain interesting information, discarding those that are unimportant given an information need. To develop this understanding researchers have paid attention to many different features, both at collection and document levels. Eventually the research community has come up with descriptions - namely retrieval models - to match the characteristics of relevant documents. An important example is the okapi BM25 by Robertson et al. (1995), which was presented as a submission to TREC-3 in 1994. Since then it has become a reference retrieval model both for its simplicity and retrieval effectiveness.

During the conception of BM25 the authors explored the characteristics of web documents leading to the formulation of two hypothesis describing a relationship between document length and the frequency of query terms in it.

Likewise other retrieval models were developed as different ways to understand the relevance of a document with respect to a query were conceived. These state of the art retrieval models include Divergence From Randomness (DFR) (Amati et al., 2003); Hiemstra's Language Model (HLM) (Hiemstra, 2001); or Dirichlet Language Model (DLM) (Zhai and Lafferty, 2001).

However information retrieval in Microblogs can be extremely challenging due to the documents morphology and limited content. In the case of Twitter, messages known as *Tweets*, are limited to 140 characters in length and of varied linguistic quality (Teevan et al., 2011). These unforeseen retrieval conditions and challenges propelled novel solutions, mainly spearheaded by the TREC initiative.

The main line of research was to utilise features which are inherent to microblog documents, such as hashtags, metions or URLs coupled with query/document expansion techniques. There were even efforts on learning to rank (L2R) tweets such as Duan et al. (2010). Consequently, there was very little work on analysing the behaviour of state of the art retrieval models in microblog retrieval conditions. Examples of such works are Naveed et al. (2011); Singhal et al. (1996) were they identify mainly how document length is detrimental towards for microblog retrieval. However it is particularly interesting to uncover *how/whether microblog features affect their performance* in order to significantly develop our understanding of the underlying retrieval problems.

To this end, we first elaborate an in-depth study of the behaviour of state of the art retrieval models in the context of Twitter retrieval. The main outcome of the study shows how document length can be effectively leveraged by a retrieval model - contrary to previous belief - thus leading to the conception of a new microblog-specific retrieval model namely **MBRM**, which outperforms the best known baselines in microblog retrieval. We set the focus of our work in the context of these research questions:

- **RQ1.A:** Why do certain models perform better than others in the Microblog domain?
- **RQ1.B:** What are the best parameters for each state of the art retrieval model in the Microblog domain?
- **RQ1.C:** Can we build a custom retrieval model to better capture the relevance of documents?

In order to answer these research questions we simulated behaviour of state of the art retrieval models under microblog conditions. Then we experiment to improve their behaviour through a series of experiments. We then extract our conclusions which will then lead to the creation of a microblog specific retrieval model.

The rest of the chapter is organised as follows. First, we cover a number of related works regarding microblog retrieval and introduce the concepts utilised throughout this work (Section 3.2). Section 3.3 sets the evaluation environment in which our investigation is carried out, giving way to our main analysis in Section 3.4. Section 3.5 introduces a novel microblog-specific retrieval model, and we finalise with the conclusions in Section 3.6 and future research directions.

3.2 Background

In this Section we will introduce concepts and related literature relevant to this chapter.

3.2.1 Retrieval Models

The first part of this work revolves around retrieval models and how their design affects their performance when retrieving microblogs. In our experimentation we include retrieval models such as: Okapi BM25 (Robertson and Zaragoza, 2009); Divergence From Randomness (DFR) (Amati et al., 2003); Hiemstra's Language Model (HLM)

(Hiemstra, 2001); and Dirichlet Language Model (DLM) (Zhai and Lafferty, 2001). These models are introduced in more details in Section 3.4, and their behaviour described individually against microblog conditions. However we first introduce some basic background to ease the understanding of the following sections.

Retrieval of microblogs is hard. Retrieval models are reliant by design on term frequency and document length as the variables to quantify whether a document is more important than other. From a simplified perspective and assuming similar document lengths, a retrieval model will give more importance to a document that contains query terms more frequently than another document. Likewise, when query terms appear a similar number of times, a document will be deemed less or more informative based on the document length. However, microblog documents are limited in length to 140 characters in the case of Twitter. This limitation obviously challenges the above-mentioned rationale, which unfortunately forms the basis of most - if not all - retrieval models. The new medium and the low retrieval performance achieved by state of the art retrieval models gave way to an extensive area of research spearheaded by the Text Retrieval Conference (TREC) through its microblog track. Over recent years, numerous approaches have been proposed which significantly improve retrieval performance in diverse ways.

However to the best of our knowledge no significant progress has been made to understand *why are retrieval models failing* in microblogs. Due to their limited size, document length and term frequencies are often loosely blamed with the underperformance of retrieval models (Ferguson et al., 2012; Naveed et al., 2011). We believe it is important to explore, and properly assess the interaction of such features. Better understanding could lead to improving the performance of existing retrieval models, or new bespoke models altogether.

3.3 Experimental Setting

Datasets. In this evaluation we have used the four collections (2011-2014) from the TREC Microblog track. The 2011 and 2012 collections share the same corpus but have different topics and relevance assessments. On the other hand the 2013 and 2014 collections share the same corpus. The later corpus is an order of magnitude bigger than previous collections. In total there are 225 topics with query lengths ranging from 2 to 3 tokens, in line with the literature (Teevan et al., 2011). Refer to Table 3.1 for an

Table 3.1: Descriptive statistics for the collections being used in this study

TREC Microblog track collection year	2011	2012	2013	2014
Number of topics	50	60	60	55
# documents	16M		260M	
# assessed documents	40855	73073	71279	57985
# assessed non-relevant documents	38124	66893	62268	47340
# assessed relevant documents	2731	6180	9011	4753
Ratio $\frac{Relevant\ Docs}{Non-Relevant\ Docs}$	0.07	0.09	0.14	0.10
Avg. relevant documents per topic	58.45	106.54	150.18	79.22

extended overview of these collections.

Behaviour Simulations. We will study the behaviour of retrieval models in the context of microblog ad-hoc retrieval. To that end, we will explore their parameters with respect to query term frequencies within from 1 to 15 and document lengths of up to 30, since the average length for a tweet is $\tilde{1}5$ after stop words removal.

Evaluation measures. We pay attention to precision at different ranks, with a maximum cut-off point at rank 30. Future evidence is accepted only at the collection statistics level as agreed by TREC organisers disregarding any documents after the query issuing time when computing evaluation measures [1].

Baseline selection. Table 3.2 contains evaluation results for the state of the art retrieval models considered in this study, when applied to Twitter TREC collections from 2011 to 2014. The models considered in this evaluation are IDF (TF-IDF[2]), BM25, DFRee, Hiemstra's LM (HLM) and Dirichlet's LM (DLM) since it was the baseline for the Microblog Tracks in 2013 and 2014. Moreover, we adhere to the implementation and default settings found within the Terrier IR platform (Ounis et al., 2005). Finally, since DFRee and IDF are generally the best performing models we will use them as our baselines.

[1] https://github.com/lintool/twitter-tools/wiki/TREC-2013-Track-Guidelines

[2] $Where\ TF = 1$. Results worsen considerably if we do not set TF to a constant.

Table 3.2: Evaluation results for the state of the art models considered. (Bold denotes the best performing system)

(a) 2011 collection

	Precision				
	@5	***@10***	***@15***	***@20***	***@30***
BM25	0.54	0.48	0.45	0.41	0.38
DFRee	0.61	**0.58**	**0.54**	**0.50**	0.45
DLM	0.50	0.47	0.45	0.42	0.37
HLM	0.54	0.48	0.45	0.42	0.38
IDF	**0.63**	0.56	0.52	0.49	**0.46**

(b) 2012 Collection

	Precision				
	@5	***@10***	***@15***	***@20***	***@30***
BM25	0.40	0.37	0.34	0.34	0.31
DFRee	**0.46**	**0.45**	**0.42**	**0.39**	**0.36**
DLM	0.34	0.33	0.32	0.29	0.27
HLM	0.38	0.37	0.35	0.33	0.31
IDF	0.44	0.39	0.36	0.36	0.34

(c) 2013 collection

	Precision				
	@5	***@10***	***@15***	***@20***	***@30***
BM25	0.58	0.51	0.46	0.42	0.38
DFRee	**0.67**	0.60	0.55	0.51	**0.45**
DLM	0.27	0.28	0.26	0.26	0.24
HLM	0.44	0.38	0.35	0.33	0.31
IDF	0.66	**0.62**	**0.56**	**0.52**	**0.45**

(d) 2014 collection

	Precision				
	@5	***@10***	***@15***	***@20***	***@30***
BM25	0.69	0.62	0.58	0.57	0.52
DFRee	0.73	0.68	0.65	0.63	0.60
DLM	0.35	0.35	0.34	0.34	0.33
HLM	0.55	0.49	0.46	0.44	0.41
IDF	**0.75**	**0.73**	**0.69**	**0.67**	**0.62**

(e) All collections

	Precision				
	@5	***@10***	***@15***	***@20***	***@30***
BM25	0.55	0.49	0.46	0.43	0.39
DFRee	**0.62**	**0.57**	**0.54**	**0.51**	**0.46**
DLM	0.36	0.35	0.34	0.32	0.30
HLM	0.47	0.43	0.40	0.38	0.35
IDF	**0.62**	**0.57**	0.53	**0.51**	**0.46**

3.4 Investigating Retrieval Model Problems

The literature has identified **document length normalization** as the main culprit for the under-performance of retrieval efforts in microblogs. The work by Naveed et al. (2011) suggests that the **Verbosity** and **Scope** hypotheses do not hold for microblog retrieval.

The **verbosity** hypothesis supports that some authors are more verbose than others, thus applying length normalization by dividing by the length of the document is beneficial to better capture relevance, as repetition of terms is superfluous. On the other hand, the **scope** hypotheses states that some authors simply have more to say, thus naturally adding more relevant information to the topic. As a result documents are longer but more extensive and rigorous in their content than shorter ones. The added value should be accounted for and thus the documents should be promoted over shorter ones.

In the context of Microblog retrieval, Naveed et al. (2011) carried out a number of experiments using a logistic regression model over a number of tweet features as the retrieval methodology. They showed significant improvements in performance when their algorithm did not perform document length normalization over its normalised counterpart. However, since their ranking approach takes into consideration multiple other features, it is not clear if their finding about document length normalization is generalisable.

Furthermore, although it is been often assumed, it is not known if length normalisation is bad altogether for microblog retrieval, or maybe it is just how it is currently interpreted in this particular case, what makes it harmful.

Intuition tells us that document length normalization might not interact well with the limitations imposed in microblog documents. The **Verbosity** and **Scope** hypotheses seem not to properly model the behaviour of users publishing microblogs as they are generally challenged to fit their messages within the strict character limit. Consequently, retrieval models designed under scope and verbosity or similar premises - such as BM25 (Robertson and Zaragoza, 2009) - are likely to exhibit unexpected behaviour.

The first step into developing our understanding of the behaviour of retrieval models is to study the elements that compose them. To this end we have compiled Table 3.3 which summarises the different components involved in the score computation of a variety of state of the art retrieval models. The top row of the table indicates whether the component relies on collection statistics (I.e. Collection feature) or the document statistics (I.e. Document feature). The second row contains acronyms for each of the features, which are expanded as:

ND. **Number Of Documents:** Total number of documents in the collection.

DF. **Document Frequency:** Number of documents in which the term appears (I.e. A term's posting list size).

ADL. **Average Document Length:** This is the average document length in number of tokens, for all documents in the collection.

NT. **Number Of Tokens:** Number of different tokens in the collection.

Table 3.3: Features involved in the computation of retrieval models.

	Collection Features					Document Features	
	ND	***DF***	***ADL***	***NT***	***CTF***	***TF***	***DL***
IDF	*	*					
DFRee				*	*	*	*
BM25	*	*	*			*	*
HLM				*	*	*	*
DLM				*	*	*	*

CTF. **Collection Term Frequency:** Frequency of a term in the whole collection. (I.e. Total number of occurences of a term in the collection)

TF. **Term Frequency:** Frequency of the query term in the document being evaluated.

DL. **Document Length:** This is the document length, in number of tokens, for the document being scored.

Each of the remaining rows contain the name of the retrieval model as well as whether a component involved in its computation (Denoted by *). For example, DFRee uses Number Of Tokens (NT), Collection Term Frequency (CTF), Term Frequency (TF) and Document Length (DL).

In the following sections we investigate the behaviour of the abovementioned retrieval models in terms of these features. We perform our analysis mainly by means of simulating their behaviour with a range of different values common under microblog retrieval conditions. We then contextualise the model's actual performance with respect to its simulated behaviour, and draw generalised conclusions across these experiments.

3.4.1 The BM25 Case

The work by Ferguson et al. (2012) examined the performance of BM25 when used under a microblog retrieval scenario. Their findings showed how the closer to zero the free parameters were set in BM25, the better the performance achieved. However, they did not connect this finding to the design of BM25 and what these settings meant in terms of the affected components. In this section we exemplify and connect these findings to the theory by simulating the behaviour of BM25 under microblog retrieval conditions.

First, we observe in Table 3.3 how BM25 relies on document length by using both ADL and DL components in its computation. Furthermore, BM25 has two free parameters, namely b and k_1, which control the effects of the "saturation function" over the final score. The saturation function in BM25 encodes the document length evidence as part of the score as follows:

The first version of the saturation function is given by:

$$\text{Version 1: } \frac{f(q_i, D)}{f(q_i, D) + k_1} \text{ for some k\_1} > 0 \tag{3.1}$$

Once we take into consideration the Verbosity and Scope hypotheses, we derive the following saturation function:

$$\text{Version 2: } \frac{f(q_i, D)}{f(q_i, D) + k_1 * ((1 - b) + b * dl/avdl)} \text{ for some k\_1} > 0 \tag{3.2}$$

The main difference between these equations is that **Version 2** reduces the effect of term frequency with respect to the document length and its collection average, whilst **Version 1** only relies on the k_1 free parameter. Secondly, the free parameter b ponders between the Verbosity and Scope hypotheses. Setting b to 0 effectively disables the Verbose hypothesis, giving full weight to Scope, in other words, the longer the document the better. Thus when b is set to 0, *Version 2* of the saturation function becomes *Version 1*.

As we introduced before, the study carried by Ferguson et al. (2012) explored the best parameters for b and k_1 concluding that best performance is achieved as both parameters tend to 0. However, the authors did not mention that by setting those parameters close to 0, we are disregarding the document length normalisation component altogether. Thus for all intents and purposes BM25 becomes IDF. This can be proved mathematically by substituting b and k_1 by 0 as follows.

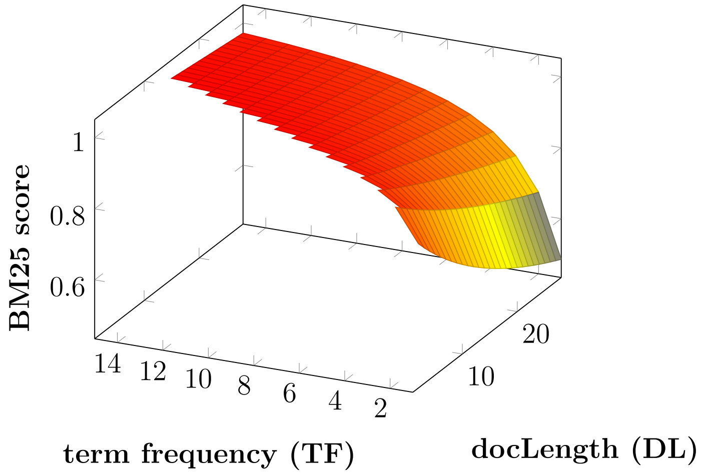


Figure 3.1: Term Frequency (TF) vs, Doc. Length (DL)

$$
\begin{aligned}
\text{BM25}(D,Q) &= \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{\text{avgdl}})} \\
&= \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(0+1)}{f(q_i,D)+0\cdot(1-0+0\cdot\frac{|D|}{\text{avgdl}})} \\
&= \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)}{f(q_i,D)} \\
&= \sum_{i=1}^{n} \text{IDF}(q_i) \qquad (3.3)
\end{aligned}
$$

Initially it would seem that the **Scope** and **Verbosity** hypotheses do not hold for microblogs. The reasoning behind being that these hypotheses were developed for documents that were unbounded in terms of their length such as web pages or books. However, since document length has an upper bound in microblogs, authors express their ideas in a very constrained space where verbosity and scope hypotheses do not seem to hold. However we will later observe that this conclusion is partially true[1].

Furthermore, terms in microblog documents have very low document frequencies. In fact, more often than not, query terms appear at most once in each document unless dealing with spam. Thus a query term appearing more than once within a document can have a dramatic effect over the score produced by BM25. In other words, the very

[1] We later demonstrate that an interpretation of the **scope** hypothesis does hold whilst **verbosity** does not

low document frequencies result in unreliable estimations of the informativeness of a query term. Consequently, in this particular case, it is better to rely on features outside the document such as collection features.

Finally, Figure 3.1 shows the possible BM25 scores for a range of Term Frequency (TF) and Doc. Length (DL) values.[1]. We can extract two interesting behaviours which we can compare later to other retrieval models. Firstly the increase of document length is regarded as negative. In other words the more information in number of terms is encoded in the document the less relevant it is regarded. Secondly the increasing term frequency results in increased scores. This would seem counter-intuitive in a document with such a limited length, as users normally struggle to fit their messages. Additionally, there is a danger of promoting spam messages which may only contain the query terms.

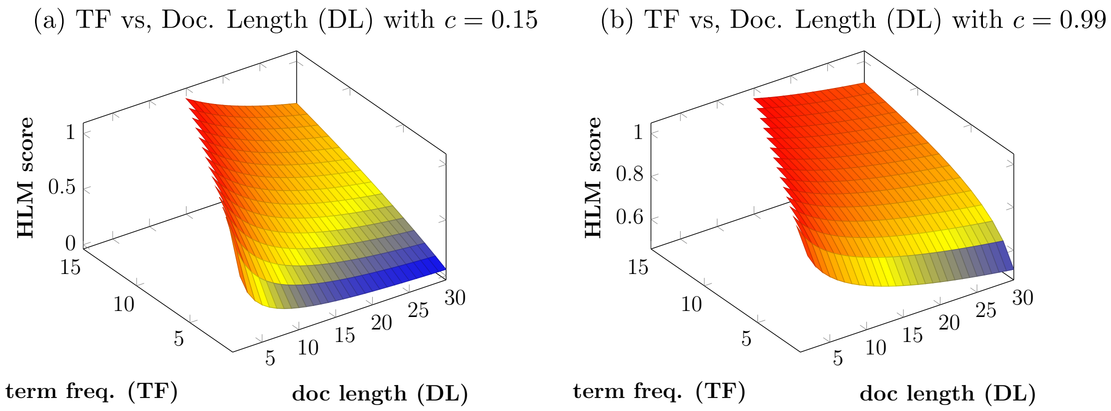
3.4.2 The Hiemstra's Language Model (HLM) Case

In this section we study the Hiemstra's Language Model (HLM) under Microblog conditions. Table 3.3 shows that HLM utilises both CollectionTermFrequency (CTF) and TermFrequency (TF) together with the total number of different tokens in the collection (NT) and document length (DL). Furthermore, if we pay attention to Table 3.2 we can observe that whilst DFR and HLM utilize the same components, HLM exhibits a more erratic performance under microblog conditions. HLM's performance for the 2013 collection is considerably lower than that of DFR or IDF, whereas it remains close to the top performing models for the 2011, 2012 and 2014 collections. HLM is formulated as follows:

$$\mathrm{HLM}(D,Q) = \sum_{i=1}^{n} \log_2 \left[1 + \frac{c \cdot f(q_i, D) \cdot ntoks}{(1-c) \cdot f(q_i, C) \cdot |D|} \right] \tag{3.4}$$

where $ntoks$ refers to the number of unique tokens in the collection (NT), c is a free parameter, and C represents the set of all documents in the collection. $f(q_i, D)$ represents the TF of a query term q_i in document D, whereas $f(q_i, C)$ is CTF of term q_i. The free parameter c regulates how HLM satisfies the conditions of **coordination level ranking (CLR)**) (Hiemstra and De Vries, 2000). CLR is a rule enforced in the design of HLM which ensures that documents containing n query terms are ranked higher than those with $n-1$ terms.

[1] Where $ND = 100k$ and $DF = 100$



(a) TF vs, Doc. Length (DL) with $c = 0.15$

(b) TF vs, Doc. Length (DL) with $c = 0.99$

Figure 3.2: HLM analysis w.r.t. term frequency (TF) and document length (DL)

Similarly to BM25, the assumption where higher term frequencies should be regarded positively, can easily result in the promotion of spam and undesired results. And this is rooted in the fact that query terms occur normally 1-2 times in a microblog document, due to length limitations.

Figure 3.2a shows a plot of the possible scores produced by HLM in its default configuration ($c = 0.15$)[1]. We can observe that for documents where the length is lower than 5 the differences between the scores are very marked. Above length 5 the progression of scores is much more subtle. In other words, shorter documents are subject to high differences between their scores due to small changes in their limited length.

Furthermore, we can observe in Formula 3.4, how the high sensitivity to low document length is a result of the model's design, since document length acts as a multiplier in the denominator. Additionally, term frequency can be found within the nominator as a multiplying component. Consequently, when higher than 1 it will result in an unreasonable boost of the score. In the case of microblog documents this can be problematic due to the scarce frequencies which average around 1.17 (±0.48)[2].

Table 3.2 shows that HLM is the second worst model overall for microblog retrieval. We hypothesise that the reason for this under-performance lies in the substantial scoring differences above-mentioned, resulting from the specific morphology of microblog documents which HLM does not account for. Thus reducing the differences in the scoring, should yield improved retrieval performance.

[1] Where $ND = 100k$, $DF = 100$ and $NT = 1000$

[2] Computed for query terms in all TREC microblog topics up to 2014 and our baseline DFR

Table 3.4: P@30 scores for HLM as we consider different combinations of dTF and dDL, and c (All collections together)

c	***dTF***	***dDL***	***P@30***
0.15			0.3475
0.15	20		0.3486
0.15		20	**0.3839**
0.15	20	20	**0.4462**
0.05			**0.2824**
0.40			**0.4009**
0.70			**0.4281**
0.99			**0.4492**
0.99	20	20	**0.4532**

3.4.2.1 Offsetting experiment

In order to test this hypotheses we simulate the behaviour of longer documents with higher term frequency by offsetting the values of TF and DL. We do this by a simple addition $TF = TF + dTF$, in this case dTF being the pondering value to offset TF. Likewise, we utilise $DL = DL + dDL$ where dDL is the variable to offset DL.

Table 3.4 shows the performance of HLM measured by Precision@30 with different configurations. The first row shows the performance of HLM with a default configuration of $c = 0.15$.

The second row with $dTF = 20$ so that $TF = TF + 20$ which denotes the offsetting of TF by +20. As stated before, the reason behind this offsetting is to reduce the differences between possible scores with respect to the actual values of TF. As we can observe offsetting just TF does no result in any significant improvement. Similarly, the third row shows the performance of HLM when offsetting DL by +20 in order to reduce the possible score differences. Consequently the results are much better than before with a Precision@30 increase of +11.76%. Finally, we experiment with the offsetting of TF and DL together to achieve yet another +15.79% Precision@30 increase over the previous combination and a very substantial increase of +29.41% over the baseline (no offsets) configuration).

It is interesting to notice how only the increase of TF does not help in retrieval, however only increasing DL does produce better results. Yet more importantly, by incrementing both TF and DL we obtain the best performance over all previous configurations. These results hint to a very subtle relationship between DL and TF values of microblog documents. Rows 5 to 8 in Table 3.4 show the performance of HLM with different values of c. As c is increased performance increases as well, reaching comparable performance to the approach which offsets DL and TF.

Finally, we compare Figures 3.2a and 3.2b which show scores produced by HLM w.r.t. TF and DL with different values of c. Figure 3.2a sets $c = 0.15$ whereas Figure 3.2b sets $c = 0.99$. It is easily observed how Figure 3.2a shows more differences across the spectrum of scores with respect to TF and DL than Figure 3.2b. We can also observe how offsetting DL and TF forces the possible values of HLM to lie in the more stable area of the Figures. Furthermore, Figure 3.2b produces the most stable scores.

From these experiments we can conclude that retrieval models require a conservative and delicate relationship with DL and TF, taking especial care to reduce the differences across the spectrum of possible scores, in order to reduce any unfair weighting differences due to scarcity in DL and TF.

3.4.3 The DLM Case

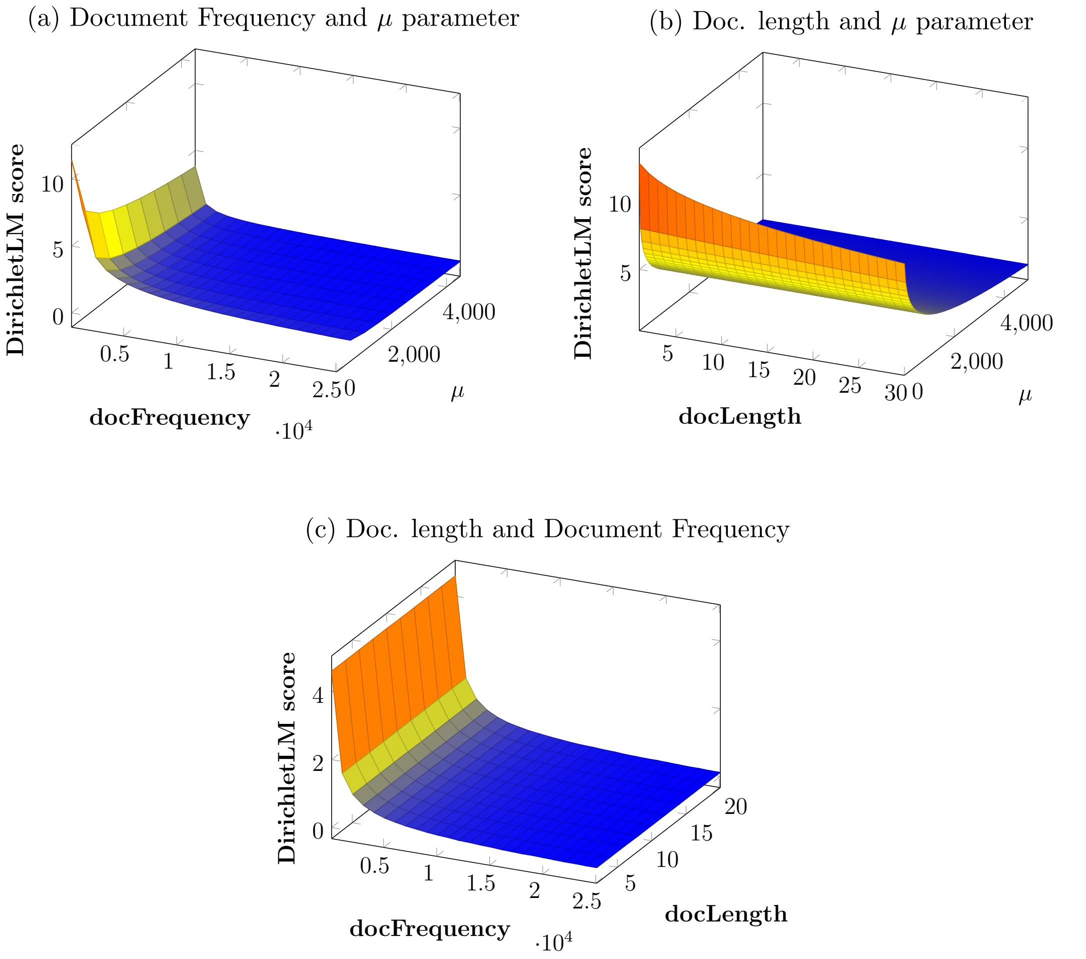
Dirichlet Smoothed language model (DLM), was the baseline retrieval model for the 2013 and 2014 instances of the microblog track. DLM was used within the "Microblog track as a service" client which managed a Lucene index in its core. DLM has a smoothing parameter named μ, which was set to 2500 by default during the 2013 and 2014 microblog tracks. Moreover, DLM scores are produced [1] by the following equation:

$$\mathrm{DLM}(D, Q) = \sum_{i=1}^{n} \log_2 \left[1 + \frac{f(q_i, D)}{\mu \cdot \frac{f(q_i, C)}{ntoks}} \right] + \log_2 \left[\frac{\mu}{|D| + \mu} \right] \tag{3.5}$$

where $ntoks$ refers to the number of unique tokens in the collection (NT), μ is a free parameter, and C represents the set of all documents in the collection. $f(q_i, D)$ represents the TF of a query term q_i in document D, whereas $f(q_i, C)$ is the collection document frequency (CTF) of term q_i.

Figures 3.3a and 3.3b show DLM scores in terms of the μ parameter, w.r.t. document frequency and document length respectively. Figure 3.3c on the other hand demonstrates the relation between document frequency and document length.

[1] As implemented in the Terrier IR platform


(a) Document Frequency and μ parameter

(b) Doc. length and μ parameter

(c) Doc. length and Document Frequency

Figure 3.3: DLM simulation figures

As we can observe from Equation 3.5 the parameter μ is closely related to the collection statistics, and the length normalization component of the equation. Moreover the lower the values of μ the higher the score differences for similar document frequencies as shown in Figure 3.3a. Similarly, we can observe in Figure 3.3b how μ interacts with document length. For low values of μ we can observe how the scores are reduced at the same time that documents become larger, as expected for normal documents. Interestingly, this behaviour is dampened with higher values of μ, as score differences are heavily reduced w.r.t. the different document lengths. Since the default value for μ is 2500, it is no surprise that document length has virtually no effect over the scores for DLM as seen in Figure 3.3c, contrary to other retrieval models.

This could be a desired feature for microblog retrieval, however let us look at the

Table 3.5: P@30 scores for DLM for a range of μ values (All collections together)

μ	***P@30***
1	0.4028
5	0.4164
20	0.4241
50	0.4099
100	0.3933
500	0.3396
1000	0.3227
2500	0.2988

performance achieved for a range of μ values in Table 3.5. As we can observe generally the higher the value of mu the worse the performance obtained, with the exception of μ within the 1 to 20 range.

In order to further understand the behaviour of DLM in the case of Microblog retrieval, we perform an analogous experiment to the previously performed for HLM. Since DLM was also designed for longer documents than microblogs, offsetting the statistics of TF and DL can be interesting experiment as it would better resemble its standard behaviour in term of the numerical values produced as scores.

The results of the evaluation are presented in Table 3.6. The first four lines contain the P@30 values for different combinations where μ is set to 20. As we can observe offsetting TF by +20 results in a substantial +7.47% increase of P@30 with respect to the default configuration. On the other hand offsetting DL by +20 results in a 8.02% decrease of performance in terms of P@30. Finally, combining the offsetting of both TF and DL results in comparable performance than that obtained by only increasing TF.

The same behaviour is obtained across all combinations when we set the $\mu = 2500$. To further develop our understanding of the behaviour, and to draw conclusions for such results, we devised Figures 3.4a and 3.4b. Figures 3.4a and 3.4b present the DLM scores produced with respect to Doc. Length (DL) and Term Frequency (TF) when $\mu = 2500$ and $\mu = 20$ respectively.

Let us analyse the results from Table 3.6 in connection with Figures 3.4a and 3.4b. As we can observe incrementing DL will result in an increased differentiation of DLM scores with respect to TF as more values are closer to the minimum and

Table 3.6: P@30 scores for DLM as we consider different combinations of dTF and dDL, and μ, (All collections together)

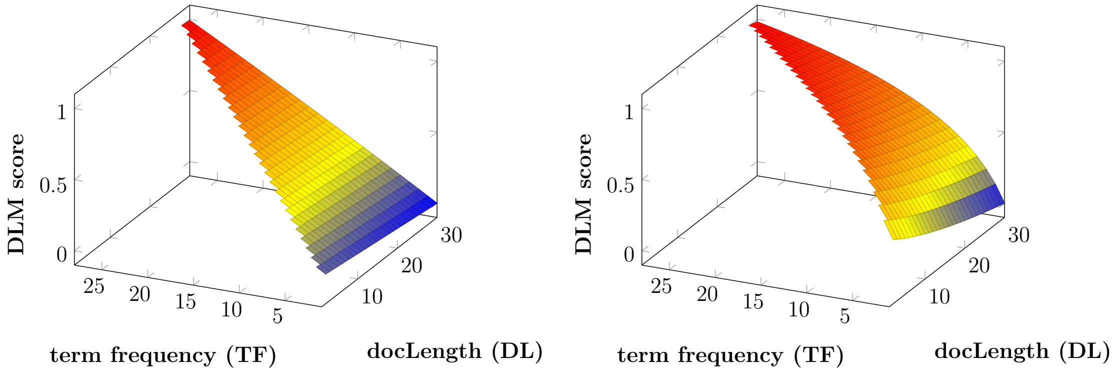
μ	***dTF***	***dDL***	***P@30***
20			0.4241
20	20		0.4558
20		20	0.3901
20	20	20	0.4547
2500			0.2988
2500	20		0.4468
2500		20	0.2892
2500	20	20	0.4466

maximum values. In other words there are less intermediate values (Light coloured areas), which ultimately reflects on heightened sensitivity to differences across the TF spectrum. Furthermore, we can also observe in Table 3.6 how incrementing DL values, results in worse performance in all cases. Consequently the increased differentiation of DLM scores with respect to the TF parameter, produced by the increment of DL is detrimental and in line with the findings in the previous section.

Additionally, Figure 3.4a shows an almost linear progression of DLM scores with respect to TF, whereas Figure 3.4b ($\mu = 20$) exhibits a logarithmic behaviour with respect to TF. The latter behaviour is more desirable because there should be a saturation point when incrementing TF at which there is very little value added to the score of the document, or could be even counter productive. In fact, if we take into consideration that term frequencies within microblogs are in the range 1-2, the pivoting value w.r.t TF should be very low, to avoid promoting spam microblogs.

The better behaviour with respect to TF is rewarded with increased performance whether the value of μ is 20 or 2500. In fact the offsetting of TF seems to overrule the effects of μ as similar results are obtained in both $\mu = 20$ and $\mu = 2500$ conditions. The effects of offsetting TF are most visually evident when looking at Figure 3.4b as differences amongst the different scores become very small, when $TF > 20$.

Extending on the findings by Naveed et al. (2011) who showed how length normalization was detrimental to microblog retrieval in an L2R retrieval framework. Our experiments have so far indicated the existence of a particular relationship between TF and DL that is most appropriate for Microblog retrieval. We believe that the score progressions with respect to *DL should modelled by a very gentle slope*, whereas there



(a) Doc. length (DL) and Term Frequency (TF) when $\mu = 2500$

(b) Doc. length (DL) and Term Frequency (TF) when $\mu = 20$

Figure 3.4: Evaluating DLM's behaviour

should be a pivoting point with respect to *TF where scores should decay* in order to account for spam. In the following sections these ideas will be further elaborated.

3.4.4 The DFRee Case

DFRee[1] is a Divergence From Randomness model implemented in the Terrier IR platform (Ounis et al., 2006). DFRee has been designed as a parameter-free model and adheres to the following implementation:

$$prior = \frac{f(q_i, D)}{|D|}, posterior = \frac{f(q_i, D) + 1}{|D| + 1} \tag{3.6}$$

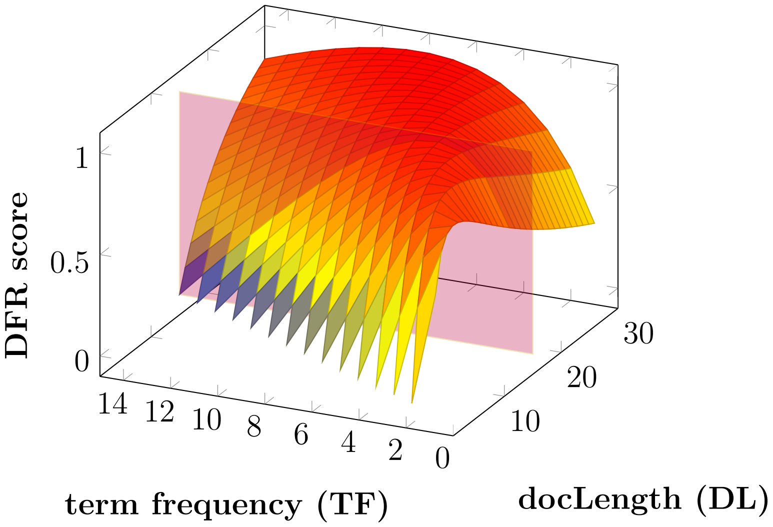
$$InvPriorColl = \frac{ntoks}{f(q_i, C)}, norm = f(q_i, D) * log_2 \frac{posterior}{prior} \tag{3.7}$$

$$\begin{aligned} DFRee(q_i, D, C) = norm * [& \\ & f(q_i, D) * (-log_2(prior * InvPriorColl)) \\ & + (f(q_i, D) + 1) * log_2(posterior * InvPriorColl) \\ & + 0.5 * log_2(posterior/prior)], \end{aligned} \tag{3.8}$$

where $f(q_i, D)$ represents the frequency of query term q_i within document D. Similarly $f(q_i, C)$ holds the collection C frequency for query term q_i. Furthermore *ntoks* is

[1] http://terrier.org/docs/v2.2.1/javadoc/uk/ac/gla/terrier/matching/models/DFRee.html

Figure 3.5: Evaluating DFR's behaviour: Doc. length (DL) and Term Frequency (TF)



the total number of unique terms within collection C and $|D|$ represents the document length of document D.

Similarly to the evaluations carried out in previous sections, we simulated the scores produced by DFRee given a range of TF and DL values. The objective is studying its behaviour in microbloging conditions, and draw conclusions about its performance. These simulated values are shown in Figure 3.5.

As we traverse the Document Length axis we can observe an interesting behaviour which is not present in any model observed so far.

For low values of TF, incrementing DL from 1 to ~ 16 results in also a higher score. This behaviour aligns with the scope hypotheses as longer documents are regarded as more informative. However, when DL reaches high enough values the scores start to decline. The latter behaviour is in line with the verbose hypotheses which assumes the extra length is due to superfluous information. Particularly when the extended document length is not accompanied by higher query term frequencies.

When dealing with documents as short as microblogs it is very difficult assert their informativeness or relevance in terms of the verbose or scope hypotheses. In fact all retrieval models observed so far follow these to some degree and perform worse than a simply using IDF as a retrieval model. Additionally, the premises in which they are built seem not to hold as they fail to perform better than simple IDF. However DFRee is an interesting exception as it performs better than all the studied retrieval models, and it performs better than IDF in some cases (Table 3.2).

We believe that the *saturation point* observed in Figure 3.5 in terms of TF and DL is responsible for DFRee outperforming other retrieval models in this task (And sometimes IDF). The score produced by DFRee can only be higher if both TF and DL increase. Thus, incrementing the value of a single component will increase the score to a saturation point after which the score will then decrease. As an example, consider an average microblog document of length 15 (blue plane in Figure 3.5). The score is maximised when TF approaches 3, after which higher TF values result in a significant reduction to the score.

This behaviour opposed to that of BM25, HLM and DLM which exhibit a positive correlation between TF and the score produced. Note that in this case a document made up of repeating query terms would be valued over others with richer, and more informative content. This behaviour is obviously problematic as it promotes spam-like documents. Fortunately DFRee has a pivoting point which attempts to alleviate this possibility, thus reducing the value of increasing TF in short documents.

Recall that users of microblog services such as Twitter, strive to fit their messages within the character limit. It stands to reason, that the more terms they fit within the character limit the higher the chances of it being informative. The pivoted behaviour of DFRee does not completely match this premise, however it does match it better than all other observed retrieval models (Including BM25, HLM and DLM) where longer documents are simply less relevant under microblog conditions.

Summarising, we believe that DFRee's behaviour is key to better understand why most retrieval models fail to capture the relevance of microblogs. Particularly important is the *saturation point* behaviour as a function of TF and DL. We can observe that promoting documents that are longer, whilst penalising documents with higher TF values than 2 may be a better fit to capture microblogs' relevance.

3.4.5 Harmonising Score differences

So far we have introduced a set of representative retrieval models, and discussed how they behave when facing microblog-like conditions. We have mainly simulated the spectrum of scores produced w.r.t. TF and DL by each model when fixing all other parameters. Moreover we have observed that retrieval models performance seems to

Table 3.7: Behaviour when harmonising score differences.(All collections together.)

Model	***configuration***	***stdev***	***P@30***
DLM	$c = 2500$	0.2639	0.2988
DLM	$c = 50$	0.2479	0.4099
DLM	$c = 20$	0.2384	0.4241
HLM	$c = 0.15$	0.2553	0.3475
HLM	$c = 0.40$	0.2365	0.4009
HLM	$c = 0.99$	0.1135	0.4492
BM25	$b = 0.75, k = 1.2$	0.1274	0.3948
BM25	$b = 0.75, k = 0.7$	0.0927	0.4399
BM25	$b = 0.9, k = 0.1$	0.0181	0.4580
PEARSON	-0.70		
KTau	-0.66		

increase when we overestimate the values of TF and DL, thus forcing the models to return values of lesser score differences.

Table 3.7 holds a summary of the results for all retrieval models with various configurations with respect to Precision@30. Additionally the third column holds the standard deviation of the simulated scores produced by the retrieval models in microblog conditions[1].

As it can be easily observed, the possible document scores are much closer together for those configurations that improve a retrieval model's performance. In fact there is a strong statistical correlation (last two columns) between reducing the standard deviation and improving the retrieval performance of the models. This observation motivates the following hypothesis:

The range of scores produced by retrieval models are unfairly different due to its behaviour w.r.t. the scarcity of TF and DL values in microblog conditions.

If this hypothesis is true, we should be able to achieve similar positive results if we reduce the scoring differences of a retrieval model by means of any other technique. To this end we decided to apply a base two logarithm, to the scoring function of each retrieval model. As an example, the formulation of HLM would be as follows:

[1] where $DL <= 30$ and $TF <= 15$

Table 3.8: Retrieval models performance with log-smoothed scores (All collections)

	Precision @ 30		
	Default	$log_2(Ret.Model)$	% difference
DLM	0.2988	0.3977	+33.10%
HLM	0.3475	0.4489	+29.18%
$BM25$	0.3948	0.4336	+9.83%
$DFRee$	0.4614	0.4531	-1.80%
IDF	0.4626	0.4626	0%

$$\mathrm{HLM}(D,Q)=\sum_{i=1}^{n}\mathbf{log_2}\left[\log_2\left[\mathbf{1}+\frac{\mathbf{c}\cdot\mathbf{f}(\mathbf{q_i},\mathbf{D})\cdot\mathbf{ntoks}}{(\mathbf{1}-\mathbf{c})\cdot\mathbf{f}(\mathbf{q_i},\mathbf{C})\cdot|\mathbf{D}|}\right]\right] \tag{3.9}$$

where the added logarithm function can be found next to the summation sign.

Table 3.8 holds a comparison between the default P@30 achieved by each model and the same model with the log function applied to it. As we can observe the results for DLM, HLM and BM25 perform considerably better than their standard, whereas DFRee performs marginally worse and IDF remains unaffected.

From these experiments we can conclude that state of the art retrieval models produce unfair scores due to the scarcity of TF and DL during microblog retrieval. This effect can be mitigated by employing techniques to reduce possible score differences such as applying a log function. To conclude, when ranking microblog documents our models should consider the existing TF and DL evidence, but should also be conservative when managing the overall effects on the produced scores.

3.5 MBRM: A MicroBlog Retrieval Model

In the previous section, we discussed a number of problems faced by state of the art retrieval models when dealing with microblogs. We presented scarcity of TF and DL as a source of high scoring differences amongst the spectrum of possible scores for a retrieval model. Additionally we started defining the requirements for a retrieval model to effectively handle microblog documents by better capturing their informativeness. These requirements can be summarised as:

1. Higher DL should be regarded positively as authors of microblogs strive to fit as much content as possible within the character limits

2. Higher TF should be regarded negatively as high TF could be a result of spam messages, and normally TF revolves around 1-2

3. Score differences with respect to DL and TF should produce gentle slopes, to not penalise/promote unfairly documents with very little differences.

Following these premises, we have designed a “MicroBlogs Retrieval Model”, namely MBRM. MBRM is composed of two parts to deal with document based evidence. Then we attach the aforementioned part to an IDF component which represents the collection’s information. Similarly to the formulation of BM25, the two main components of MBRM deal with document length and query term frequency. The first component deals with the document length and is given by the following logistic distribution:

$$DLComp(DL) = \frac{c_1}{1 + a_1 \mathrm{e}^{-b_1 DL}} \tag{3.10}$$

where a_1, b_1 and c_1 are parameters to control the growth, maximum and starting point of the distribution. Secondly, the following component given by a gaussian distribution deals with the effect of TF over the final score produced by MBRM:

$$TFComp\,(TF) = a_2 e^{-\frac{(TF-b_2)^2}{2c_2^2}} \tag{3.11}$$

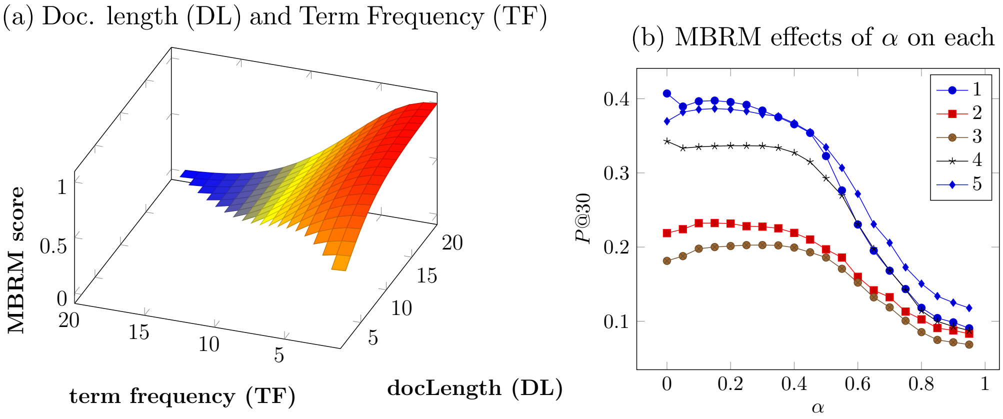
where a_2, b_2 and c_2 are similar parameters to those found in the previous function. These functions were chosen as they offer good control over the curves, and their values can be bound between 1 and 0 and we do not need to normalise them. The final formulation for MBRM is given by:

$$MBRM(D,Q) = \sum_{i=1}^{|Q|} (1-\alpha) * \mathrm{IDF}(q_i) + \alpha * DLComp(|D|) * TFComp(q_i) \tag{3.12}$$

which can be also expressed as:

$$MBRM(D,Q) = \sum_{i=1}^{|Q|} (1-\alpha) * \mathrm{IDF}(q_i) + \alpha * \left(\frac{c_1}{1 + a_1 \mathrm{e}^{-b_1 DL(|D|)}} \right) * \left(a_2 e^{-\frac{(TF(q_i)-b_2)^2}{2c_2^2}} \right) \tag{3.13}$$

Figure 3.6a shows a simulation of the behaviour of MBRM in terms of TF and DL. The parameters used to for both components (DLComp and TFComp) are shown in



(a) Doc. length (DL) and Term Frequency (TF)

(b) MBRM effects of α on each fold.

Figure 3.6: MBRM: A Microblog Retrieval Model

Table 3.9. These parameters where chosen to provide a saturation point in terms of the maximum score provided with respect to DL as DL approaches 15. Additionally, we reduce the score of query terms with frequencies higher than 1, to avoid spam behaviours. Consequently, we can observe in Figure 3.6a how the scores obtained with respect to the TF axis decrease slowly for the initial values of TF, but rapidly accelerate in their descent to then settle near 0. This behaviour is similar to that of DFRee - albeit smoother - as the highest importance is also given to low TF values ~ 2.

In terms of DL we produce a soft increasing slope to account for increasing value assigned to more informative documents. Unlike $DFRee$, the slope is always incremental. The idea behind it being that the more terms in the microblog the more comprehensive it should be, as more information is encoded regardless of the character limitation. In order to find the optimal value for the pondering value of α we divided the all the collections into 5 folds. For each of the folds we produced a P@30 result for a number

Table 3.9: MBRM recommended parameter settings

Parameter	**Recommended values**
a_1	1.5
b_1	0.3
c_1	1.0
a_2	1.0
b_2	2.0
c_2	6.0

Table 3.10: Performance of MBRM on all collections (Where * $p < 0.05$ and ** $p < 0.01$ respectively, with respect to IDF and DFRee)

	Precision				
	@5	***@10***	***@15***	***@20***	***@30***
DFRee	0.62	0.57	0.54	0.51	0.46
IDF	0.62	0.57	0.53	0.51	0.46
MBRM ($\alpha = 0.20$)	**0.64***	**0.59***	**0.56****	**0.53****	**0.48***

of α values in the 0-1 range. These can be found in Figure 3.6b. It can very easily be observed that the most optimal values for the mixing parameter α are near 0.20.

Finally Table 3.10 shows the evaluation results obtained for MBRM in terms of Precision at different levels in comparison with IDF and DFRee. As it can be observed, the performance is always significantly superior than the baselines. The main difference with respect to IDF is obviously that it takes advantage of document statistics, where IDF does not. However the main difference with respect to DFRee is that documents longer than 15 terms are not penalised following the aforementioned rationale.

These results not only demonstrate that we can make effective use of document statistics unlike previously thought by other authors (Naveed et al., 2011), but also that the scope hypotheses still holds for small documents such as microblogs. In other words, the authors of the documents will attempt to encode as much information as possible even with the obvious document limitations.

The verbose hypotheses however seems not to hold, as authors are simply capped by the character limitation with very little length variations. Thus documents are not generally longer due to style differences, or the verbosity of the author, but it is rather a reflection of the author's capacity to encode rich information in such limited constraints, which again aligns better with the scope hypotheses. And this is what we ultimately attempted to capture with our MBRM retrieval model.

3.6 Conclusions

In this chapter, we verified whether the scope and verbosity hypotheses still hold for microblog document retrieval. We initially hypothesise that the scope and verbosity hypothesis would not hold due to the character limit inherent to microblog documents. We derive this intuition from the assumption that authors of documents are able to produce documents of any length, which is behind the scope and verbosity hypotheses.

We then proceeded to analyse the behaviour of a number of state of the art retrieval models. The chosen models were BM25, HLM, DLM, DFRee and IDF. Our experimentation resulted in a better understanding of what are the shortcomings experienced by such models under microblog ad-hoc retrieval conditions. Particularly, we isolated the fact that longer documents should be promoted to account for the effort of microblog authors to encode their messages into the character limit. Then we identified that higher term frequencies than 1-2 should be penalised as they are more likely to be less informative and more reminiscent of spam documents. Based on these observations we concluded that the scope hypotheses does still hold in microblog documents, as generally longer documents are more informative, however verbosity does not due to the limitation in character length.

Finally we built a retrieval model optimised for microblog retrieval, namely MBRM, which takes intro account the observations extracted from the experimentation with aforementioned retrieval models. Our evaluation results demonstrate how MBRM significantly outperforms the best baselines (IDF and DFRee), by making better use of document-encoded evidence.

Future work will show how MBRM can be used to push further the current performance of approaches that rely on the initial results such as Automatic Query Expansion. We will also investigate which are the best parameters for MBRM in order to optimise its performance under microblog retrieval conditions.

Chapter 4

Microblog Dimensions and Informativeness

4.1 Introduction and background

In the previous chapter, we explored the performance of state of the art retrieval models in the context of microblog retrieval. We established that the *scope* hypotheses used in the design for BM25, and inspiration to many other retrieval models, does hold for microblog documents. However the *verbosity* hypothesis does not. As a contribution of our study we developed our understanding of what affects retrieval in microblog conditions and introduced a microblog specific retrieval model, namely, MBRM.

Microblog documents have more dimensions than normal documents. Aside from the textual message, microblogs contain tags such as mentions and hashtags as well as urls. These tags refer to recipients of a message (or users of interest), the topic at hand, and web links to related information respectively. In this Chapter we explore these intrinsic features of microblog documents, and attempt to further our understanding of what makes a microblog document relevant in terms of these features.

These features have been utilised before in a variety of ways. The workshop Making Sense Of Microposts (MSM) (Basave et al., 2013) presented participants with a challenge. The objective was to build systems able to identify and extract concepts from microblog documents, in a semi-supervised manner. The participant systems were to categorise concepts as belonging to the categories: person, organisation, location and miscellaneous. A similar task is that of microblog summarisation (Sharifi et al., 2010) in that tweets have to be processed and made sense of in order to produce a richer representation. Amongst the works submitted to this workshop, we can highlight the work by Tao et al. (2012). In their work they perform an in-depth analysis of both topic dependent and independent features for the MSM task. Some of the topic independent features consider the presence of hashtags, URLs and the length of the documents to be in connection with the relevance of documents. Whilst in our work we pay attention to the same features, we do so from a different angle. We study how many characters relative to the total characters in the document is dedicated to each of the microblog dimensions.

In the context of ad-hoc retrieval, the work by Massoudi et al. (2011) explores the use of these and other features to improve retrieval performance. These features include emoticons, hyperlinks, shouting, capitalization, retweets and followers. Work by Nagmoti et al. (2010) extended the study concerning the use of social features such as the number of followers and followees to further the performance gains in ad-hoc retrieval.

While all these works utilise microblog features to produce better results in their particular tasks, they do not properly attempt to explain how these features relate to the relevance of microblog documents. In our work, we consider features based purely on microblog characteristics, explain their relationship with relevance, and finally utilise and combine those features to improve the behaviour and retrieval performance of a given state of the art retrieval model. The results of our experimentation lead to the conception of a "*microblog informativeness hypothesis*" drawing inspiration from the *scope* and *verbosity* hypotheses. We then tested our hypothesis by successfully enhancing the behaviour of our baseline retrieval model. Finally we also explore how the different dimensions from microblog documents interact with each other, by modelling their co-ocurrences in relevant and non-relevant microblog documents by means of state-machines.

Finally we produce a number of experiments to demonstrate how the ordering of elements within a microblog document can also be used as a source of relevance evidence within retrieval models. To this end, we encode the observed structure of relevant and non-relevant microblog documents into two different state machines. Hence a score is produced for any unseen document, by estimating how often similar structures can be found in the state machines.

4.2 Informativeness of Microblogs

In information retrieval, the relevance of a document is modelled by the combination of statistical measures extracted from both the collection and the documents themselves, which are embodied in retrieval models. Particularly, most retrieval models take into consideration document based statistics, such as document length and term frequency, in an attempt to estimate the relevance of documents. A very prominent example of the usage of document statistics, are the scope and verbosity hypotheses posed in the design of BM25.

Recall that the *verbosity* hypotheses states that some authors are naturally more verbose than others thus leading to longer documents, whereas the *scope* hypotheses regards longer documents as being more *informative* as a result of the extended contents. These hypotheses were implemented within BM25 as the following saturation function:

$$\frac{f(q_i, D)}{f(q_i, D) + k_1 * ((1 - b) + b * dl/avdl)} \text{ for some k\_1} > 0 \tag{4.1}$$

where q_i and D stand for a query term and a document where q_i appears respectively. dl and $avdl$ are the length of document D and the average length of all documents in the collection respectively. Finally k_1 and b are free parameters that control the influence of the verbose and scope hypotheses.

Microblog documents - such as tweets - have a fixed maximum size (140 characters in the case of tweets). Consequently, authors tend to optimise their wording in order to effectively convey their messages within the character limits and constraints set by the platform. Intuition tells us that retrieval models built around assumptions similar to the **scope** and **verbosity** hypotheses are very likely to exhibit an unexpected behaviour under microblog retrieval conditions, as we previously explored in Chapter 3.4. Fortunately, microblogs are highly dimensional documents which contain various types of information encoded within the same message, following an organically and community-agreed vocabulary.

In our work we draw inspiration from the exploratory process that led to the conception of the *scope* and *verbosity* hypotheses and ultimately to the successful BM25 retrieval model. To this end we describe a novel hypotheses tailored to microblog retrieval, namely **"Microblog Informativeness"** which highlights and relies on the intrinsic characteristics of such documents as follows:

> **The informativeness of microblog documents is tightly connected to the richness of content portrayed by the rate of usage of each of its dimensions.**

Firstly, for the purposes of our study, we generalise any retrieval model $P(Q|D)$ as a particular relationship noted by "$\boxed{?}$"[1] between document length $|D|$ and frequency of query term q_i in document D given by $P(q_i|D)$ which are used to produce the score of

[1] The question mark $\boxed{?}$ is intentional

a term $TS(q, D)$. We pay special attention to the $P(q_i|D)$ and $|D|$ components since they are the main difference between microblogs and longer documents. Thus, the relationship between these two features and collection statistics is included in the $\boxed{?}$ wild-card. Our generalisation can be formulated as follows:

$$\begin{aligned} P(Q|D) &= \sum_{i=0}^{|Q|} TS(q_i, D) \\ TS(q, D) &= |D| \boxed{?}\, P(q|D), \end{aligned} \tag{4.2}$$

Since the number of terms in microblog documents is largely constrained by the 140 character limitation of services such as Twitter, we decided to measure a microblog's relevance from a different point of view. We assume that microblog documents (**D**) are 4-dimensional entities comprised of **Text** $T(D)$**; URLs** $U(D)$ (Linking to an external resource); **Hashtags** $\#(D)$ (Terms preceded by #) indicating a topical context and **Mentions** $@(D)$ (Terms preceded by @) indicating an intended audience.

We believe that authors of informative microblogs will choose shorter synonyms of terms carefully in order to reduce the character count, and dedicate the character surplus in the text content to other dimensions. Consequently the amount of characters dedicated to each of the dimensions should have a relationship with the likelihood of a microblog to be more informative than others.

Therefore we define **Microblog Informativeness** (MI) as the probability $MI(Q|D)$ for a Microblog document D to fulfill an information need expressed as a query Q. Thus $MI(Q|D)$ is made up by an unobserved combination represented by "$\boxed{?}$" of the aforementioned dimensions, as follows:

$$MI(Q|D) = \sum_{i=0}^{|Q|} T(D) \boxed{?}\, U(D) \boxed{?}\, \#(D) \boxed{?}\, @(D) \boxed{?}\, TS(q_i, D) \tag{4.3}$$

where $T(D)$, $U(D)$, $\#(D)$ and $@(D)$ are the ratios given by the number of characters spent in the document for each of the dimensions considered[1]. For example, the ratio for the text dimension $T(D)$ is given by:

[1] URL's are automatically shortened by Twitter

$$T(D) = \frac{\#ofCharsforTextDimension}{Total\#ofChars}, \tag{4.4}$$

In order to test our hypotheses and learn what characteristics relate better to relevant microblog documents, we analyse retrieval runs produced by the state of the art baseline DFRee[1]. We use the documents from actual rankings generated with DFRee, instead of all documents in the relevance judgements, in order to analyse those documents that contain query terms and we can make a difference. In other words, including all the documents in the relevance judgements could produce decontextualiased results, as documents evaluated come from a very diverse set of retrieval techniques including query expansion, machine learning, etc, where documents are matched by features not included in the original query, thus it would not be the best context for our evaluation.

To this end, we take into consideration the TREC Microblog topics 1 to 110 to observe and draw conclusions from. Then we confirm our findings through an evaluation on the newer 111 to 170 topics from the 2013 TREC Microblog search task.

Tables 4.1(a...e) introduce the mean character ratios for each of the dimensions for all documents retrieved by DFRee at the cut-offs @10, @20, @30, @50 and @100 respectively. The star indicates statistically significant differences between relevant and non-relevant documents for that dimension. Additionally, the last row on each table, indicates the average document length in number of characters for both relevant and non-relevant documents.

As we can observe in Tables 4.1(a...e), the differences between relevant and non-relevant documents in terms of document length (DocLength) are not statistically significant in any case. However, we can observe how relevant documents tend to be shorter than non-relevant documents for cut-offs @10 and @20, whereas then they become longer than non-relevant documents for any cut-off after @20. We can conclude that it is difficult to rely on this feature to discern between relevant and non-relevant documents, as the differences contradict each other depending on the chosen cut-off point.

The **URLs** dimension in Table 4.1 is statistically significantly larger on relevant documents than in their non-relevant counterparts across all cut-off points. This is in

[1] We chose DFRee as it is the best performing model - together with IDF - as shown in Table 3.2

Table 4.1: Ratio of each dimension (Dim) for relevant (Rel) and non-relevant (Non-Rel) documents at different cutoffs for DFRee runs on the 2011 and 2012 Microblog collections. DocLen is given by the mean number of characters for all documents in the group.

(a) Cutoff @ 10

Dim	**Rel**	**Non-Rel**
Hash	1.96	1.619
Ment	2.75	2.444
Urls	17.32	14.16 *
Text	77.95	81.77 *
DocLength	97.47	100.2

(b) Cutoff @ 20

Dim	**Rel**	**Non-Rel**
Hash	2.626	1.861 *
Ment	2.453	2.402
Urls	17.54	13.54 *
Text	77.37	82.18 *
DocLength	96.50	97.38

(c) Cutoff @ 30

Dim	**Rel**	**Non-Rel**
Hash	2.514	1.999
Ment	3.061	2.671
Urls	17.13	14.28 *
Text	77.29	81.04 *
DocLength	96.21	95.76

(d) Cutoff @ 50

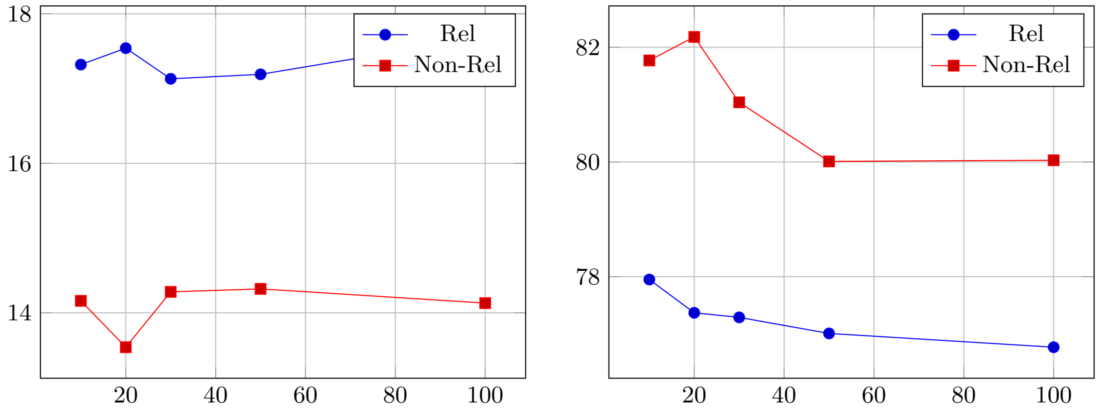
Dim	**Rel**	**Non-Rel**
Hash	2.820	2.518
Ment	2.968	3.136
Urls	17.19	14.32 *
Text	77.01	80.01 *
DocLength	95.90	94.45

(e) Cutoff @ 100

Dim	**Rel**	**Non-Rel**
Hash	2.638	2.514
Ment	2.893	3.315 *
Urls	17.69	14.13 *
Text	76.77	80.03 *
DocLength	93.96	92.56

line with previous works suggesting that the presence of URLs increases the likelihood for a document to be relevant (Massoudi et al., 2011). Additionally Figure 4.1a shows the changes in space dedicated to the URL dimension as we go down the result list. An interesting behaviour can be observed as relevant documents behave in exactly the opposite way to non-relevant documents. Traversing the different cut-off points show how the characters dedicated to URLs in relevant documents increases whereas, it decreases for non-relevant documents.

The **Text** dimension on the other hand, remains statistically significantly lower for relevant documents, than for non-relevant documents, across all cut-offs. However, as observed in Figure 4.1b, the behaviour as we traverse the list towards lower cut-off



(a) Urls (b) Text

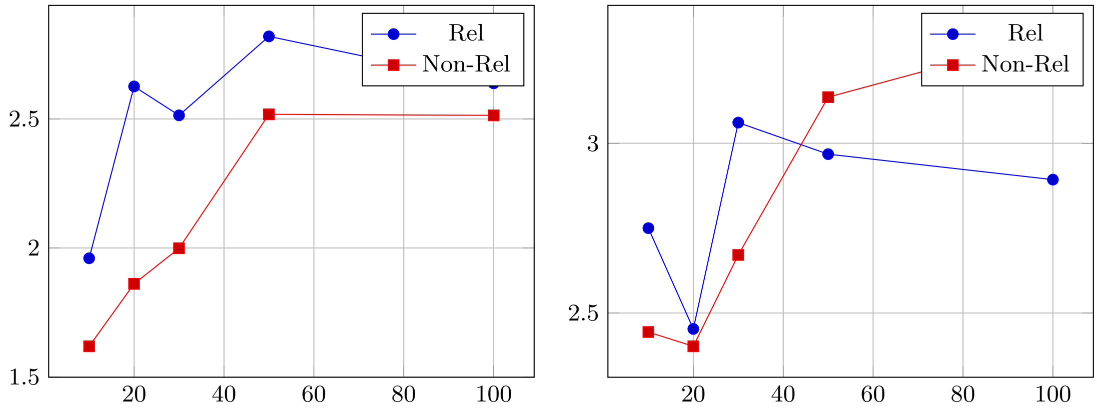
Figure 4.1: Rate (%) of characters dedicated to Urls and Text in Relevant and Non-Relevant documents at different cut-off points.

points is similar for both relevant and non-relevant documents. Thus the differences in characters dedicated to this dimension remain stable between relevant and non-relevant documents.

The stability of the differences observed for both the **URLs** and **Text** dimensions across all cut-off points make them a especially interesting set of features to be further studied and leveraged towards improving the behaviour of retrieval systems.

Figure 4.2 shows the behaviour for the **Hash** and **Mention** dimensions. In terms of the **Hash** dimension, differences are only significant when looking at the @20 cut-off. Then, as we traverse the result list, the presence of hashtags becomes more pronounced for both relevant and non-relevant documents. Additionally, Figure 4.2a shows how relevant documents dedicate a higher portion of the content to this dimension than non-relevant documents in average. This clear difference shows how hashtags are an interesting feature that could help in promoting relevant documents over non-relevant ones.

Finally, we observe the behaviour of the **Mention** dimension in Figure 4.2b. For the first three cut-offs @10; @20 and @30, relevant documents seem to spend more characters on the intended audience than non-relevant documents. After the @30 cut-off the roles are swapped and non-relevant documents spend more space in referring to the target users than relevant documents. Additionally, the differences in terms of the space dedicated to the **Mentions** dimension only becomes significant once we are at the



(a) Hashtags

(b) Mentions

Figure 4.2: Rate (%) of characters dedicated to HashTags and Mentions in Relevant and Non-Relevant documents at different cut-off points.

much lower cut-off point @100. This could reflect that many non-relevant documents may be conversational in nature, instead of introducing facts interesting to a wider audience. Additionally non-relevant documents could be spam messages including only mentions in the text as we approach higher cut-off points.

4.2.1 Modelling Microblog Informativeness

In the previous section we observed that relevant Microblog documents present different characteristics to those non-relevant in terms of the aforementioned dimensions (Figure 4.3). More specifically, relevant documents tend to use less characters for text, but more characters to contain the URLs and hashtag dimensions than non-relevant documents.

We cannot assume that the less space dedicated to text the more relevant the document will be, as that would make a text-less document the one with the highest likelihood of being relevant. Therefore, we estimate that a relevant document has an optimal amount of space dedicated to the text dimension which ranges from 76% to 78% as observed in Figure 4.1b. Thus we can model informativeness in terms of the term scoring function of a retrieval model $TS(q_i, D)$ for any given query Q, document D and its Text dimension $T(D)$ as:

$$MI(Q|D) = \sum_{i=0}^{|Q|} TS(q_i, D) + \lambda[\ 1 - |T(D) - 0.76|\ ], \qquad (4.5)$$

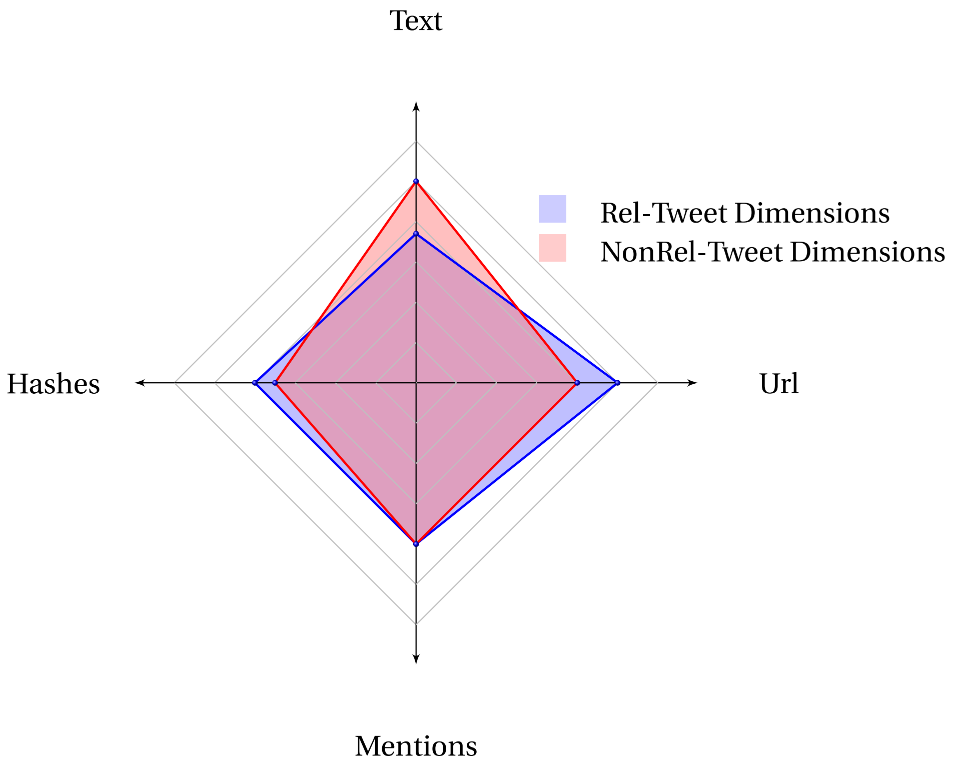


Figure 4.3: Dimensional differences between relevant and non-relevant documents. Statistically significant differences are exaggerated for easier visualization.

where a lower score is given to those documents diverging from the optimal text dimension rate 0.76[1]. We test this formulation using DFRee to produce the $TS(q_i, D)$ score over the microblog 2013 collection, which was **not** used in producing the analysis results in the previous section. Moreover the results are produced with the λ parameter set to 1.

The results are shown in the **text** row within Table 4.2. As we can observe, the performance of DFRee is enhanced by taking into account the textual dimension of the microblog documents, being statistically significantly better in terms of P@20. Similarly, we combine the rate of characters dedicated to the URL dimension with the score of the retrieval model as follows:

[1] The 76% rate for the text dimension specified above, which we normalise between 0 and 1.

$$MI(Q|D) = \sum_{i=0}^{|Q|} TS(q_i, D) + \omega U(D), \tag{4.6}$$

where we set the free parameter ω to 1. The results obtained for the experiments with this model are shown in Table 4.2 in row *url*. The use of the URL dimension on its own also improves the performance over the DFRee itself, most significantly for P@10 and P@20. Furthermore, it produces slightly better results than the Text approach. Additionally we combined both models to produce:

$$MI(Q|D) = \sum_{i=0}^{|Q|} TS(q_i, D) + \lambda[1 - |T(D) - 0.76|] + \omega U(D), \tag{4.7}$$

The results for this combination are shown in Table 4.2 as row *text-url*. Further improvements with respect to previous approaches are introduced at all cut-offs except P@10, where *url* performs slightly better than the combined approach. Finally we also added components to account for the hash and mention dimensions, producing the following two models:

$$\begin{aligned} MI(Q|D) = \sum_{i=0}^{|Q|} TS(q_i, D) + \lambda[1 - |T(D) - 0.76|] \\ + \omega U(D) + \gamma \#(D), \end{aligned} \tag{4.8}$$

$$\begin{aligned} MI(Q|D) = \sum_{i=0}^{|Q|} TS(q_i, D) + \lambda[1 - |T(D) - 0.76|] \\ + \omega U(D) + \gamma \#(D) + \delta @(D), \end{aligned} \tag{4.9}$$

where the free parameters are set to 1[1]. The results for both models 4.8 and 4.9 are shown in Table 4.2 as *text-url-hash* and *text-url-hash-ment* respectively. The performance achieved by adding the hash component over the previous models is further increased specially for P@10, whereas it performs slightly worse than *text-url* in terms of P@30. The addition of the mentions component in *text-url-hash-ment* reduces retrieval performance across P@10, P@15 and P@20 with respect to the last model.

If we consider Figures 4.1a, 4.1b, 4.2a and 4.2b and Table 4.2 we can see how the dimensions that showed constant differences across all cut-offs between relevant and non-relevant documents are the features enhancing the performance of the baseline. The only feature which results in poorer retrieval performance is the **mentions** dimension, which as observed in Figure 4.2b follows an erratic behaviour. For lower cut-off points more space is dedicated to the mentions in relevant documents, however it is the opposite case after the @40 cut-off point.

Table 4.2: Results when experimenting with the different dimensions over the 2013 TREC Microblog collection (*$p < 0.05$ over DFR).

Model	P@5	P@10	P@15	P@20	P@30
DFRee	0.65	0.59	0.54	0.51	0.45
text	0.65	0.59	0.54	0.52*	0.45
url	0.65	0.61*	0.54	0.52*	0.46
text-url	0.66*	0.61*	0.55*	0.52*	**0.47**
text-url-hash	**0.66***	**0.62***	**0.56***	**0.53***	0.46
text-url-hash-ment	0.66*	0.61*	0.55	0.52*	0.46

Based on our experimental results, we can assert that there are structural differences between relevant and non-relevant documents in terms of the dimensions defined in this work. More specifically, we have come up with a possible instantiation of our *Microblog Informativeness* hypotheses which leverage Microblog specific characteristics and is expressed by Equation 4.8. The implications of these findings and experiments are that users produce Microblog documents in different ways, with certain formats more likely to satisfy the information need of a prospective searcher. In the following Section, we expand our analysis by taking into consideration the order of the dimensions.

[1] Parameter optimisation could lead to substantial performance gains in future work, but it was not needed to answer the research questions set in this work

4.3 Dimensions Interaction.

To further our analysis in the structure of microblog documents we studied how the different dimensions interact with each other. Apart from the presence of the dimensions above discussed, we believe that the order in which they appear, and the interactions between them are also important. In fact, there are several documents on the web [1] which are meant to assist in writing the perfect tweet to grab the attention of readers.

In our study we utilised all documents in the relevance judgements from the Tweets 2013 collection as our training set. Each tweet is tokenised, and each token is categorised as representing each of the "text", "hashtag", "mention" and "url" dimensions, with the help of simple regular expressions matching. Moreover we quantify the frequency that a dimension is followed by another one. For example, we count the number of times when text leads to a hashtag, or a mention leads to a url. The frequencies of each dimensions leading to another dimension of the microblog documents are then utilised to build a simple state machine (or automata). Figure 4.4 shows an example, denoting how state 1, can transition to other states, such as state 2, with the probabilities stated above the arrows [2].

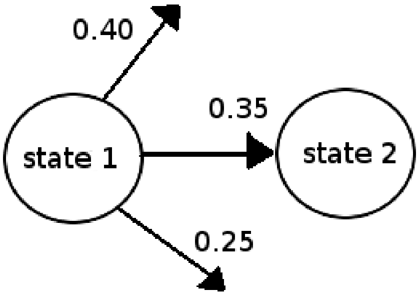


Figure 4.4: State machine example.

Figures 4.5a and 4.5b show state machines for both relevant and non relevant documents respectively. Both these figures contain a node to represent each of the dimensions studied in previous sections. Additionally they contain a "**start**" and "**end**" nodes, to denote the beginning and ending of the microblog document. Consequently, every existing tweet can be characterised by a particular path from the **start** to the **end**.

[1] http://blog.hubspot.com/marketing/tweet-formulas-to-get-you-started-on-twitter

[2] Notice that all transition probabilities for a node add up to 1.

While both figures look very similar, there are important differences that are worth noting. Firstly, looking at the transition from mentions to the end of the document, we can see that the probability for relevant documents is more than double (+21%) than that for non-relevant documents. This means that relevant documents are more likely to finish mention than non-relevant microblogs. Likewise the probability of ending a relevant document with a token of text is 12% less than for non-relevant documents. Moreover the chance of transitioning from a text token to a url token is 13% higher for relevant documents compared to non-relevant microblogs. Finally the chances to start a document with a mention is half (6% less) for relevant documents with respect to non-relevant ones.

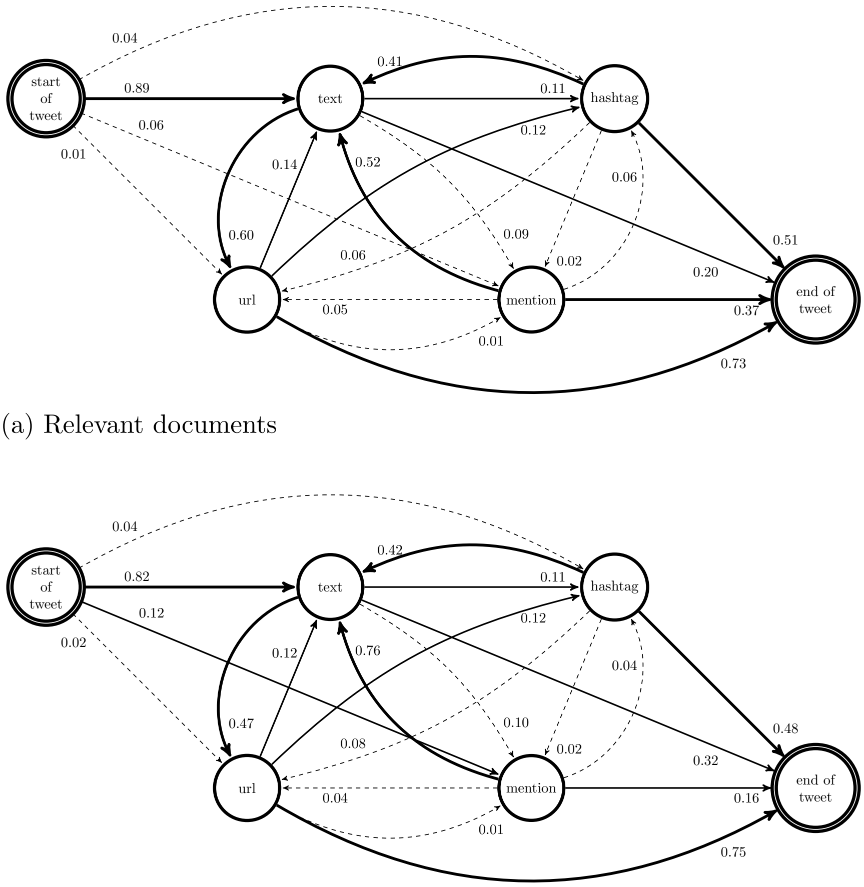
In order to test whether we can use this evidence for producing better rankings, we devised our **"State"** approach. The State approach is a re-ranking method that linearly combines the score given by any retrieval method with the aggregation of probabilities from start to end nodes w.r.t a microblog's structure.

As an example, consider the following tweet: *"Astronomers discover ancient system with five small planets. Details: http://go.nasa.gov/1wCpkJn @NASAKepler"*. Following the approach described above, we can infer the following structure: "$[start] - > [text] - > [url] - > [mention] - > [end]$". If we take the automata for relevant documents (Figure 4.5a) as the source of probabilities it would produce the score: $0.89 + 0.60 + 0.01 + 0.37 = 1.87$.

The "State" score therefore is given by the following equation:

$$\begin{aligned} State(D,Q) = (1-\alpha)P(Q|D) \\ +\alpha * (R\_Score(D) - NR\_Score(D)), \end{aligned} \tag{4.10}$$

where $R\_Score(D)$ and $NR\_Score(D)$ are the scores computed by traversing the automatas in Figures 4.5a and 4.5b respectively and α is a weighting factor which balances the linear combination with the score given by a retrieval model $P(Q|D)$. Notice the subtraction of the score given by the automata based on non-relevant documents with respect to the score based on relevant documents. The intuition is that we want documents that agree with the structure observed for relevant documents, whilst diverging from that of non-relevant documents.



(a) Relevant documents

(b) Non-Relevant documents

Figure 4.5: Tweet automatas for the 2013 collection

Table 4.3 shows the retrieval results for our re-ranking approach over the 2011 and 2012 collections. P@5 to P@30 represent Precision at the different cut-off points, whereas MAP denotes Mean Average Precision at cut-off 30. The first column contains the model being evaluated. Baseline represents a simple retrieval run using DFR only for ranking, whereas "State_n" contain the results for our "State" approach with different values of α. As we can observe, retrieval effectiveness is improved significantly for a number of measures. Specifically the "State_0.05" configuration achieved a p value below 0.01 for both P@10 and P@15. We can see how the most prominent improvements are achieved at the top cut-off points. This result suggests that taking into consideration the structure of documents, helps in bringing more relevant documents

Table 4.3: Experimental results for the State retrieval method on the 2011 and 2012 collections. (* $p < 0.05$ and † $p < 0.01$)

	P@5	P@10	P@15	P@20	P@30	MAP
DFRee	0.458	0.432	0.399	0.382	0.362	0.109
State_0.02	0.451	0.434	0.408	0.396*	0.358	0.108
State_0.03	0.475	0.452†	0.414*	0.395*	0.362	0.108
State_0.05	0.478	**0.469**†	**0.428**†	0.395*	**0.369**	**0.110**
State_0.07	**0.481**	0.454	0.416	**0.398***	0.361	0.107
State_0.10	0.458	0.424	0.397	0.377	0.349	0.103

to the very first few documents to be read, which is a highly desirable outcome due to the fast-paced environment that is microblog search.

We can conclude from these experiments that the structure of tweets can be extracted and leveraged to produce better rankings. We can confirm that not only it is the relative space in terms of characters dedicated to each dimension that links to relevance, but also how these dimensions relate to each other within the document.

4.3.1 Additional notes

The simplicity of the state modelling allows for it to be conveniently stored and re-used in real-time. The states are stored as a set of precomputed heuristics which include the transitions between dimensions and the associated probabilities based on the observed data. The model itself could be updated from time to time to accommodate any shifting in the structuring and style of micro-bloggers. However it is not expected to change considerably, as it is a reflection of the consequences brought by the medium limitations.

4.4 Conclusions

In this work, we defined a microblog document as a 4-dimensional entity. In the case of Tweets, the document contains 4 distinct dimensions namely, Text; Url; Mentions and Hashtags. Then, we proposed the notion of “Microblog Informativeness”, which states that a microblog document’s relevance - or interestingness - with respect to a user’s *information need* expressed as a query, has a significant relationship with the structure of the document in terms of how many characters are dedicated to each dimension.

In order to test our hypotheses, we propose a number of techniques which utilise the number of characters used for each microblog dimension to re-weight the retrieval score of a microblog document. By doing so, we were able to significantly improve the performance of a state of the art retrieval model in the context of ad-hoc microblog retrieval.

Finally, we extend our analysis to account for the different variations in the ordering of microblog dimensions. We devised state machines to model the structure of known relevant and non-relevant documents. Then we developed an approach that makes use of the probabilities provided by such state machines to produce scores which reflect on the structure of the documents. Our experimentation, shows with statistical significance that it is possible to utilise the structure of tweets to improve their ranking in an ad-hoc retrieval scenario.

Future work will further expose the relations between these dimensions as well as finding further applications of the features described in this work for other purposes, such as Automatic Query Expansion.

Part III

Query Performance Prediction

Chapter 5

Query Performance Prediction on Microblogs

5.1 Introduction

Most information retrieval systems experience a high variability in retrieval performance across different queries. Whilst many queries are satisfied successfully, the system produces poor results for many others. Since a number of retrieval approaches rely on the initial set of results, it would be highly desirable to predict when queries are not being properly satisfied, in order to address them accordingly.

This task is known as query performance prediction (QPP), and has been an active and challenging area of research over the last decade. Multiple predictors have been proposed in the literature with varying degrees of success. These predictors fall mainly into two categories: pre-retrieval and post-retrieval predictors. Pre-retrieval predictors are computed before retrieving any documents, thus relying solely on features related to the query terms. On the other hand, post-retrieval predictors, rely on features extracted from the retrieved documents. Post-retrieval predictors mainly estimate how well a query is represented by retrieved documents.

In this work we study pre and post retrieval predictors for microblog retrieval tasks. Although much work has been done in predicting the performance of queries over web collections, to the best of our knowledge, no work has been done in the context of microblogs. Microblogging platforms such as Twitter have gained momentum over recent years providing a new way of sharing information and broadcasting short messages over a network of users. Microblogs present many differences with respect to web documents both in morphology and content. Mainly, microblogs constitute a time ordered stream of very short documents as they are published. Moreover, microblogs contain community defined tags to refer to certain topics (hashtags), or people (mentions), which we intent to investigate in our QPP study.

The motivation behind studying QPP for microblogs resides in increasing the robustness of existing retrieval approaches. More specifically, QPP can be especially handy for selectively applying pseudo relevance feedback (PRF) based automatic query expansion (AQE) approaches. PRF-based AQE approaches rely on the initially retrieved set of documents. Thus if these documents loosely represent the initial information need, PRF-based approaches most likely result in unexpected behaviour, and worsened results.

Effective QPP represents an opportunity to estimate the performance of a system for a particular query, based on pre-retrieval and post-retrieval features. In turn, this would allow an IR system to selectively perform AQE when the circumstances are most propitious, based on estimates given by predictors.

Our work in this Chapter is driven by two research questions. **(RQ1)** To what extent we can predict the performance of a retrieval model in the context of microblogs?. **(RQ2)** To what extent, the combination of predictors can improve overall prediction performance, in the context of microblogs?.

In this Chapter, we investigate the performance of previously proposed predictors by Hauff et al. (2008) in the context of microblogs. We subsequently show that they fail to perform effectively which prompts the need for better predictors. Consequently, we propose a number of predictors, which take into consideration the characteristics of microblogs. Our evaluation findings show how our predictors outperform those found in the literature in the context of microblogs. Finally we further improve our performance by producing a machine-learned prediction model which combines our predictors by means of a support vector machine (SVM) for regression.

5.2 Related Background

One of the main works in query performance prediction is that by Cronen-Townsend et al. (2002). In their work they proposed a predictor is based in the Kullback-Leiber divergence between the query's and the collection's language models. This predictor attempts to quantify the "clarity" of the query. In other words, the non-ambiguity of the query which in turn should reflect on how well it represents a particular topic. Their evaluation shows good correlation of their predictor with average precision, using Spearman's ranking correlation tests.

Work by He and Ounis (2004) extended previous work by suggesting other predictors such as the standard deviation of IDF values within the query. They also defined a simplified version of the "Clarity Score" proposed by Cronen-Townsend et al. (2002) namely Simplified Clarity Score (SCS). Finally they also proposed an alternative to SCS called query scope (QS). Their main objective was to investigate pre-retrieval predictors, as post-retrieval predictors are normally computationally more expensive to use.

In order to predict query difficulty, He et al. (2008) proposed a query coherence score (QC-1) which attempts to quantify how related are the query terms to the retrieved set of documents as well as measuring the differences between the language used in the retrieved set and query, with respect to the collection. They found that their approach correlates well with average precision, using Spearmans's rank correlation test. Furthermore they also suggested two other versions of this score but their performance was poorer than their simpler first version. For their evaluation they used a number of retrieval models, including BM25 and TFIDF to retrieve documents from the TREC Robust track collection.

Work by Zhao et al. (2008) proposed a series of pre-retrieval performance predictors. One of the most succesful was SCQ. The aim of SCQ is to compute a similarity score between the queries and the collection. Moreover they also proposed a variability measure relying on the standard deviations of TFIDF scores for the query terms. Furthermore they also proposed a joint predictor using both previous approaches together. Their evaluation showed how their joint predictor outperformed all their approaches as well as previous work. It is important to note that their joint approach is slightly better than their simple SCQ, only when the linear interpolation gives most of the weight to SCQ, being much more complex, and computationally much more expensive. In this work we will evaluate the performance of SCQ in our particular context.

A short but comprehensive survey of performance predictors was produced by Hauff et al. (2008). Moreover they proposed a WordNet based predictor, which uses the number of senses of terms, as a measure of their ambiguity. The higher the number of senses associated with a term in the ontology, the most likely it is to produce poor results. Their approach did not outperform previous predictors, nonetheless, it is an interesting approach, that may prove useful since other components such as TF and IDF, may not be informative enough in Twitter corpora. This study helped on deciding which were the best performing predictors as a starting point of our study in the context of microblogs.

Evaluation methodology in the literature. As introduced by Hauff (2010), the evaluation of query performance prediction approaches can be formalised as follows:

$$f_{perf}(q, C, E, R) \longrightarrow \mathbb{R}$$

where f_{perf} denotes a numeric estimation of the performance of a query q in the $\mathbb{R}$, in terms of the information provided by C, E and R. C refers to a corpus of documents whereas R denotes a ranking method and E an external source. These estimations are used for ranking the predicted performance of these queries, and measure its alignment with respect to the actual effectiveness measured by some evaluation metric.

The "de facto" evaluation procedure in previous work has been the statistical correlations between the predictors and the evaluation metric results for a given system. More often than not, the evaluation metric used was Average Precision (AP). The better the predictor estimates the performance of the system in terms of an evaluation metric, the higher the correlation scores.

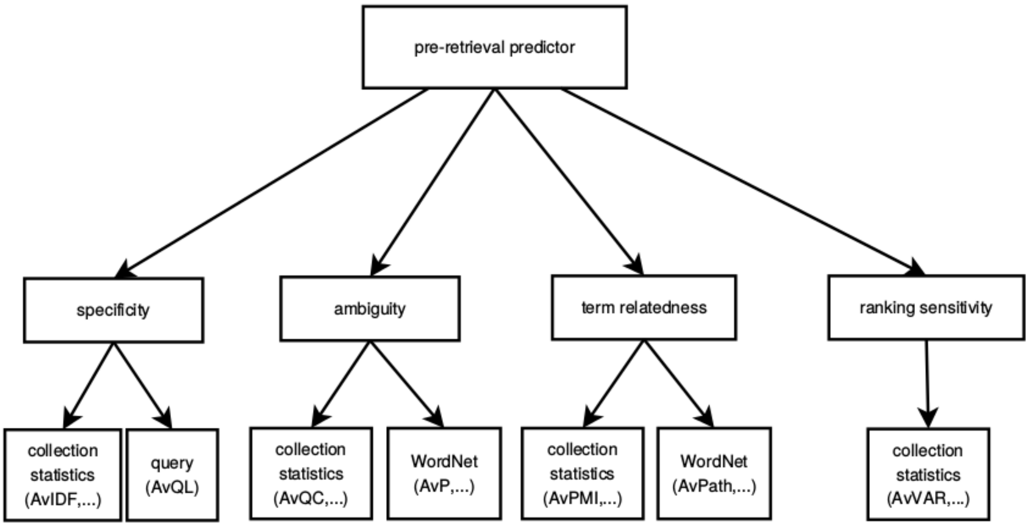
The most used correlation metrics are Kendall-Tau (K.Tau) and Spearman's (SP.Rho) rank correlation coefficients. The SP.Rho correlation coefficient is a measure of statistical dependence between two variables, by which it is estimated how well their relation is represented by a monotonic function (I.e.: grows/decreases always in the same direction). SP.Rho uses Pearson's correlation coefficient in such a way that is much less sensitive to outliers. Kendall-Tau's correlation coefficient is slightly different, as it does not rely on the values of the variables themselves, but it rather measures the similarity in the ordering of the data provided when ranked by each of the variables.

State of the art prediction. The correlation coefficients obtained for AP in web collections vary wildly. The Kendall-tau coefficients, with respect to AP, for the best performing pre-retrieval predictors range from 0.30 to 0.49 depending on the collection (Carmel and Yom-Tov, 2010). On the other hand, the Kendall-tau coefficients for post-retrieval predictors are generally higher.

It is important to note the high variability in terms or predicting performance, with respect to the collection. The collections used in the literature include "TREC Vol. 4+5"; "WT10g" and "GOV2", where it is often the case for a particular predictor to be the best for a particular collection and the worst for another.

Selective Query Expansion. One of the main applications of QPP is selective Query Expansion (Carmel and Yom-Tov, 2010). It refers to selectively applying automatic

Figure 5.1: Pre-Retrieval predictor taxonomy by Hauff (2010)


pre-retrieval predictor
specificity
ambiguity
term relatedness
ranking sensitivity
collection statistics (AvIDF,...)
query (AvQL)
collection statistics (AvQC,...)
WordNet (AvP,...)
collection statistics (AvPMI,...)
WordNet (AvPath,...)
collection statistics (AvVAR,...)

query expansion (AQE) whenever predicted performance is above a certain threshold. This serves as a warranty for PRF-based AQE approaches, as they rely on the top N retrieved documents to perform optimally.

5.3 Predictors

In this section, first we describe the predictors we will be considering in our evaluation, including our proposed ones. Secondly we introduce the evaluation approach followed to benchmark and compare their performance.

5.3.1 Predictors in the literature

Many predictors have been proposed in the literature. They are mainly defined as pre-retrieval, or post-retrieval predictors. Pre-retrieval predictors rely only on information associated with the query terms and their collection statistics, as well as external information such as that provided by semantic taxonomies. On the other hand, post-retrieval predictors rely on information, extracted from the documents retrieved. The later predictors, therefore highly depend on the retrieval model used.

The work by Hauff (2010) provided a taxonomy that organises pre-retrieval predictors in terms of the information they depend upon and the features they are trying to estimate. The taxonomy is presented in Figure 5.1. This will be useful for organising previous work, as well as putting our proposed predictors in context, thus allowing for their comparison with previous work. As we can observe there are four main groups namely: **specificity**; **ambiguity**; **term relatedness** and **ranking sensitivity**.

Specificity predictors. Firstly we introduce the **QueryTermIdf** predictor. This predictor utilizes the IDF values of query terms as a means to estimate the system's retrieval performance. The intuition is that the higher the IDF value the more specific a term is. Furthermore, score variations across terms may indicate drifting concepts, negatively affecting performance. We derive different predictors considering the mean, median, standard deviation (Std), max, min and diff($max - min$) IDF scores from each query (Hauff et al., 2008). Moreover, **Simplified Clarity Score (SCS)** proposed by He and Ounis (2004), attempts to model the clarity of a query, i.e. how well it targets a particular topic based on collections metrics. An homologous predictor to SCS is **Query Scope (QS)**, which was also proposed by He and Ounis (2004).

Similarity of Collection w/ Query (SCQ) is another specificity predictor, which was proposed by Zhao et al. (2008). SCQ simply computes the similarity between the collection and the query at hand.

Ambiguity predictors. This category refers to those measure the semantic ambiguity of query terms. The intuition is that the more ambiguous query terms are, the worse the retrieval results will be. In the work by Hauff et al. (2008), a predictor to measure the semantic ambiguity of query terms was proposed using a semantic ontology. The **Ambiguity** predictor relies on the hyponym relation between terms found in WordNet. Hyponyms relations are homologous to being the sub-class of something. (E.g. Dog is a hyponym of mammal). Intuitively, the more hyponyms a term has, the higher its ambiguity, thus increasing its likelihood to harm retrieval performance.

Term Relatedness predictors. Term relatedness predictors measure how related pairs of terms are across the collection. One of such predictors is point mutual information (PMI). PMI computes the co-occurrence of all terms in a collection, and assumes the more query terms co-occur the more likely they treat a particular topic, and therefore results are more likely to be satisfactory. PMI is given by:

$$PMI(t_1, t_2) = \log \frac{P(t_1, t_2|D)}{P(t_1|D)P(t_2|D)} \tag{5.1}$$

where $P(t_1, t_2|D)$ gives the number of documents where t_1 and t_2 co-occur, and $P(t_1|D)$ and $P(t_2|D)$ are the number of documents where t_1 and t_2 occur.

Ranking Sensitivity This category of predictors, attempt to measure the query's effectiveness in discriminating documents. The intuition is that if query terms appear in similar documents, then these documents become undistinguishable by the retrieval system, and the query is predicted to be ineffective. These predictors work exclusively on collection statistics. One of such predictors is **Term Weight Variability (VAR)**, which measures the variability of weights for a term across the collection. Zhao et al. (2008) hypothesises that the higher the standard deviation for a term, the more discriminative it is, leading to better performance than terms with lower standard deviation.

Post retrieval predictors. Furthermore, in the work by Carmel and Yom-Tov (2010) four post_retrieval predictors were introduced, namely **NQC**, **WIG**, **QF** and **Clarity**. NQC measures the normalized standard deviation of the top scores. The intuition behind this predictor is that relevant documents are assumed to have a much higher score than that of the mean score. WIG works in a similar fashion, by measuring the divergence of retrieval scores of the top-ranked results from that of the documents in the corpus.

Finally, QF and Clarity are predictors that take into account the actual content of the documents. QF measures the divergence between the original top results for the query and the results that would be obtained for a query constructed from the top results. Finally, Clarity measures the KL divergence between a (language) model induced from the result-list and the corpus model.

5.3.2 Proposed Predictors

In this subsection we introduce our proposed predictors, which are mainly based on post-retrieval features.

Query Coverage Predictors. A common property to all retrieval models is that documents covering more query terms will obtain a higher score than those covering the query partially. Equation 5.2 exemplifies this by representing a model scoring function $P(Q, D)$.

$$P(Q, D) = \sum_{q_i=0}^{|Q|} P(q_i, D) \tag{5.2}$$

where Q is the set of all query terms q_i and D is the document being scored. As we can observe the higher the number of query terms found in the document the higher the score resulting from the sum. The intuition behind is that documents including the highest number of query terms are most likely to satisfy the user's information need as the match it more closely. Particularly for microblog retrieval, given the scarcity of term frequencies, makes the presence or absence of query terms a very determinant feature towards estimating the relevance of a microblog document. Based on this assumptions we define two predictors: **CoveredQueryTerms** and **TopTermsCoverage**.

The **CoveredQueryTerms (QTCov)** predictor measures how well the query is being represented by the documents in the result list.

$$cov(q_i, d_j) = \begin{cases} 1, & \text{if } tf(q_i, d_j) \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

where $cov(q_i, d_j)$ is a function that returns 1, whenever a term q_i is present in document d_j, and 0 otherwise. Moreover, QTCov may be defined as:

$$QTCov(q_i, D) = \frac{\sum_{j=0}^{|D|} cov(q_i, d_j)}{|Q|}, \tag{5.4}$$

where the rate of query terms in Q appearing on each document is aggregated and normalized between 1 and 0. (1 being a document that completely fits the query). This predictor attempts to directly model the intuition that drives every retrieval model, producing a higher value when the query is being properly matched.

Similarly we defined **TopTermsCoverage (TTCov)** which measures the coverage of the top N terms in the result list. When documents describe, or revolve around a particular topic, they will inevitably share a common vocabulary. This relation between query terms and terms within the top retrieved documents is already exploited in other contexts, such as PRF-based AQE.

Furthermore, Cranfield experiments usually rely on topical relevance, for producing the relevance assessment set. The topical relevance assumption is that a document is relevant to a query if it contains information about the topic at hand, regardless of the utility of the document. The Cranfield evaluation paradigm is used widely for creating the relevance assessments of the test collections, and TREC's microblog collections are no exception. Therefore, finding documents which are on topic, by containing top appearing terms, should be a good indication of the system's retrieval performance. TTCov is thus defined as:

$$TTCov(t_i, D) = \frac{\sum_{j=0}^{|D|} cov(t_i, d_j)}{|T|}, \tag{5.5}$$

where t_i is a term contained within the set of top occurring terms T. The set T is parametrised, and during this experiments was defined to contain the top 3 most occurring terms.

Time Specific Predictors. Time is of the essence in microblog search. As millions of Tweets are published, others become obsolete because users are only interested in the most up to date information available for their topics of interest. This family of predictors uses publication times to estimate the quality and representativeness of the documents being retrieved.

An example of the distribution of relevant documents in time can be seen in Table 5.1, which shows statistics for the differences in publication times for both relevant and non-relevant tweets from query number 36 in the microblog 2011 collection. The tweets were ordered with respect to time, and the difference between publication time of d_i and $d_(i+1)$ was computed. All the statistical measures shown in the table are computed from these differences. As we can observe there are substantial differences, between the relevant documents set and the non-relevant set. The non-relevant set of documents is considerably more spread out throughout the time whereas the relevant documents, are much closer together with respect to time. This observation motivated the introduction of the time-based predictors named **TimeCohesion** and **QueryTimeDistance**.

Table 5.1: Differences between the publication times of tweets (Scaled down: $time(d_i) \cdot 10^{-9}$). Differences are statistically significant $p < 0.001$

Time Diffs	median	avg	lower percentile	higher percentile
Rel. Docs	439.59	897.97	237.07	974.79
NonRel. Docs	645.94	2378.29	57.32	2457.23

TimeCohesion (TimeCH) is a predictor which taps into the distribution of retrieved tweets over time. We assume that the closer documents appear with respect to time, the more likely they refer to the same event or topic.

$$TimeCH(D) = \sum_{i=0}^{|D|-1} time(d_{i+1}) - time(d_i), \tag{5.6}$$

where $time(d_i)$ is a function returning the publication time of document d_i contained in set D.

To compute it, we take the differences between retrieved document timestamps. Differences are taken only between contiguous documents in the rank.

QueryTimeDistance (TimeDist). This predictor takes into account the real-time nature of microblog search. Users submitting queries to a microblog search engine are interested in knowing about up to date information, which often has not even reached traditional sources of media. Therefore, often the queries are issued very close in time to the publication time of the relevant documents that may satisfy it. TimeDist is defined as:

$$TimeDist(D) = \sum_{i=0}^{|D|} time(Q) - time(d_i), \tag{5.7}$$

where $TimeDist(D)$, is the aggregation of differences between the time the query was issued, and the publication time of documents in D.

Microblog Specific Predictors. Microblog documents have specific features, that have been shown to have some connection with the relevance of documents in previous work. Examples are the presence of Urls and length of the tweets (Gurini and Gasparetti, 2012).

We define the **Http** predictor to exploit the presence of Urls. A common behaviour by microblog users is to provide a short description of the information to be published followed by a Url, which in turn points to a relevant article expanding the information they referred to (Teevan et al., 2011). Thus, intuitively, the presence of a Url in a microblog document, where the information is quite limited, is often an indication that important information is to be communicated. Particularly, this predictor measures how common is to find a Url in the retrieved set of documents. To this end we compute the rate of documents with a Url in the result list as follows:

$$Http(D) = \frac{\sum_{i=0}^{|D|} hasUrl(d_i)}{|D|}, \tag{5.8}$$

where $hasUrl(d_i)$ is a function returning 1 if the document d_i contains a Url, and 0 otherwise. The result is the number of documents containing a Url divided by the total number of documents contained in set D (I.e the rate of Urls). Finally,to find and match the Urls, we utilize regular expressions taking into account every possible web protocol used to define a Url (e.g.: http, https, ftp, etc).

HashTagCount similarly to the Http predictor is defined as the rate of documents with hashtags in the retrieved results set. Hashtags are important in the context of Twitter as they refer to particular topics. Thus the presence of similar hashtags repeatedly across documents might indicate that the different users are speaking about the same topic. The HashTagCount predictor is given by:

$$HashTagCount(D) = \frac{\sum_{i=0}^{|H|} HashTagFreq(h_i, D)}{|D|}, \tag{5.9}$$

where $HashTagFreq(D)$ returns the frequency of a hashtag h_i appearing in the result set D. The rate of hashtags in the result set is found by dividing by the total number of documents in D. In this work, apart from the sum, we also consider the mean, median and max of these rates as standalone predictors.

Finally, we proposed the **TweetLength** predictor, which is defined as the number of terms in each retrieved tweet, after stop-words removal. As tweets are very small in length, the variations between document lengths can have a greater effect in those retrieval models that depend on document length normalization, with respect to retrieving web documents. Moreover, document length has been shown some connection with the relevance of documents (Gurini and Gasparetti, 2012).

5.4 Evaluation

In the literature, evaluations have mainly taken into account Pearson, K.Tau or SP.Rho as correlation measures with respect to average precision (AP). The user model considered in the literature when investigating QPP approaches, takes into consideration a vast number of documents per topic, thus AP represents is a good choice. However, in Microblog retrieval, it is most important to optimise performance for the first retrieved documents due to its real-time nature. It has been agreed in the literature that a user will not look further than the first 30 documents, thus AP might not be appropriate for this task. This is specially true if we want to optimise our QPP approaches to improve PRF-based AQE. Therefore, due to these constraints we focus on the very top retrieved documents paying attention to P@15 for query performance prediction purposes.

In this evaluation we utilize the Tweet2011, 2012 and 2013 collections for a total of 170 topics. The collections have been merged together to produce enough evidence for learned predictor models.

5.5 Results and Discussion

In this chapter we introduce and discuss the results obtained for the different stages of this work. Firstly we introduce the hypothetical benefits of a system that could determine, in advance, whether applying a PRF-based AQE technique would result in improved performance. Secondly, we study the performance of predictors introduced in the literature and compare them to our proposed predictors. Then, we utilize some machine learning regression approaches to combine the different predictors to optimize performance. Finally, we show the results obtained when classifying the different retrieval runs into three classes according to their performance.

5.5.1 Is Selective Automatic Query Expansion useful?

The utility of PRF-based AQE approaches has been demonstrated in many tasks. However it is not clear which queries are most likely to benefit from it, and which ones would have their performance hindered. In the literature, it is assumed that the better

Table 5.2: **Oracle** Selective Automatic Query Expansion performance on microblog collections 2011 - 2013. (** $p < 0.01$

	DFR	DFR+ROCCHIO	ORACLE-AQE
P@10	0.5274	0.5696	**0.6000****
P@30	0.4087	0.4534	**0.4718****

the performance of the initial search on a PRF scenario, the better the final results will be. While this is a valid intuition, we carried out experiments to measure to what extent does this assumption hold in the context of microblog ad-hoc retrieval.

We produced runs using DFRee as a baseline retrieval system, and expanded the queries by a traditional Rocchio query expansion approach (Carpineto and Romano, 2012). Then, we analysed the distribution of topics being benefited/hindered out of a total of 168 topics from TREC's 2011, 2012 and 2013 microblog collections:

- 21.42% topics had worse performance
- 32.14% topics had better performance
- 46.44% topics remained unaffected

As we can see, the topics being negatively affected by the AQE algorithm represent a considerable amount with only 11% of difference with respect to those being improved. Whilst overall AQE provides better system performance, there is much room for improvement.

Additionally, if a retrieval system was able to predict when a query is going to suffer from applying AQE on it, we could avoid the negative effects of those 21.42% failed topics.

Table 5.2 shows results obtained by such hypothetical system, namely **ORACLE-AQE**. This run was obtained by, using the initial run instead of the expanded run, whenever the performance has dropped after the AQE step. Additionally we show the results for runs utilising DFR as a baseline and Rocchio for PRF-based query expansion denoted as **DFR+ROCCHIO**.

The oracle results provided by **ORACLE-AQE** show a significant improvement of +5.39% and +4.06% over the DFR+ROCCHIO runs. Whilst the improvement does not seem huge in terms of P@10 and P@30 achieved, we have to keep in mind that 21.42%

of the topics are being better satisfied by ORACLE-AQE than by DFR+ROCCHIO. Therefore there is much room for improvement in terms of robustness.

Furthermore, we want to investigate whether the topics failing are those with low initial performance. To this end, we split the topics into three groups with respect to the P@10 performance obtained over the initial set. The groups are defined as: **Low** ($P@10 < 0.25$), **Medium** ($P@10 > 0.25$ *and* < 0.75) and **High** ($P@10 > 0.75$). This grouping provides a fair split in terms of number of topics per group (55,48 and 65 respectively). Results are as follows:

- Low: 29.09% improved; 25.45% worsened.
- Medium: 43.75% improved; 20.83% worsened.
- High: 26.15% improved; 18.46% worsened.

As we can observe the topics follow an intuitive trend. In the Low group the percentage of failing topics is highest, then failure reduces gradually as performance increases for groups Medium and High. In the case of the topics that improved performance, we would expect an inverse relationship as there should be a higher percentage of topics improving in the High group than in the Low group, compared to those topics that failed.

However, in the case of the High performing group, the difference between improved and worsened topics is not very different from the Low performing group. This behaviour is likely due to those cases in which topics already have the best terms that could be found in top retrieved documents, therefore could only be improved using an external source. On the other hand, terms in the initial query may be so discriminative of a group of documents, that additional terms do not contribute to the selection of documents in the top ranks.

We can conclude from this analysis that, high performing topics show a similar behaviour to those in the low end. Therefore, we must target both the High and the Low performance groups through AQE if we want to have a significant impact with our proposed approaches.

MAP correlations			
Predictor	**K.Tau**	**SP.Rho**	**Pearson**
post_TTCov_Mean	0.302 **	0.447 **	0.403
post_TTCov_Median	0.253 **	0.312 **	0.274
post_TTCov_upper	**0.356 ****	**0.463 ****	**0.434**
post_TTCov_Lower	0.178	0.218 **	0.197
post_TimeCH_Lower	-0.202 **	-0.300 **	-0.273
post_TimeCH_Median	-0.200 **	-0.291 **	-0.310
post_TimeCH_Upper	-0.122 *	-0.188 *	-0.231
post_TimeCH_Mean	-0.197 **	-0.288 **	-0.281
pre_SCQ_Sum	0.094	0.138	0.254
pre_QueryTermIdf_Diff	0.140 **	0.209 **	0.200

Table 5.3: Predictor correlations with MAP for retrieval runs using DFRee (**$p < 0.01$ & *$p < 0.05$)

5.5.2 Evaluating Query Performance Predictors

Firstly we analyse the performance of existing predictors, as well as our proposed predictors for QPP in microblogs.

Tables 5.3 and 5.4 show the correlation coefficients in terms of K.Tau, SP.Rho and Pearson for a subset of predictors. Since it was not possible to show all the predictors in this thesis, we have chosen to include only those achieving a Pearson coefficient higher than 0.19.

The predictors are prefixed with either ”pre_” or ”post_” to indicate whether they are pre-retrieval or post-retrieval predictors. Furthermore, the suffixes: Mean, Median, Std, Max, Min, Lower and Upper; denote mean, median, Standard Deviation, maximum, minimum, lower percentile and upper percentile, of the predictor values respectively. Moreover, Sum refers to the Sum of all predictor values, whereas Diff is the difference between Max and Min.

Table 5.3 shows the correlations coefficients in terms of MAP. In the literature most work has been carried out to predict this particular evaluation metric, thus we provide this table for reference purposes. In the survey done by Hauff et al. (2008) the maximum correlation achieved using K.Tau ranged from 0.30 to 0.49 depending on the collection. As we can observe, the correlations coefficients obtained in our case are slightly weaker in terms of AP than what it has been obtained in the literature.

P@10 correlations			
Predictor	**K.Tau**	**SP.Rho**	**Pearson**
post_http	0.163 **	0.206 **	0.213
post_QTCov_mean	0.291 **	0.382 **	0.375
post_QTCov_median	0.305 **	0.382 **	0.373
post_QTCov_upper	0.325 **	0.404 **	0.392
post_QTCov_lower	0.266 **	0.336 **	0.312
post_TTCov_mean	0.301 **	0.416 **	0.429
post_TTCov_median	**0.365 ****	**0.456 ****	**0.441**
post_TTCov_upper	0.264 **	0.355 **	0.374
post_TTCov_lower	0.253 *	0.303 **	0.298
post_TimeCH_lower	-0.212 **	-0.286 **	-0.236
post_TimeCH_median	-0.145 **	-0.199 *	-0.239
post_TimeCH_mean	-0.170 **	-0.233 **	-0.212
post_TimeCH_diff	0.192 **	0.269 **	0.198

Table 5.4: Predictor correlations with P@10 for retrieval runs using DFRee ($^{**}p < 0.01$ & $^{*}p < 0.05$)

State of the art predictors **SCQ**, **VAR**, **SCS** and **QS** (Described in Section 5.3) performed poorly in the context of microblogs, as their K.Tau coefficient values ranged between 0 and 0.16, thus are not shown in Tables 5.3 or 5.4. This under-performance demonstrates how challenging query performance prediction is in the context of microblog retrieval, and the very need for tailored predictors to this new task.

On the other hand, the best results in terms of AP and Precision are produced by our predictors (Refer to Appendix A for the values obtained by other predictors). **post_TTCov_upper** is one of such predictors, achieving a K.Tau coefficient of 0.356, being the best correlation with respect to MAP. This predictor takes the upper percentile of the rate at which top terms appear in the retrieved set of documents.

Whilst the results in terms of AP are important, our main focus is the study of QPP for the purpose of PRF-based AQE, thus we pay special attention to the correlation coefficients with respect to P@10 in Table 5.4. As it can be observed, amongst the top performing predictors we find those relying on microblog specific features, namely **post_TimeCH** which measures how close in time are the retrieved tweets and **post_http** measuring the presence of URL's in documents. Additionally, the correlations achieved by these predictors with respect to P@10, are generally higher than what was achieved for AP, with **post_TTCov_Median** being the best performing predictor.

An interesting observation regarding **post_TTCov_upper** and **post_TTCov_Median** is that they may be referring to the same documents, as with MAP a larger set of documents is considered compared to P@10.

We can conclude from these results that, whilst overall the correlations obtained are yet not strong enough as a predictive tool, we have improved over state of the art predictors, thus we are one step closer to making Microblog query performance predictions a reality.

5.5.3 Linearly Combining Best Predictors

In the previous section we experimented with single predictors, producing results that outperform state of the art approaches. Since, predictors are inherently different in terms of their design, it is possible to combine them, so as to cover the limitations of one with the strengths of the other. As a proof of concept we attempt to combine the best two predictors together **QTCov_median** and **TTCov_median**. These predictors are complementary as one deals with query terms, whilst the other deals with top occurring non-query terms.

To combine them we used linear regression with respect to the score achieved in terms of P@10. Moreover, the experiments were carried out using 10-fold cross validation to reduce any effects on data bias. The combination of these predictors, namely **QEAndQT_50**, produced the following model:

```
QEAndQT_50 =
0.5656 * QTCov_median +
0.5487 * TTCov_median +
-0.0355
```

The combined effort of both predictors results in considerably improved performance, giving a Pearson correlation value of 0.5387. This result translates to a +22.15% improvement when compared to that achieved by the best single predictor. As we can observe, the weights assigned to each predictor are very close to each other, which suggests that they contribute almost equally to the predictions.

5.5.4 Combining Predictors by SVM

Similarly to the previous section, we attempt to combine different predictors together to enhance the overall performance. In this particular case we used the popular Support Vector Machine (SVM) for regression to predict the value of P@10, avoiding any biasing by performing a ten-fold cross-validation. The resulting learned prediction model is defined as follows:

```
P_10 =  0.3028 * TTCov_upper
 + 0.3494 * QTCov_median + 0.3701 * QTCov_upper
 - 0.4745 * twids_median - 0.2641 * TTCov_mean
 + 0.5014 * twids_mean + 0.3394 * TTCov_median
 + 0.2318 * TTCov_lower + 0.3122 * twids_diff
 + 0.2429 * http - 0.1651 * QTCov_lower
 - 0.2745
```

The correlation coefficients obtained for this model, are 0.412 (+12.88%), 0.559(+22.59%), and 0.539 (+22.22%), for K.Tau, SP.Rho and Pearson respectively. As it can be observed the performance achieved in terms of Pearson is comparable to the previous approach which linearly combined **QTCov_median** and **TTCov_median**, thus suggesting these are a set of very prominent features which may be eclipsing the effects of other non-complementary features. Finally the choice of SVM is due to its popularity within the IR community, thus any other approach can be used which could yield different results.

5.5.5 Feature Selection and SVM-SMO

In Subsection 5.5.1 we observed that for those topics with Medium to High performance the percentage of improved topics by means of AQE was considerably higher than for those runs with low performance ($P@10 < 0.25$). Therefore, if we managed to predict when a topic belongs to each class, we could disable AQE for those in the low performance group, as the likelihood of the query being successfully expanded is almost the same as failing to expand it, thus reducing the randomness of the algorithm. To this end, we try to classify these three classes (low,medium,high) with respect to the previously mentioned predictors. Before combining features, the most appropriate ones must be selected. We do this by pruning those that contribute weakly to the predictions or have no contribution at all. The contributions of each predictor in isolation with respect to an SMO classifier are as follows:

```
=== Attribute selection 5 fold cross-validation ===

number of folds (\%)  attribute
           1( 20 \%)     1 SCQ_max
           5(100 \%)     2 SCQ_sum
           0(  0 \%)     3 VAR_max
           0(  0 \%)     4 VAR_sum
           1( 20 \%)     5 QE_SETCS
```

Table 5.5: Classifying different levels of performance

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.782	0.354	0.518	0.782	0.623	0.732	Low
0.208	0.117	0.417	0.208	0.278	0.528	Medium
0.677	0.165	0.721	0.677	0.698	0.744	High
0.577	0.213	0.568	0.577	0.554	0.678	Avg.

```
4( 80 \%)     6 QEAndQT_25
0(  0 \%)     7 QEcoveredQueryTerms_upper
0(  0 \%)     8 coveredQueryTerms_median
1( 20 \%)     9 twids_lower
0(  0 \%)    10 coveredQueryTerms_upper
0(  0 \%)    11 twids_median
0(  0 \%)    12 QEcoveredQueryTerms_mean
1( 20 \%)    13 twids_mean
0(  0 \%)    14 QEcoveredQueryTerms_cond
0(  0 \%)    15 QEcoveredQueryTerms_median
1( 20 \%)    16 QEcoveredQueryTerms_lower
0(  0 \%)    17 coveredQueryTerms_mean
1( 20 \%)    18 twids_diff
4( 80 \%)    19 QEAndQT_50
0(  0 \%)    20 QEAndQT_75
3( 60 \%)    21 http
1( 20 \%)    22 coveredQueryTerms_lower
```

Having found those predictors with no contribution (0%), we performed a classification run by means of an SMO classifier. The results achieved in this case are shown in Table 5.5. As we can observe, some of the classes are more predictable that others. In particular, those topics with **high** performance are correctly classified with good precision. The **middle** performance class however, seems to be much more difficult to predict obtaining really low results in terms of precision and recall, amongst other metrics.

Furthermore, recall and the True Positive (TP) rate is quite high for the **low** performance group. This group is of especial interest as it has practically the same rate of successful/failed query expansion runs, thus it is a main contributor in the algorithm's lack of robustness.

5.6 Conclusions

In this chapter, we studied the performance of the state of the art predictors in the context of microblogs. In our study we focus on predicting query performance for increasing the robustness of PRF-based AQE approaches. Consequently we paid especial attention to the query performance prediction in terms of evaluation metrics regarding the top retrieved documents. Our evaluation suggests that predictors in the literature perform poorly in the context of microblogs, thus we need to come up with predictors that are better fit for purpose. To this end, we defined a number of predictors relying on microblog features and characteristics. We then benchmarked their performance and showed that most of them outperform those in the literature, with **TTCov** being the most correlated with MAP and P@5.

Whilst some of the predictors we proposed, such as TTCov and QTCov considerably outperformed state of the art prediction models in the context of microblogs, their performance on their own is still insufficient for effective **selective** query expansion. In order to improve over the performance of our best predictors we devised a set of experiments to combine them together. The first of such experiments used support vector machines for regression to learn a prediction model based on the best performing predictors. The resulting model further increased performance by a +22% in terms of the Pearson correlation coefficient, and +12.88% for K.Tau.

Secondly, we looked at the same problem from a classification point of view. To this end we defined three different topic groups, according to performance ranges measured by P@10. We then attempt to classify topics into each of this categories, in order to decide whether to apply AQE or not. Our evaluation experiments show promising results in classifying low ($P@10 < 0.25$) and high ($P@10 > 0.75$) performance topics, whilst topics with medium performance ($P@10 > 0.25\ and P@10 < 0.75$) are much harder to predict.

This chapter represents an initial step into bringing query performance prediction to the domain of microblog ad-hoc retrieval. The main goal is to provide robust mechanisms to improve automatic query expansion. We do so by finding predictors that may help in estimating how appropriate is the pseudo relevant set before applying query expansion techniques. Thus this allows to selectively apply AQE - or similar - approaches when the conditions are most propitious.

Future work will put these findings to a practical application in selective approaches to PRF-AQE, or in the selection of a baseline model to optimize a system's overall performance given the conditions of a particular query. Furthermore, we will study the performance of other predictors which will consider more microblog specific features.

Part IV

Automatic Query Expansion

Chapter 6

Automatic Query Expansion on Microblogs

6.1 Introduction

The character length restriction policies found in microblog documents, such as tweets, results in term sparsity, which in turn leads to what is known as the **"term (or vocabulary) mismatch problem"**. This problem has been studied as early as in 1987 by Furnas et al. (1987) and is produced by the difference between the vocabulary used in the formulation of the query, and that of the relevant documents desired by a searching user. Thus it is no surprise that the length limitations in microblog documents exacerbates this problem, mainly due to the very limited useful information available in them to match the query.

In recent years, two main paths have been followed to bridge the representational gap between queries and the desired documents, namely **automatic query expansion** (AQE) and **document expansion**. AQE refers to modifying and/or enriching an initial query with new terms. These terms are often mined either from an initial set of retrieved documents (Pseudo Relevance Feedback), or some external source. Document expansion, on the other hand, attempts to enhance the representation of documents, normally relying on external sources.

However we hypothesise that AQE approaches reliant on the scores produced by retrieval models can produce unreliable behaviour in the context of microblog ad-hoc retrieval (Related to **RQ4**). To this end we propose a novel **term selection approach** which promotes terms whilst not relying on the score value assigned to documents by a retrieval model (Contribution **C7**). Our approaches utilise instead the rank number assigned to the documents to estimate the importance of terms contained within them. Additionally we apply two different normalisation techniques with respect to the rank to gradually reduce the value of terms within. These functions - namely linear and logarithmic - are coupled with the document frequency value for a term in the pseudo relevant set or collection based statistics such as IDF. We compare our approach to the state of the art methodology RM3 which utilises the score assigned by the retrieval model to the documents. Our evaluation shows how our rank-based approaches perform significantly better than the any given baseline more often than the state of the art RM3 approach.

The second part of this chapter studies an alternative approach to estimating the quality of prospective terms for query expansion (Related to **RQ5**). We isolate features based on IDF values extracted terms from the query and relevant and non-relevant documents. We assume "high quality" terms appear only on relevant documents; "medium quality" can be found both in relevant and non-relevant documents, and finally "low quality" terms can only be found in non-relevant documents. We then utilise this classes to build a classifier from training data to predict these term categories. The classifier information is then utilised to determine a boost parameter, which we then utilise to enhance the behaviour of RM3 (Contribution **C8**).

We show how our approach - namely RM3_TQP - significantly outperforms the behaviour of RM3 for multiple evaluation metrics, on testing sets. Thus we demonstrate that it is possible to predict the quality of term before it is used for automatic query expansion.

In this chapter we first introduce the common background which will be used throughout the rest of the chapter. Secondly we describe our discounting approaches in Section 6.3, the experimental framework in Section 6.4 and we discuss the results in Section 6.5. Finally we explore a technique to predict the quality of terms in order to improve the behaviour of a state of the art AQE method in Section 6.7, introduce the evaluation setting in Section 6.8 and discuss the results in Section 6.9. Finally we conclude the Chapter in Section 6.10.

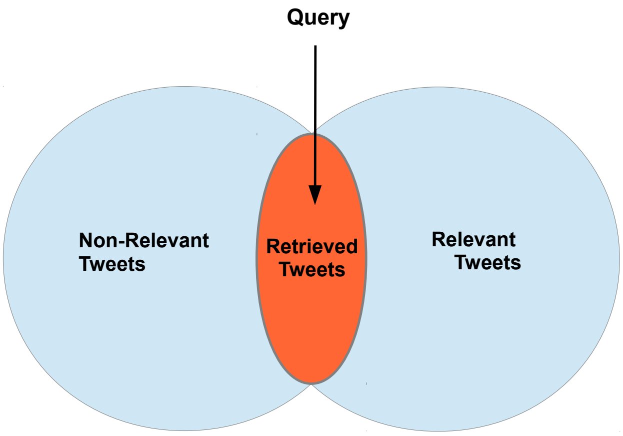
6.2 Related Background

In this section we cover the related background to this chapter including the vocabulary mismatch problem, automatic query expansion (AQE) or pseudo relevance feedback.

The **vocabulary mismatch problem** refers to the reduced chances of matching a query terms with relevant documents due to the differences in terms of the vocabulary utilised to express the query and that used in the document. This problem, also known as the **"term mismatch problem"**, has been studied as early as in 1987 by Furnas et al. (1987).

Figure 6.1 illustrates how an issued query will - inevitably - retrieve documents from the intersection of both the relevant and non-relevant set of documents. If we could determine which terms are more common to the relevant set, we would be able to shift the focus of the query and better capture the information need of the user. This problem is specially pronounced in the context of microblog ad-hoc retrieval, due

Figure 6.1: Vocabulary/term mismatch problem. Documents are inevitably pulled from both sets due to the vocabulary used



to the limited information contained within microblog documents. Consequently the queries are often poorly matched leading to poor ad-hoc retrieval performance. The main goal of our work in this Chapter, is to bridge representational gap between the queries and relevant documents. To this end we study the characteristics of the terms contained within relevant and non-relevant documents and then devise techniques to capture their differences.

A common approach to alleviate the term mismatch problem is automatic query expansion (AQE). The objective of Automatic query expansion techniques is to expand the initial representation of a textual query by including new terms and/or balancing the weight of existing terms through some **term selection** mechanism. The source of the proposed new terms, may be the document collection itself or an external source.

Automatic Query expansion. Automatic query expansion approaches (AQE) have been the focus of research efforts for many years. Work by Carpineto and Romano (2012) introduce a comprehensive study about these approaches, giving insight on the challenges faced by these techniques. Most importantly it introduces critical issues such as parameter setting, efficiency and usability of the approaches. Moreover, in their work they propose a comprehensive description of the steps involved in any query expansion approach. These steps include:

1. **Preprocessing of Data Source:** This involves the tokenization, stop word removal and stemming of those terms found in the initial set of retrieved documents.

2. **Generation and Ranking of Candidate Expansion Features:** This stage refers to estimating the relatedness of terms found in the initial set of retrieved documents, with respect to the initial query.

3. **Selection of Expansion Features:** After the ranking of terms, a number of top ranked terms is selected following a given policy.

4. **Query reformulation:** At this stage, terms are added to the initial query following a policy, and normally weights are assigned.

Pseudo Relevance Feedback. An important concept in query expansion approaches is Pseudo Relevance Feedback (PRF) (Xu and Croft, 1996). AQE approaches such as Rocchio (Carpineto and Romano, 2012) rely on the knowledge of relevant documents, to ensure the source of expansion terms is reliable. However in many cases, we do not have explicit knowledge as to which documents may be relevant to the query, and thus we need to find an alternative reliable source. PRF is a technique by which the top N documents retrieved for a given query, are assumed to be relevant. The set of top N retrieved documents is thus named the "Pseudo relevant set", and is used by AQE approaches as a source of prospective terms to expand an initial textual query. Consequently, PRF represents a lightweight and reliable feature source to score and determine the best terms for expansion, for a given initial query. However, the reliance of PRF on top retrieved documents is not bullet-proof and its success depends greatly on the performance in gathering a good pseudo relevant set. Therefore, its can be unstable and a source of topical drift as there are no warranties for the top document to be relevant. Nonetheless it has experienced wide use in microblog retrieval as it has been shown to perform effectively on average by previous work such as Whiting et al. (2011) and Lau et al. (2011).

Automatic Query Expansion in Microblog Retrieval. Numerous participants including the top performing ones in both 2011 (Amati et al., 2011; Li et al., 2011; Metzler and Cai, 2011) and 2012 (Aboulnaga et al., 2012; Han et al., 2012; Kim et al., 2012) TREC Microblog tracks employed AQE methods for the ad-hoc task, reporting significant improvements on retrieval effectiveness. However these approaches often fail to filter unrelated terms to the original query, which ultimately can hinder retrieval effectiveness particularly for those topics performing badly from the start. Alternatively, some approaches utilise external evidence in order to find new terms. The work by Gurini and Gasparetti (2012) successfully used Wikipedia as a source of query expansion terms by finding associations within terms in the articles and those of the original query. A different approach is that proposed by Bandyopadhyay et al. (2012), which devised an approach to query commercial search engines such as Google or Bing, and extracted prospective terms from the generated result list.

Relevance-Based Language Model 3 (RM3) was initially introduced by Lavrenko and Croft (2001) and later became popular in works such as Efron et al. (2014) in which they investigate the temporal cluster hypothesis, i.e. how similar documents appear together in time. The assumption behind the RM3 approach is that expansion terms are as good as the documents they are found within. Thus the relevance of a term is derived directly from the retrieval score assigned to the document holding it, as well as, the document frequency of the term in the pseudo relevant set, as portrayed by the following formulation:

$$RM3(t|R_Q) = \sum_{d \in D} P(d)P(t|d)\prod_{i=1}^{n} P(q_i|d) \tag{6.1}$$

where t is any given term present in the pseudo relevant set R_Q, d is a document in the set of all documents D and q_i is the i^{th} query term. Finally, we use a mixing parameter of 0.5 in line with the work by Efron et al. (2014) to regulate the effect of newly added terms with respect to the original query, as it also produced the best results within our experiments.

6.3 Discounting AQE

As we previously introduced, Pseudo Relevance Feedback (PRF) based AQE approaches rely on the assumption that terms found at the top N retrieved documents (Pseudo relevant set) are the most suitable for expanding queries. One of the most successful

PRF based AQE approaches is RM3 (Lavrenko and Croft, 2001). As we previously introduced RM3 makes a very simple yet elegant assumption: “A term is only as good as the document that holds it”. Thus the computation of RM3 directly relies on the score produced by the baseline retrieval model. However, in microblog ad-hoc retrieval, the scores produced by the retrieval models can be very misleading as discussed in our Chapter 3 and it can be observed by the evaluation results produced for Table 3.2. In this table we can appreciate how the presence of relevant documents decreases substantially as we traverse from cut-off points 5 to 30. Our hypothesis is formalised as:

> **H1.** AQE approaches reliant on the scores produced by retrieval models result on unreliable behaviour in the context of microblog ad-hoc retrieval.

This observation, motivated the modelling of a different term selection mechanism which does not directly rely on the scores produced by a retrieval model, but on other more independent features. To this end, in this work we propose a number of term selections strategies to help in estimating the importance of terms, relative to their position in the initial result set.

Table 6.1 helps to illustrate the situations in which a discounting approach could help. Terms A,B,C, and D appear arbitrarily on each of the documents of the result set in Table 6.1. Hypothetically, AQE approaches not accounting for the importance of documents where terms are found within, could produce the ranking $Rank : A, B, D, C$. We hypothesise that a better term ranking should be $Rank : A, B, C, D$, since C appears almost as many times as D in the ranking, but closer to the top. In some cases, it might even be better to have C rank higher than B as it appears in the first document, which has the highest chance to relate to the topic, thus producing the rank $Rank : A, C, B, D$, even when appearing half the times.

6.3.1 Linear And Logarithmic Discount Functions

Our first instance to model this behaviour, considers a linear discounting function. We exploit the rank of the document itself as a linear function, and use it to decrease the relevance of terms as they are found closer to the bottom of the search result list. This approach can be formalized as follows:

Table 6.1: Pseudo relevant set to illustrate term selection techniques. (* denotes the presence of the term in the document)

	Prospective Expansion Terms			
	A	***B***	***C***	***D***
Doc 1	*		*	
Doc 2	*	*	*	*
Doc 3	*	*		*
Doc 4	*	*		*
Doc 5		*		
...				
Doc N				

$$LinearDiscount(d) = \frac{rank(d)}{maxRank}, \tag{6.2}$$

where $rank(d)$ is the rank of the document d in the pseudo relevant set and $maxRank$ is the total number of documents in the pseudo relevant set.

As an alternative to the above-mentioned model we utilise a logarithmic discounting function. This function provides a more smooth discount than the linear discount, making differences between prospective terms less pronounced with respect to the rank. We formally describe the logarithmic discounting approach as:

$$LogDiscount(t, d) = \frac{1}{1 + \log_b rank(d)}, \tag{6.3}$$

where $\log_b rank(d)$, is the logarithm with base b of the rank position $rank(d)$ of document d. We experimented with different logarithmic bases in the range from 1.1 to 3.0 but no substantial differences were found. Thus we decided to set $b = 2.0$.

In the following Section we implement the discount functions in a Rocchio based AQE approach as well as a combination with collection statistics provided by IDF.

6.3.2 Automatic Query Expansion Approaches

In other to test our hypotheses, we implemented our discounting approaches as part of the well known AQE approach **Rocchio** (Carpineto and Romano, 2012) which has been successfully utilised in the context of microblogs. Rocchio's approach to term weighting for AQE can be formalized as follows:

$$AQE\_ROC(t,R) = \sum_{d \epsilon R} w(t,d), \tag{6.4}$$

where R is the set of pseudo relevant documents, and $w(t,d)$ is the weight of term t within document d given by its in-document frequency. To produce the discounting versions of Rocchio, we combine it with each of the above mentioned discounting functions as follows:

$$AQE\_ROC\_Log(t,R) = \sum_{d \epsilon R} w(t,d) * \frac{1}{1 + \log_b rank(d)}, \tag{6.5}$$

$$AQE\_ROC\_Lin(t,R) = \sum_{d \epsilon R} w(t,d) * \frac{rank(d)}{maxRank}, \tag{6.6}$$

Furthermore we derived a similar approach, namely **AQE_IDF**, which utilizes IDF values instead for producing the initial term weights. These term scores are then modified by applying any of the discounting formulae to it. The AQE_IDF approach is formalized as follows:

$$AQE\_IDF(t,R) = \sum_{d \epsilon R} Idf(t), \tag{6.7}$$

where $Idf(t)$ is the Idf score for term t. The main difference with respect to Rocchio's algorithm is that the score produced takes into consideration Idf as a source of collection-based evidence. The addition of the discounting component is formalised as follows:

$$AQE\_IDF\_Log(t,R) = \sum_{d \epsilon R} Idf(t) * \frac{1}{1 + \log_b rank(d)}, \tag{6.8}$$

$$AQE\_IDF\_Lin(t,R) = \sum_{d \epsilon R} Idf(t) * \frac{rank(d)}{maxRank}, \tag{6.9}$$

The following Section introduces the experimental setting which will drive the evaluation and the prospective conclusions.

6.4 Experimental setting

Baseline systems. A number of state of the art retrieval models were evaluated, and their results are presented in Table 3.2. When deciding on a baseline for PRF-based AQE, it is vital to consider the best and more consistent performance at upper ranks. As we can observe DFR fits the description providing the best performance in terms of P@10 in 2 out of 3 collections, thus we selected it as the baseline for our AQE experiments

Parameters. The parameters for the retrieval models used in this work have been set accordingly to their recommended implementation within the Terrier IR platform. [1]

In order to find the best configuration for the AQE approaches we reserved the first 90 topics for training (Out of 225 topics in total). The parameters to optimise are the number of terms to add to the original query, the number of documents to consider in the pseudo relevant set, as well as the initial retrieval model. We studied the behaviour of these AQE approaches with respect to the three best performing retrieval models, namely DFR[2], IDF[3], and BM25.

The training results for all considered configurations at this stage are included in Appendix B. We chose the best performing configuration for each pair of AQE approach and retrieval model to produce the experimental results on the remaining 135 topics which will be discussed on the following Section.

6.5 Results and Discussion

The rest of this Section studies the effects of the different discounting functions introduced in Section 6.3 when applied to PRF-based AQE compared against the state of the art RM3.

In Chapter 3 we showed how the features inherent to microblogs negatively affected the performance of state of the art retrieval models in producing representative scores

[1] http://www.terrier.org/docs/v3.5

[2] DFRee: DFR free of parameters as implemented in http://www.terrier.org/docs/v3.5

[3] TFIDF where TF=1 to reduce adverse effects of small TF variations

Table 6.2: Evaluation of AQE approaches on test set (164 topics) using IDF as baseline. (Significance denoted by * ($p < 0.05$) w.r.t. baseline.)

	Max Terms	Max Docs	P@5	P@10	P@15	P@20	P@30	MAP
IDF			0.661	0.622	0.573	0.546	0.496	0.296
RM3	1	20	**0.664**	**0.634**	**0.600***	0.569*	0.521*	0.317*
AQE_IDF	1	30	0.643	0.617	0.587	0.562	0.516	0.315*
AQE_IDF_Lin	1	10	0.655	0.629	0.596*	**0.571***	**0.525***	0.321*
AQE_IDF_Log	3	20	0.630	0.619	0.599*	**0.571***	**0.525***	**0.325***
AQE_ROC	1	30	0.642	0.616	0.588	0.562	0.516	0.315*
AQE_ROC_Lin	1	30	0.645	0.617	0.588	0.563	0.516	0.315*
AQE_ROC_Log	1	30	0.643	0.619	0.588	0.563	0.516	0.315*

Table 6.3: Evaluation of AQE approaches on test set (164 topics) using DFR as baseline. (Significance denoted by * ($p < 0.05$) w.r.t. baseline.)

	Max Terms	Max Docs	P@5	P@10	P@15	P@20	P@30	MAP
DFR			**0.658**	0.601	0.559	0.532	0.487	0.292
RM3	1	30	0.615*	0.593	0.572	0.550	0.501	0.309
AQE_IDF	1	30	0.615*	0.592	0.572	0.550	0.501	0.308
AQE_IDF_Lin	1	30	0.618	0.595	0.572	0.551	0.502	0.309
AQE_IDF_Log	1	10	0.633	0.604	0.573	0.544	0.499	0.312*
AQE_ROC	1	30	0.613*	0.592	0.571	0.550	0.501	0.308
AQE_ROC_Lin	1	10	0.627	0.600	0.573	0.544	0.500	0.310*
AQE_ROC_Log	3	20	0.613*	**0.605**	**0.588**	**0.554**	**0.514***	**0.317***

when ranking. As a result, this observation led to our main hypothesis which suggests that the scores produced by retrieval models can be misleading when utilised by Automatic Query Expansion approaches like RM3 when considering new expansion terms. Consequently, in order to test our hypothesis, we developed a set of AQE approaches that rely on other features than the score produced by the retrieval model, and we compared it to the state of the art RM3 approach. The main evaluation results to allow this comparison are presented in Tables 6.2, 6.3 and 6.4 which include results obtained by all considered AQE approaches with respect to the DFR, IDF and BM25 baselines respectively.

Table 6.2 holds the results for all our experiments utilising IDF to produce the pseudo relevant set. The results are presented in terms of precision at different cut-offs

Table 6.4: Evaluation of AQE approaches on test set (164 topics) using BM25 as baseline. (Significance denoted by * ($p < 0.05$) w.r.t. baseline.)

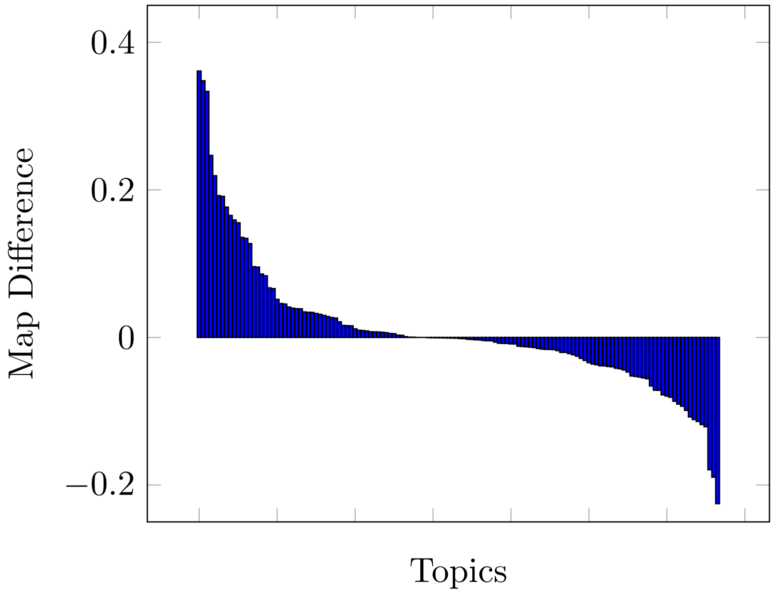
	Max Terms	Max Docs	P@5	P@10	P@15	P@20	P@30	MAP
BM25			**0.582**	0.522	0.485	0.460	0.419	0.235
RM3	3	30	0.573	0.528	0.516	0.491	0.451*	0.273*
AQE_IDF	5	30	0.567	0.537	0.513	0.489	0.447*	**0.274***
AQE_IDF_Lin	3	30	0.569	0.532	**0.519***	0.493*	0.454*	0.273*
AQE_IDF_Log	5	30	0.569	**0.542**	0.517	0.492*	0.445*	**0.274***
AQE_ROC	3	30	0.576	0.537	**0.519***	**0.494***	**0.455***	0.273*
AQE_ROC_Log	3	30	0.572	0.534	0.513	0.490	0.452*	0.273*
AQE_ROC_Lin	5	20	0.569	0.537	0.517*	0.492*	0.442	0.269*

as well as MAP. We can observe in row two how RM3 does not behave statistically different from the baseline at P@5 and P@10. However as we traverse higher cut-off points, RM3 starts to produce statistically significantly improved results w.r.t the IDF baseline. A similar pattern can be observed for AQE_IDF_Lin and AQE_IDF_Log which achieve very performances to RM3 on average for these metrics. Interestingly, AQE_Lin and AQE_Log exhibit almost the same behaviour, thus showing that the different normalisations used do not have a significant effect in this particular case. Furthermore, the approaches based on the Rocchio method (AQE_ROC; AQE_ROC_Lin and AQE_ROC_Log) only achieve statistically significant results for this baseline at much lower ranks as shown by the MAP evaluation metric. In this particular case we can extract that the most successful AQE approaches include RM3, AQE_Lin, AQE_Log.

The next Table 6.3 holds similar results but utilising a DFR baseline. The average performance of DFR is very similar to that achieved by the IDF baseline in the previous Table 6.2. However, the AQE approaches exhibit an entirely different behaviour. Firstly RM3 does not achieve any significantly better results than the DFR baseline. On the other hand it significantly worsens the results for the early document ranks as shown by the P@5 evaluation metric. In fact, this also happens to three of our proposed AQE approaches, namely AQE_IDF; AQE_ROC and AQE_ROC_Log. However, the results obtained by our approaches AQE_IDF_Log; AQE_ROC_Lin and AQE_ROC_Log do show significant improvements at later ranks as exhibited by the MAP metric, as well as, P@30 in the case of AQE_ROC_Log.

We can extract a number of important observations from this table. The RM3

Figure 6.2: Per topic MAP difference between RM3 and AQE_IDF_Log with an **IDF** baseline

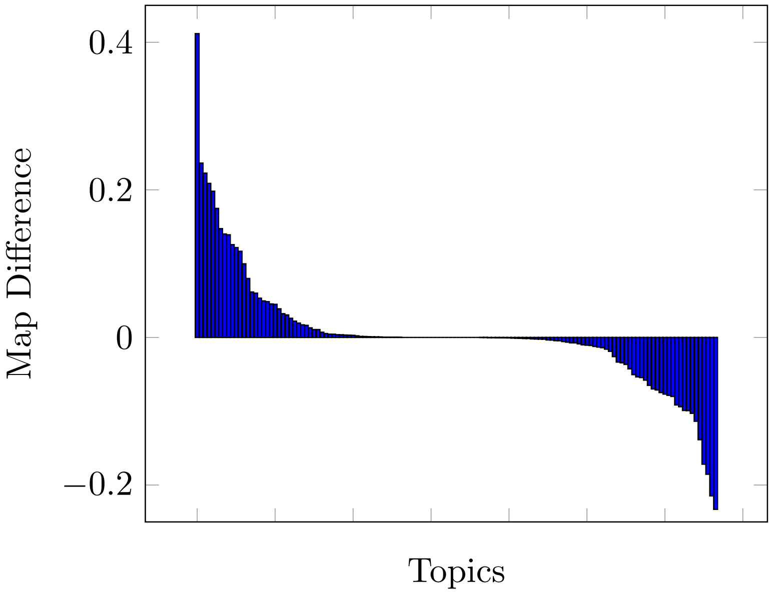


methods does not obtain any significant improvements over the baseline, whereas some of our methods achieved significantly better results in terms of MAP, without the expense of significantly worsened results measured by P@5 and exhibited by RM3.

Finally we study Table 6.4 which presents results in terms of a BM25 baseline. This table shows the behaviour of the considered AQE approaches when operating on a significantly weaker baseline. The first observation we can extract is that there is a higher number of cases in which the AQE approaches achieved significantly better results than the baseline. This shows how AQE approaches in general, can help to better capture or steer the initial results into the intended relevant document set. Looking at the results obtained by RM3 we can confirm how it only achieved significantly better results for MAP and P@30. On the other hand, our approaches obtained more consistently better results at different cut-off points as well as MAP. Three cases deserve special attention, namely AQE_IDF_Lin and AQE_ROC which achieved significantly better performance in terms of P@15; P@20; P@30 and MAP. This is followed by AQE_IDF_Log which achieved significantly better results for P@20; P@30 and MAP and AQE_ROC_Lin with better results at P@15; P@20 and MAP.

As a conclusion from these results, we can confirm that a approaches such as AQE_IDF_Log which rely on a mixture of IDF and a normalisation method w.r.t. the rank

Figure 6.3: Per topic MAP difference between RM3 and AQE_IDF_Log with an **DFR** baseline



in which terms are found, can produce similar result in average, however in a more consistent manner thus achieving significantly improved results in more scenarios than RM3.

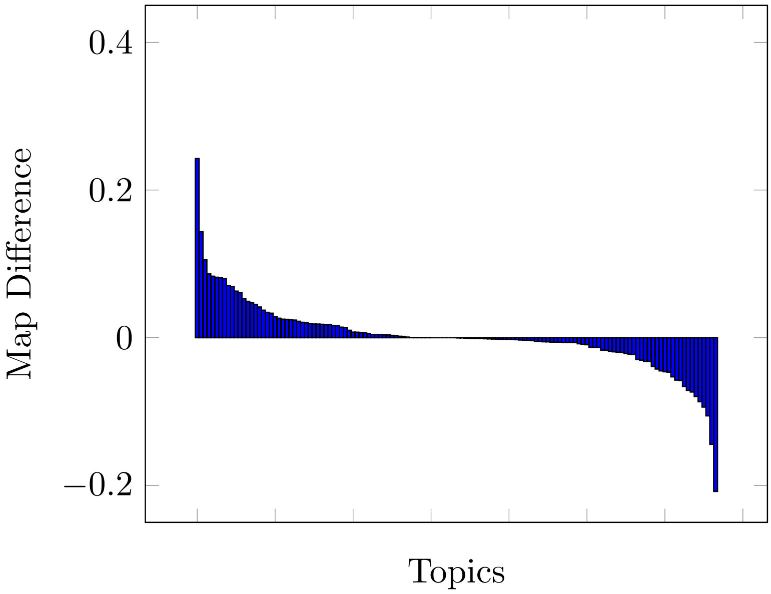
Additionally we compiled Figures 6.2, 6.3 and 6.4 in which we can observe the differences in performance between our best performing approach AQE_IDF_Log and RM3 in the context of the different baselines. We utilised MAP as the evaluation metric and subtracted the MAP value achieved for each topic utilising RM3 to that of AQE_IDF_Log. Therefore, the values above zero indicate the topics in which AQE_IDF_Log performs better than RM3 and vice-versa.

On the three figures, we can observe a large area in the middle with practically unaffected topics. There are a number of possible reasons behind this behaviour:

A. The documents returned already represent very closely the topic, and the addition of new terms has not added any information that was not previously considered. (Easy topics)

B. The initial query is already quite distant from the relevant topic, thus returned documents in the pseudo relevant set are not related to the topic, and conse-

Figure 6.4: Per topic MAP difference between RM3 and AQE_IDF_Log with an **BM25** baseline



quently any expansion terms do not add any useful value to improve the search. (Difficult topics)

C. Both AQE approaches have selected the same set of terms for expansion, thus producing the same results.

D. Both AQE approaches have selected different terms, however they are similarly related to the relevant documents retrieved achieving similar results.

In order to further examine this results we have compiled Figure 6.5. We include the topics from Figure 6.2 where the MAP difference between RM3 and AQE_IDF_Log is lower than 0.01, in order to account for all topics where both AQE approaches achieve similar results. This Figure shows the differences between RM3 and AQE_IDF_Log with respect to the IDF baseline in terms of MAP, as well as the initial MAP value achieved by the IDF baseline itself. Results are ordered in increasing MAP value of the IDF baseline (Used to generate the pseudo relevant set). As we can observe, there are many cases in which RM3 and AQE_IDF_Log do have an effect, however they receive a similar MAP score in line with **C** and **D**. Moreover, we can also observe how the effects on the left half of the figure are a lot less prominent than that of the right. This is due

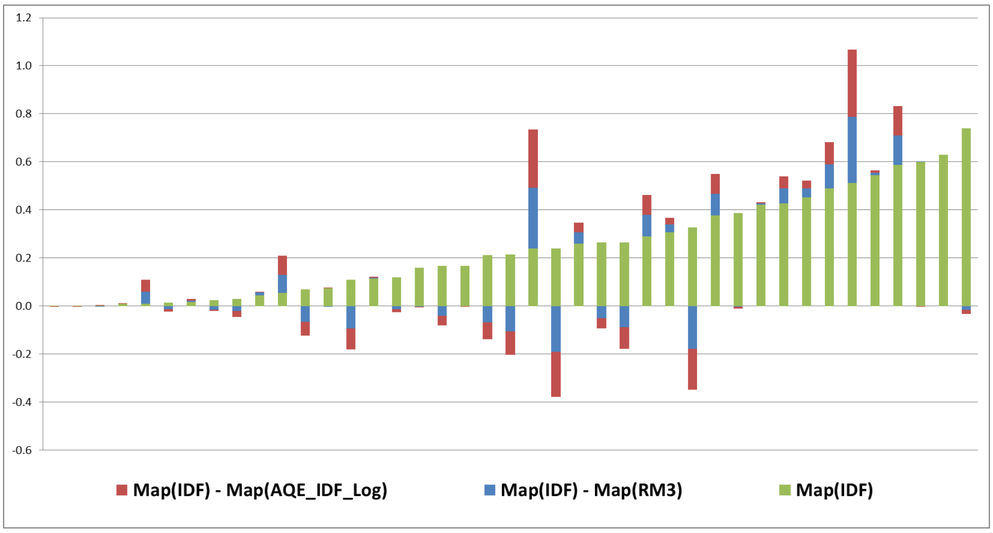


Figure 6.5: MAP difference of RM3 and AQE_IDF_Log w.r.t. initial MAP obtained by the IDF Baseline

to the pseudo relevant set either properly covering the topic, thus the AQE approach has no effect (**A**), or the initial result is not representative of the topic thus the AQE approach cannot find appropriate terms for expansion (**B**).

Moreover, we can appreciate how the area occupied in Figure 6.2 by the metrics for both AQE approaches is larger than in Figures 6.3 and 6.4. This can be linked to the performance of the baselines. IDF is the best performing baseline, and it is very closely followed by DFR, whereas BM25 is significantly worse. Therefore we can observe how a better starting point noticeably benefits both systems. More important yet are the differences between both AQE approaches observed in these figures. We can clearly see how each AQE approach affects particular sets of topics differently. The split for which each of the approaches is beneficial and detrimental with respect to the other is very close in magnitude, thus explaining why they are producing similar averages in the results exposed in Tables 6.2, 6.3 and 6.4 whilst behaving very differently.

Summarising, our hypothesis suggests that the scores produced by retrieval models can be misleading in determining whether a term is optimal to be used for query expansion, in the context of microblog ad-hoc retrieval. As a representative of this methodology we experimented with RM3 and compared it against our own methods which do not rely directly on the retrieval model scores to perform such computations. Our experimental results confirm the superiority of our approaches by demonstrating improved consistency in achieving significantly better results independently from the baseline utilised. Both methodologies often achieve comparable results in terms of average performance measured by precision and MAP. However each approach clearly affects a particular set of topics differently than the other as seen in Figures 6.2, 6.3 and 6.4. These results present an opportunity for future researchers to determine what are the differentiating features for each set of topics, which can lead to the prospective application of selective AQE methodologies. Finally these conclusions motivated the following Section in which we investigate features that make terms most effective in the context of AQE.

6.6 Predicting Term quality for optimised AQE

As we observed in the previous Section, most AQE approaches experience high variability in retrieval performance across many different queries. Whilst many queries are expanded successfully, the system often produces poor results for many others. The reason behind this variability is that there is no clear and effective way estimate which terms are more linked to a particular topic, thus more likely to produce better results.

Selective Query Expansion. In Chapter 5 we introduced the Query Performance Prediction (QPP) task. The aim of QPP approaches is to attempt to measure the level of success a system will have in retrieving the appropriate documents for a given query, without the certainty provided by relevance judgements. One of the practical applications of QPP is selective query expansion (Carmel and Yom-Tov, 2010). Selective query expansion attempts to alleviate the above-mentioned cases where applying AQE to a particular set of queries leads to worsened results.

There are number of works in which selective AQE has been applied. In the work by Amati et al. (2004) they selectively applied automatic query expansion (AQE) whenever predicted performance is above a certain threshold. This serves as a warranty for PRF-based AQE approaches, as they rely on the top N retrieved documents to perform optimally thus achieving a significantly more robust system. Moreover, Yom-Tov et al. (2005) trained a classifier to identify queries for which PRF should produce satisfactory results. The classifier was trained on a dataset where queries were annotated as being successfully expanded or not. Additionally work by Cronen-Townsend et al. (2006) introduced an approach by which they compared the language model of the initial retrieved set against the result set retrieved issuing the expanded query. If the expanded retrieved set language model was too far from the initial set, it indicated that the query had drifted too much thus being interpreted as a worsened result, and discarded in favour of the initial result. Finally, the work by He and Ounis (2007) combined a metasearch engine with selective query expansion to provide a selection mechanism. Such mechanism was in charge of selecting a document source from a number of collections when producing the search results to respond a given query.

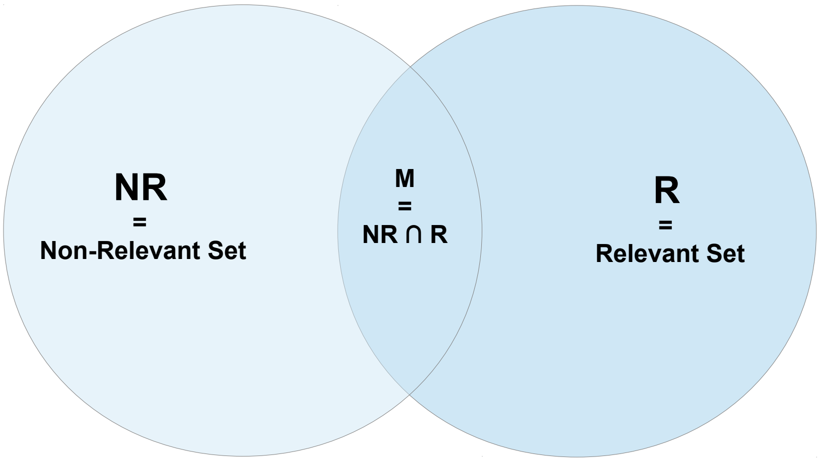
In our study aligns with the work by Yom-Tov et al. (2005), as we will be using features to build a classifier, that in turn will predict the suitability of a term for query expansion. Also, this task has many points in common with selective query expansion, as we utilise predictors to train the classifier which in turn will be used to improve the behaviour of an automatic query expansion stage.

Classification of good PRF terms. The most related work to ours was presented by Cao et al. (2008) in the context of Web retrieval. In their work, they employed an SVM classifier which attempted to predict whether a retrieval system would perform better when a particular term was included. Their features included term proximity and co-ocurrences, together with term distributions and document frequencies. They demonstrated the feasibility of this classification through their significantly improved retrieval performance, thus served as a motivation to attempt a similar classification effort and assess its suitability for query expansion in the context of Microblog documents.

In this Section, we propose a classification methodology to discriminate those terms that are most beneficial for expansion, from those that may produce topical drift. To this end we define a set of features appropriate to capture the differences in the quality of terms with regards to their suitability for query expansion. Our features are mainly based on the inverse document frequency (IDF) values of terms and the relative differences with respect to the values of other co-occurring terms. In order to drive the rest of the study we pose the following research questions:

RQ1. Is it possible to infer the quality of a term, and its suitability to expand an initial query using an AQE approach in the context of microblog ad-hoc retrieval?

Figure 6.6: Document set diagrams. Classes are extracted from here.



RQ1.1. Are there significant differences between terms in known relevant documents and those in non-relevant documents?

RQ1.2. To what extent is it possible to classify terms as being optimal to be used for query expansion based on a given set of features?

RQ1.3. Can we employ a classifier to improve the performance of state of the art automatic query expansion approaches?

6.7 Approach

In this work, we hypothesise the existence of significant differences between those terms appearing only in relevant documents, non-relevant documents, and those appearing in both groups. As part of our approach, we attempt to characterise such differences and leverage them for improving the behaviour of AQE approaches. In this Section we first define the features we explored, then we build a classifier based on them and finally introduce the implementation of an AQE approach to utilise such classifier.

6.7.1 Identifying classes

The first step in our work is the formal definition of the type or classes of terms we want to distinguish. Figure 6.6 introduces a representation of the sets of interest considered in this study. The Non-relevant set NR at the left contains all non-relevant documents we have knowledge of in any set of relevant judgements. Moreover, on the right hand side of Figure 6.6 we can find the Relevant set which contains all known relevant documents. We then define three classes of interest: The "Low-Value" class which contains all those terms found exclusively in terms belonging to the NR set; The "High-Value" class which holds only those terms appearing exclusively in known relevant documents. Finally, the "Medium-Value" set, which contains those terms that appear in both relevant and non-relevant documents. We can express the classes formally as:

$$\text{Low-Value} = \{t|t \in NR \wedge t \notin R\} \tag{6.10}$$

$$\text{High-Value} = \{t|t \in R \wedge t \notin NR\} \tag{6.11}$$

$$\text{Medium-Value} = \{t|t \in NR \wedge t \in R\} \tag{6.12}$$

where t is any given term. Now that we have a formal definition of what we want to characterise, we move on to defining features to capture their class differences.

6.7.2 Describing Features

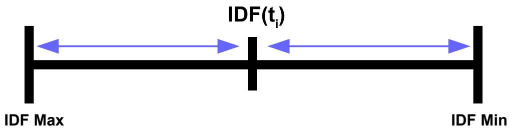
In this work, we rely on IDF-based features to establish the differences between terms. We believe that the difference between the IDF values of terms can be leveraged to evaluate their membership to the above-mentioned classes.

Absolute Features. Firstly we define **idfScore** to be the average IDF score of the terms belonging to a given class. Query and document terms are included in this feature.

idf_query_max and **idf_all_max** are features which capture the average maximum IDF for each of the classes. Our intuition is that terms appearing in documents or queries where the maximums are higher, they have a higher chance to be of importance. We devised two features, **idf_all_max** which computes the maximum IDF out of all terms in the documents, and **idf_query_max** which considers only query terms.

Similarly we defined **idf_query_min** and **idf_all_min** to hold the average minimum IDF values for query terms only and all terms respectively. Next, we introduce **idf_query_mean** and **idf_all_mean**, which follows the same idea as previous features, just that this time we capture the mean IDF values for query terms, or all terms in the documents respectively.

Figure 6.7: Visual representation of a scale of IDF values. Most of our features are defined as distance measures between the different thresholds.



Relative Features. Figure 6.7 introduces a graphical representation of an IDF scale. **IDF Max** represents the maximum value of IDF within a query or a document. **IDF Min** on the other hand refers to the minimum value of IDF within a query or a document. The features in this subsection exploit the relative differences or distances in Figure 6.7 expressed as blue arrows between the maximum and minimum IDF values.

First features in the "relative" family are **(idf-minQuery)**, **(idf-minAll)** and **(idf-minNQT)**. These features measure the absolute difference, or distance between the IDF value of a given term and the minimum IDF value, whether we consider only query terms (idf-minQuery), non-query terms (idf-minNQT), or all terms (idf-minAll). The rationale behind these features is that terms discriminative of similar topics should hold similar IDF values, thus lower IDF could be a reflection of the unrelatedness of a term with respect to others. I.e. terms have less in common with the particular topic, even if the IDF value is relatively high in comparison to other terms. The main advantage of computing relative measures, is that we may be able to obtain a contextualised measure of the importance of terms, when comparing to other terms within the document or query terms.

Similarly to the last set of features, (maxQuery-idf), (maxNQT-idf) and (maxAll-idf) measure the distance of the IDF value of term from the maximum IDF value, when we consider only query terms, non-query terms, or all terms respectively. The motivation behind these features is the opposite to the last set. The closer terms are to the maximum in terms of IDF values, the closer the match they should be with respect to the topic. Sometimes, terms are found that contain an even higher IDF value than query terms, thus it is interesting to take into consideration measures that account for such terms and others that do not.

Other features. The final set of features contain **coveredQueryTerms**, **doclength_chars** and **doclength_terms**. The **coveredQueryTerms** (Rodriguez Perez and Jose, 2014) feature measures how well the query is being represented by the documents in the result list. CoveredQueryTerms is defined as:

$$coveredQueryTerms(q_i, D) = \frac{\sum_{j=1}^{|D|} cov(q_i, d_j)}{|Q|}, \tag{6.13}$$

where the rate of query terms in Q appearing on each document is aggregated and normalized between 1 and 0. (1 means that all query terms are present in the document). Furthermore $cov(q_i, d_j)$ is defined as:

$$cov(q_i, d_j) = \begin{cases} 1, & \text{if } tf(q_i, d_j) \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{6.14}$$

where $cov(q_i, d_j)$ returns 1 whenever the term frequency $tf(q_i, d_j)$ of term q_i in document d_j is higher than 1, and 0 otherwise.

The final two features **doclength_chars** and **doclength_terms** refer to the length of the document in which the words are found, measured by the number of characters and number of terms respectively. For more information about the features in this work refer to Table 6.5.

6.7.3 Term Quality Prediction

The next step in our work is to build a classifier that takes into consideration above features to determine the value of a term. The objective of the classifier is to perform Term Quality Prediction (TQP), which involves determining the membership of a term to one of the groups above-mentioned.

Since the number of samples per class in our test collection was very unbalanced (I.e. the number of terms belonging to the "Low-Value" class was much greater than for the "Medium-Value" and "High-value" classes), we applied a filter to our training data named Smote. Smote (Chawla et al., 2002) re-samples a dataset by applying the Synthetic Minority Oversampling TEchnique (SMOTE), which in simple terms generates synthetic data for the under-represented classes, which in turn helps produce more accurate classification models.

Our choice of classifier was the Weka (Hall et al., 2009) implementation of the J48 Decision Tree which provided us with reasonably good results as we will introduce in Section 6.9. While we did not explore further the choice of classifier, it is worth exploring in the future as other classifiers may yield better classification performance. Furthermore, we performed feature selection before building the classifier by means of the *BestFirst* method together with the *ClassifierSubSetEval* evaluator implemented in Weka. The feature selection step showed that the (maxNQT-idf) and (idf-minNQT) features were not helpful for the J48 classifier, thus were not considered in our final feature choice.

The final classifier was built using a 10-fold cross validation over the 2013 TREC Microblog collection. This collection was chosen as it was sampled much deeper than previous collection, thus making it a better source of learning data.

6.7.4 Classifying Terms for PRF-AQE

The Automatic Query Expansion RM3 approach (Lavrenko and Croft, 2001) based on PRF has been proven to produce significantly improved results when expanding queries under microblog retrieval constrains (Efron et al., 2014). The assumption behind the RM3 approach is that expansion terms are as good as the documents they are found within. Thus terms found within a document are more important than another terms, if they are found within a document with a higher retrieval score. However, this is not often the case and terms that do not hold relationship with the topic at hand are often selected, over more related terms. Thus including further knowledge in order to distinguish between the different terms could be highly beneficial.

To this end, combining the knowledge provided by our classifier with the scores produced by RM3 seems like a good opportunity to validate the usefulness of our classifier as well as devising an alternative version of RM3 tailored to microblog retrieval.

Furthermore, we define the $Class(t, \vec{F})$ as a wrapper function to hold the classifier. In this case the $Class(t, \vec{F})$ function makes use of the J48 Classifier to estimate the class of t as follows:

$$Class(t, \vec{F}) = J48(t, \vec{F}) \tag{6.15}$$

where $\vec{F}$ is the feature vector for term t. The value returned by the $Class(t, \vec{F})$ function is the predicted class of term t. In our case it will return 0,1 or 2 for "Low-Value", "Medium-Value" or "High-Value" respectively. Then we define a function boost to act as a boosting parameter selection depending on the predicted class for term t.

$$boost(t, \vec{F}) = \begin{cases} 1, & \text{if } Class(t, \vec{F}) = 0 \\ 1.25, & \text{if } Class(t, \vec{F}) = 1 \\ 1.50, & \text{if } Class(t, \vec{F}) = 2 \end{cases} \tag{6.16}$$

The final RM3_TQP [1] approach utilises the boosting parameter to increase the importance of terms that are predicted to belong to the "Medium-Value" and "High-Value" classes and it is formalised as follows.

$$RM3\_TQP(t, \vec{F}, R_Q) = RM3(t|R_Q) * boost(t, \vec{F}) \tag{6.17}$$

where the score produced by $RM3(t|R_Q)$ is simply multiplied by a boosting factor given by the $boost(t, \vec{F})$ function. The boosting parameters are 1, 1.25 and 1.5 as shown in Equation 6.16. Whilst these parameters are heuristically selected and enough for the purposes of this work, further exploration should lead to more appropriate values which could enhance the performance.

6.8 Evaluation Settings

In this section we introduce the details of our evaluation methodology including specifications of the datasets and evaluation metrics for each task.

Datasets. In this evaluation we utilize the Microblog 2011, 2012 and 2013 TREC collections totalling 170 topics. We reserve the topics from the 2013 collection (60 topics) for training our classifier, thus leaving the 2011 and 2012 collections for testing our

[1] RM3 with Term Quality Prediction (TQP)

AQE approach.

Retrieval Model Used. Table 3.2 shows evaluation metrics for a number of retrieval models including DFRee, DLM and HLM which stand for Divergence From Randomness free of parameters, Dirichlet Language Model and Hiemstra's Language Model respectively[1]. We justify the use of DFRee by Amati and Van Rijsbergen (2002) as our baseline as it provides the best results across Microblog the 2011 and 2012 collections (Table 3.2) so as to provide the best baseline for a PRF-based AQE approach. Furthermore, it is free from parameters which helps in simplifying our experimental procedure and its interpretation.

Classification Evaluation metrics. In order to understand the behaviour of our classifier, we consider Precision and Recall measures for each of the classes being predicted as well as metrics mixing precision and recall such as F1 measure and ROC-Area.

Run Evaluation metrics. Since we are exploring the extent to which we can improve the retrieval performance of our approach, we consider Precision at cutoffs from 5 to 100 as well as Map@30 to obtain a broad view of our results.

6.9 Results and Discussion

In this section we introduce and discuss our experimental results. First we observe the behaviour of the features individually and evaluate their suitability to discern between terms that are most likely to belong to relevant than to non-relevant documents. Secondly, we examine the results and the capabilities of the supervised classifier we built for Term Quality Prediction purposes utilising the J48 decision tree. Finally, we discuss the results obtained when combining our classifier with the state of the art AQE approach **RM3** to enhance its behaviour.

6.9.1 Feature Analysis

Table 6.5 shows mean values for all features we have defined in Section 5.3 with respect to the classes we have defined. The First column contains the values for "High-Value" terms, whereas the second and third column contain the values for the "Medium-Value"

[1] http://terrier.org

Mean values for the term features per quality group.			
Feature	**High Quality**	**Medium Quality**	**No Quality**
coveredQueryTerms	2.09	2.13	1.69* †
doclength_chars	59.21	56.92 *	58.29 * †
doclength_terms	10.70	10.68	10.58 * †
idfScore	0.75	0.50 *	0.71 * †
idf_query_max	0.54	0.54	0.55 * †
idf_query_mean	0.51	0.51	0.53 * †
idf_query_min	0.48	0.48	0.51 * †
idf_all_max	0.84	0.73 *	0.84 †
idf_all_mean	0.51	0.48 *	0.52 †
idf_all_min	0.31	0.31	0.31
Distance Features			
(idf-minQuery) [only query terms]	0.28	0.12 *	0.24 * †
(maxQuery-idf) [only query terms]	0.23	0.12 *	0.21 * †
(idf-minAll)	0.43	0.19 *	0.40 * †
(maxAll-idf)	0.09	0.24 *	0.13 * †
(idf-minNQT) [no query terms]	0.44	0.19 *	0.40 * †
(maxNQT-idf) [no query terms]	0.10	0.24 *	0.14 * †

Table 6.5: Feature mean values from the 2013 TREC Microblog collection. (* $p < 0.05$ w.r.t High Quality; † $p < 0.05$ w.r.t Medium Quality)

and "Low-Value" terms respectively. The $*$ symbol denotes statistically significant differences with respect to the "High-Value" class, whereas the † symbol denotes statistically significant differences with respect to the "Medium-Value" group.

The first observation drawn from the results is that almost all features have significant differences with respect to at least one of the groups. Most importantly the "Low-Value" group is almost always statistically significantly different from the High and Medium-Value groups. This finding is very promising, as it represents the first indication of the possibility to discriminate terms which appear only on non-relevant documents. However, the "idf_all_min" is the only feature that does not produce any significant differences between any of the classes defined as its mean remains stable across all classes.

Moreover, we observed no differences between the high and medium value groups in terms of the features "idf_query_max", "idf_query_mean" and "idf_query_min". However, the "Low-Value" group is significantly higher for the three features than the other classes. These features take into consideration only the idf values of the query terms found in documents. Therefore, it is a surprising result to see that the values for the "Low-Value" group are significantly higher, as this translates to the "Low-Value" group having a generally higher discriminatory power, than the other two groups. On the other hand, the decreased IDF mean values for “High-Value” and “Medium-Value” may be due to the inclusion of more query terms that have lower IDF values.

On the contrary, if we take all terms in the tweets into account, (I.e. not only query terms) we obtain the results for the “idf_all_max” and “idf_all_mean” features. In this case, we can observe a significant difference between the values for the “Medium-Value” class with respect to the other two classes. As we can see IDF values for “High-Value” terms, are generally as high as those for “No-Quality” ones. This may be due to terms in the “Low-Value” class belonging to other unknown topics for which they may be quite discriminative of, thus having similar characteristics to the “High-Value” ones.

As previously introduced, we believe that the difference or distance between IDF values of terms from the maximum and minimum values can be useful towards discriminating the usefulness of a term for automatic query expansion. Thus, to evaluate this idea we observe the values of the “Distance features” and their significant differences between the defined classes.

Consequently, our attention was brought to the features measuring the distance from the minimum (idf-minQuery), (idf-minAll) and (idf-minNQT). We observed that the values are substantially and significantly greater for those terms in the “High-Value’ set than those in the other sets. Additionally, the “Medium-Value” metrics were closer to the minimum, suggesting that they may be less discriminative of a single topic than both the “High-Value” and the “Low-Value” classes.

Finally, looking at the distances from the maximum an interesting behaviour emerges. If we look at the (maxQuery-idf) feature which only considers query terms, the distance from the maximum IDF value in the query is greater for the “High-Value” and “Low-Value” groups. However, the opposite happens when we consider all terms in the tweets (maxAll-idf), or only those terms that are not in the query (maxNQT-idf). The conclusion of these results are that some non-query terms contained in the retrieved documents are often more discriminative than the query terms themselves.

We believe that the above-mentioned statistically significant differences between the classes with respect to the features defined in this work, may be sufficient to enable the classification of terms in the defined categories.

Class	Precision	Recall	F1	ROC-area
Low-Value	0.808	0.818	0.813	0.825
Medium-Value	0.582	0.578	0.58	0.787
High-Value	0.672	0.657	0.665	0.807
Weighted Avg.	0.731	0.732	0.731	0.813

Table 6.6: Classification results for a J48 decision tree on the 2013 Trec Microblog collection.

6.9.2 Classification quality results

In this subsection we discuss the classification of terms as belonging to each of the defined classes. Table 6.6 shows the results of the classification runs by the J48 decision tree implemented in Weka. Each of the rows contain the evaluation values for each of the classes. On the other hand each column contains the values for each evaluation metric.

Firstly, we can observe that the values for all measures when predicting the "Low-Value" class are quite high. We believe that it is due to the predominance of this class in terms of the number of representative samples in the test set, even though we have reduced the numerical differences amongst classes by re-sampling using the SMOTE filter.

Furthermore, the values in terms of precision and recall for the "Medium-Value" class are on the mid-range, suggesting that it is the hardest class to classify with respect to our features. The most probable reason as to why this class is the most difficult one, is the fact that it lies in a grey zone, as terms appear often in both relevant and non-relevant documents. However in terms of the ROC-area it looks that it may be strong enough to classify correctly at least more than half of the terms of this class.

The "High-Value" class on the other hand shows substantially better results than the "medium-value" class. The precision for this class is particularly good considering it is substantially less represented than the "Low-Value" class, reaching a 67%. Recall is also quite high, leading to good results as well in terms of ROC-Area and F-measure, suggesting a good balance between precision and recall.

Whilst the classification of terms seems promising within the 2013 microblog collection, further experimentation is required with other collections. To this end we performed a final validation of our classification capabilities in a practical setting by combining our classifier with a state of the art AQE approach.

6.9.3 Enhanced AQE with TQP results

Whilst the classification of terms as beloging to relevant documents is an interesting tool by itself, we further experimented with enhancing the behaviour of a state of the art AQE approach, such as RM3. Table 6.7 contains the evaluation results for the runs produced using the DFRee baseline. Another baseline was produced using RM3 over DFRee, which provides a point of comparison with an AQE approach. Finally, the RM3_TQP is the modified version of RM3 which combines the AQE approach with the classifier, through the use of a boosting factor which depends on the predicted class of a term. All the AQE runs take into consideration the top 10 terms within the top 30 documents.

The table is subdivided into two sections. The first section contains the results for all systems over the 2011 collection whereas the second section contains results for the 2012 counterpart. Notice that we did not use the 2013 collection to produce runs, as we employed it exclusively to train the classifier thus avoiding any biases in the model and our evaluation.

Looking at the differences between RM3 and the DFRee baselines, we can see how RM3 performs worse in terms of early precision (@5 and @10) in the 2011 collection. However small differences are hardly significant at these stages, as very little documents are considered and randomness plays a big role. For the 2012 collection on the other hand, RM3 does perform considerably better over the early precision measures reporting statistically significantly improved results for P@10 and P@15.

Later precision values (@15 to @100) demonstrate consistently better performance for RM3 with respect to the DFRee baseline, with the exception of P@100 for the 2012 collection where they behave similarly. In terms of MAP@30, RM3 performs worse in the 2011 collection, possibly due to the weight of the early precision values. For the 2012 collection RM3 obtained the same score in terms of MAP@30 than the DFRee baseline.

As a summary, RM3 does generally improve performance over the DFRee baseline specially for cut-offs higher than 10, but there is still much room for improvement. Now we compare the results obtained by RM3_TQP which combines our Term Quality

	Evaluation metrics for the AQE Runs						
Run	**P@5**	**P@10**	**P@15**	**P@20**	**P@30**	**P@100**	**MAP@30**
2011 collection							
DFRee	**0.60**	**0.57**	0.52	0.49	0.43	0.25	**0.24**
RM3	0.57	0.55	**0.53**	**0.50**	0.45	0.27*	0.22
RM3_TQP	0.58	0.55	**0.53**	**0.50**	**0.47*†**	**0.28*†**	0.23
2012 collection							
DFRee	0.46	0.42	0.39	0.37	0.35	0.24	0.11
RM3	0.50	0.46*	0.42*	0.40*	0.38	0.24	0.11
RM3_TQP	**0.52***	**0.50*†**	**0.46*†**	**0.45*†**	**0.40*†**	**0.26*†**	**0.12*†**

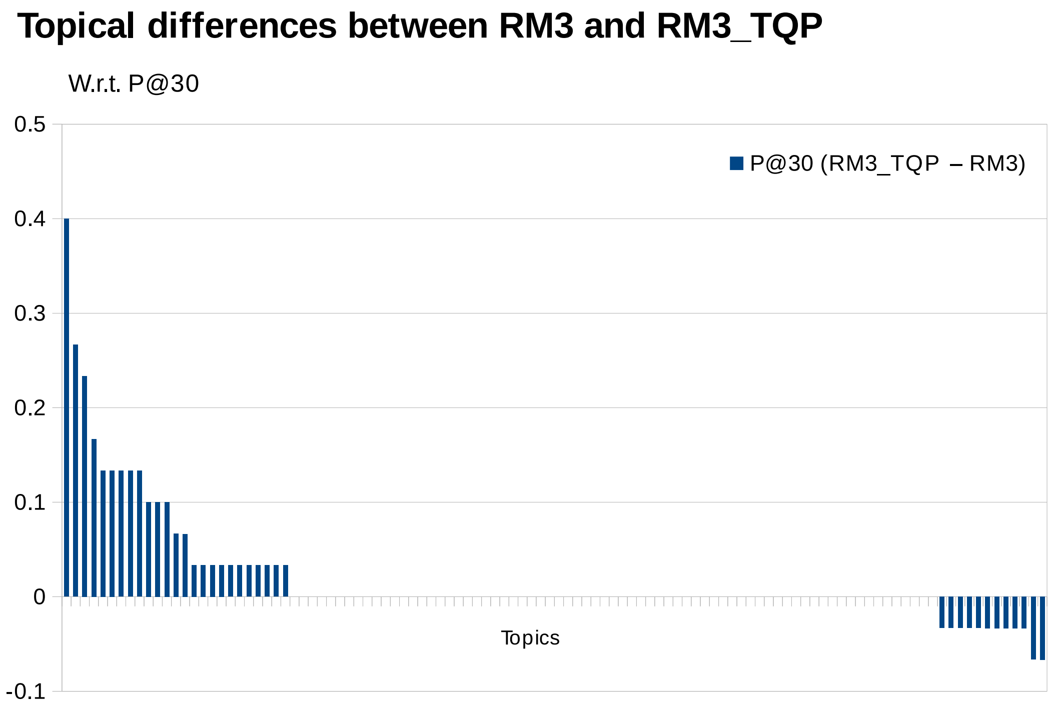
Table 6.7: Retrieval Runs for 2011 and 2012 collections (*$p < 0.05$ w.r.t. DFRee)(† $p < 0.05$ w.r.t. RM3)

Predictor classifier with the RM3 technique for AQE. Looking at the results for the 2011 collection, the differences between RM3 and RM3_TQP are not significant at earlier precision points, and remain almost the same throughout all measures. The exception with P@30 and P@100 where our approach RM3_TQP does behave significantly better. Moreover, looking at the results for the 2012 collection we can see how the RM3_TQP approach, performs significantly better than RM3 alone, for all metrics except P@5. Furthermore, RM3_TQP significantly outperforms DFRee in terms of P@5, which demonstrates its increased stability in scoring.

To further understand the behaviour of our approach with respect to RM3 we show performance differences for each topic in Figure 6.8 in terms of P@30. As we can observe, considerably more topics are improved substantially (25 topics, 10% better on average) than those worsened (Only 12 topics 3.9% worse on average). There is also a large group of topics that are unaffected by the new approach. We can conclude from these results that the selection and boosting of term scores is better capturing the importance or quality of terms for expansion, than in the case of RM3 alone.

Looking at the results for both the 2011 and 2012 collections, we can see that expanding topics for the 2011 collection is considerably more challenging than for the 2012 counterpart. This could be due to the relevance judgements not having many documents assessed containing other terms of interest aside from query terms. However we have demonstrated that our approach produces significantly better results over the AQE baseline.

Figure 6.8: Topic by topic differences in terms of P@30 for both 2011 and 2012 collections.



It is important to note, that there are still many actions we could take to improve the performance. Such as devising more fitting boosting factors for the each of the classes or finding out which is the optimal number of terms to take into the RM3 method as well as the number of documents to take into account for the pseudo relevant set. However this is out of the scope of the work, since we are only demonstrating the feasibility of our classification and AQE approaches.

Summarising, our results confirm that it is possible to predict the usefulness of a word to be used for query expansion. We demonstrated significantly better performance with our RM3_TQP approach, due to the better capturing of the relevance of a term with respect to a topic. Our work is specially promising, as the applications of our classifier are many-fold. Our work could be easily adapted to improve the behaviour of any AQE approach or as a part of re-ranking techniques, acting as a validation layer to determine the importance of terms.

Finally, the results of this work represent a successful step ahead in reducing the uncertainty about the importance of terms in PRF-based AQE approaches which is a long known problem. Any step in this direction is vital for improving the retrieval performance of microblog ad-hoc search due to the popularity of AQE approaches.

6.10 Conclusions

In this chapter we have addressed the vocabulary mismatch problem through the application of Automatic Query Expansion (AQE). The first part of the chapter challenges the use of document scores produced by retrieval models as a reliable source of information towards selecting the best expansion terms. We hypothesised that based on our Chapter 3, document scores may not be representative of the actual relevance of the document, and can lead to misleading decisions by AQE approaches.

In order to test our hypothesis we introduced a novel approach for PRF-based automatic query expansion. To estimate the value of a term, we pay attention to the rank of the document it is found within. We then derive its usefulness towards being used as a query expansion term utilising a rank-based function, sometimes coupled with collection statistic evidences gathered by IDF. A number of different approaches were developed combining linear or logarithmic normalisation methods with respect to the rank value of a document and usage or absence of the IDF statistic.

The performance of our approaches was compared to the state of the art AQE approach RM3 which utilises the score assigned to a document by a retrieval model in its computation. Our experimental results demonstrate how utilising rank-based features can be more effective as our AQE_IDF_Log approach achieved more often significantly better results with respect to the baseline than RM3, thus positively resolving **RQ4**. It is worth mentioning that in average our approach behaved similar to RM3 given the chosen evaluation metrics, however it affected a very different set of topics than RM3. This opens the possibility to utilise each method selectively to the type of topic at hand as authors like Amati et al. (2004) have already accomplished in other contexts.

The second part of the chapter dealt with reducing the topical drift which often undermines the behaviour of Pseudo Relevance Feedback based approaches to Automatic Query Expansion.

We hypothesised that we can differentiate terms that are optimal for query expansion from those that are not. To test our hypotheses we followed number of steps. Firstly we introduced a number of features derived from IDF, which allowed us to draw statistically significant differences from a training set between those terms appearing most often in relevant documents, those in non-relevant documents, and those that appear equally in both groups of documents. Moreover, we built a classifier that leveraged such features in an attempt to characterise terms as being of "Low-Value", "Medium-Value" and High-Value" with the view of prospectively using them for query expansion. Our features and classifier achieved good performance but the final test of our hypotheses was done in a practical setting, on our testing dataset. To this end, we extended the definition of the RM3 technique to include a boosting factor defined by the predicted class given by our classifier. Our experimental results demonstrate statistically significant better results when utilising our RM3_TQP approach over the original RM3. These results served as a strong confirmation of our hypotheses and as an answer to **RQ5**, as we proved that it is possible to find terms which are optimal for query expansion by means of our classifier based on IDF-related features.

Finally our contributions - in terms of new features and the quality classifier - open up new possibilities. The further definition and testing of new features may improve the classification effectiveness, which would in turn lead to better AQE performance. The classifier itself could also be deployed as part of other approaches, such as a re-ranking technique in order to better assess the importance of terms.

Part V
Conclusions

Chapter 7

Conclusions

7.1 Conclusions and Future Work

The main objective of this thesis was the exploration of issues affecting ad-hoc retrieval of microblog documents. Furthermore, we studied and proposed a number of techniques to enhance the ad-hoc results of state of the art approaches. The thesis was structured around three main parts as described in the following sections:

7.1.1 Relevance and Informativeness of Microblogs

Part II starts by studying the reasons behind the erratic performance of state of the art retrieval models in ad-hoc microblog retrieval tasks in Chapter 3. In this chapter we posed the following research questions and reached the following conclusions:

- **RQ1:** How are state of the art retrieval models affected by the morphology of microblog documents in an ad-hoc retrieval scenario?
  - **RQ1.A:** Why do certain models perform better than others in the Microblog domain?
  - **RQ1.B:** What are the best parameters for each state of the art retrieval model in the Microblog domain?
  - **RQ1.C:** Can we build a custom retrieval model to better capture the relevance of documents?

In order to answer these questions we analysed the behaviour of the state of the art retrieval models BM25, HLM, IDF, DFRee and DLM, in the context of microblog ad-hoc retrieval. Our first outcome was the expansion of our understanding regarding the shortcomings experienced by these models when utilised for microblog ad-hoc retrieval. Particularly, we learned that longer documents should be promoted in order to account for the effort out by authors to encode their messages within the character limit. Moreover we identified that documents containing higher query term frequencies than 1-2 should be penalised as this is reminiscent of spam content. We concluded that the scope hypotheses does still hold for microblog ad-hoc retrieval as generally longer documents are more informative. However the verbosity hypothesis does not hold due to the limitation in character length (**RQ1.A**). Additionally we performed an exhaustive examination of the best parameters for each considered retrieval model under microblog conditions (**RQ1.B**). Finally as a product from our study we designed a retrieval model optimised for microblog ad-hoc retrieval, namely MBRM. Our experimental evaluation demonstrated how MBRM significantly outperforms the best baselines (IDF and DFRee), by giving preference to longer documents with query terms term frequencies closer to 1 (**RQ1.C**). Future work will demonstrate how MBRM can be used to further improve the current performance of pseudo relevance feedback based approaches such as Automatic Query Expansion. Furthermore, we will explore all possible parameters in order to optimise its overall performance on microblog collections.

On the other hand Chapter 4 studied the relationship between the four dimensions of a microblog document and relevance and was driven by the following questions:

- **RQ2:** Can we define informativeness for microblogs in terms of their inherent features?
  - **RQ2.A:** Can we exploit microblog specific features in order to improve ad-hoc retrieval searches?
  - **RQ2.B:** Are there differences between relevant and non-relevant microblogs in terms of their structure? Can we leverage their structure to produce better rankings in ad-hoc searches?

Dimensions consist of the inherent elements to any tweet, namely Text; URL; Mentions and Hashtags. Consequently we developed the notion of "Microblog Informativeness", which connects the relevance of microblog documents with their structure, in order to better satisfy a user's *information need* expressed as a query. We then tested our hypotheses, by proposing a number of techniques which utilise the number of characters used for each microblog dimension to re-weight the retrieval score of a microblog document. Our technique allowed us to significantly improve the performance of a state of the art retrieval model in the context of ad-hoc microblog retrieval, thus confirming **RQ2.A**.

Finally, we extended our study to account for the different variations in the ordering of microblog dimensions. We built state machines to capture the structure of known relevant and non-relevant documents. Subsequently we developed an approach that derives scores from the state machines based on the similarity of an unobserved document to each of the state machines. Our experimentation, demonstrated with statistical significance that it is possible to utilise the structure of tweets as evidence of their relevance in order to improve the ad-hoc retrieval of microblogs, which validated **RQ2.B**.

Future work will further expose the relations between these dimensions as well as finding further applications of the features described in this work for other purposes, such as Automatic Query Expansion. Novel approaches to model the transitions between microblog elements more closely could also lead to improved performance. Moreover, at the time of submission of this doctoral work the character count limit of Twitter has been extended to 280 (Previously 140). This will have an effect over some of the findings in this Part II. However, we believe it will not fundamentally change the conclusions, as only a small fraction of users seem to take advantage of this extension[1]. This is probably due to how people have been accustomed to compose their messages over the years in this medium. Future work should test the extent of the effect of this fundamental change over the conclusions we have reached in this work.

7.1.2 Query Performance Prediction

Part III comprehends the experiments on query performance prediction introduced in Chapter 5. Query performance prediction is the estimation of the level of success in retrieving the right documents for any given query without human intervention. In this Chapter, we studied the performance of the state of the art predictors in the context of microblogs, and it was driven by the following research question:

- **RQ3:** To what extent can we predict query performance during ad-Hoc retrieval of microblog documents?

The most evident outcome of predicting query performance is increasing the robustness of PRF-based AQE approaches, as we could estimate when it is most appropriate to apply AQE to a given topic. Consequently we focused on the performance prediction in terms of the top retrieved documents measured by evaluation metrics such as Precision@10 (P@10).

Our evaluation concluded that predictors described in the literature perform poorly in the context of microblogs, thus it prompted the need for predictors that are better

[1] http://money.cnn.com/2017/11/07/technology/twitter-280-character-limit/index.html

suited for this purpose. Consequently, we defined a number of predictors which rely on features and characteristics more closely related to microblogs. We benchmarked their performance and demonstrated how most of them outperform those in the literature, with TTCov being the most correlated with MAP and P@5. However, whilst some of the predictors we proposed, such as TTCov and QTCov considerably outperformed state of the art prediction models in the context of microblogs, their performance not enough to enable effective **selective** query expansion.

In order to improve over our best performance we performed a set of experiments to combine predictors together. The first of such experiments employed support vector machines for regression to learn a model based on the best performing predictors. The resulting model increased performance by a +22% in terms of the Pearson correlation coefficient, and +12.88% for K.Tau.

Secondly, we looked at the same problem from a classification point of view. We divided the topics into three categories with respect to the $P@10$ obtained. the categories were defined as **low** ($P@10 < 0.25$), **medium** ($P@10 > 0.25\ andP@10 < 0.75$) and **high** ($P@10 > 0.75$). This time we attempted to study whether it was possible to predict such classes in order to selectively apply AQE to those topics with medium and high P@10. Thus we can avoid the topical drift produced when applying AQE to low performing topics. Our evaluation experiments show promising results in classifying low performing topics (0.78 True positives rate with 0.518 precision) and high (0.68 True positives rate with 0.721 precision) performance topics, whilst topics with medium performance ($P@10 > 0.25\ andP@10 < 0.75$) are much harder to predict.

Our experiments suggest that we can manage reasonable prediction performance, particularly when combining our predictors improving upon the predictors in the literature (**RQ3**). However it is still to be confirmed if current performance could be useful in a practical scenario.

Future work will put these findings to a practical application in selective approaches to PRF-based AQE, or in the selection of a baseline model to optimize the overall performance of a system given the conditions of a particular query. Furthermore, we will study the performance of other predictors which will consider other microblog specific features.

7.1.3 Automatic Query Expansion

Part IV deals with the topic of automatic query expansion and is driven by the following research questions:

- **RQ4:** Are retrieval model scores unreliable when determining the importance of terms in a pseudo relevant set, when utilised by automatic query expansion techniques?
- **RQ5:** Is it possible to predict the importance of a term within a pseudo relevant set before it is used for query expansion? Can this evidence improve AQE approaches?

In Chapter 6 we challenged the use of document scores produced by retrieval models as a reliable source of information to be used by AQE approaches in selecting the best expansion terms. Based on our Chapter 3 we hypothesised that document scores may not be representative of the actual relevance of the document, and can lead to misleading estimations by AQE methodologies.

We tested our hypothesis by introducing a novel approach for PRF-based automatic query expansion, which does not rely directly on the scores produced by retrieval models. Instead, to estimate the value of a term, we paid attention to the numerical rank of the documents. We then derive the usefulness of prospective query expansion terms by means of a rank-based function, which sometimes is coupled with collection statistic evidences provided by IDF. A number of different approaches were derived which combined either linear or logarithmic normalisation methods with respect to the rank value of a document and the usage or absence of the IDF statistic. Then we benchmarked the performance of our approaches to the state of the art AQE approach RM3 which does utilise document scores in its computation. Our experimental results demonstrated how utilising rank-based features can be more effective and stable as our AQE_IDF_Log approach achieved more often significantly better results over the baseline than RM3 thus validating **RQ4**. We also discovered that RM3 and our approaches affected a very different set of topics than RM3 which opens the possibility to selectively apply each method depending on the type of topic.

The second part of this chapter dealt with reducing the topical drift which often undermines Automatic Query Expansion approaches. Thus, we hypothesised that we could differentiate terms that are optimal for query expansion from those that are not. In order to test our hypotheses we firstly introduced a number of features derived from IDF, which allowed us to draw statistically significant differences between different types of terms. We annotated these terms from a training set as “High-Value” (Those appearing most often in relevant documents), “Low-Value” (Those in non-relevant documents), and “Medium-Value” (those that appear equally in both groups of documents). Moreover, we built a classifier that leveraged such features in an attempt to characterise terms as belonging to the above-mentioned categories, in order to estimate their quality when using them for query expansion. Our features and classifier achieved very good performance but the final confirmation of our hypotheses was performed in a practical setting, on a testing dataset. Consequently, we modified the definition of the RM3 technique to include a boosting factor given by the class predicted by our classifier. Our experimental results demonstrated statistically significant improved results when utilising our RM3_TQP approach over the original RM3, which served as a strong confirmation to our hypothesis, and validation of **RQ5**.

Future work, will explore other document score independent features in order to further relieve PRF-methods from the unreliable behaviour of retrieval models on microblog collections. Furthermore, we will explore the performance of our term quality classifier coupled with other AQE approaches, as well as part of a re-ranking mechanism.

References

Younos Aboulnaga, Charles L. A. Clarke, and David R. Cheriton. Frequent itemset mining for query expansion in microblog ad-hoc search. *TREC Microblog*, 2012. 19, 21, 96

G. Amati, G. Amodeo, M. Bianchi, A. Celi, C. Di Nicola, M. Flammini, C. Gaibisso, G. Gambosi, and G. Marcone. Fub, iasi-cnr, univaq at trec 2011 microblog track. *TREC Microblog*, 2011. 19, 21, 96

Giambattista Amati, Claudio Carpineto, and Giovanni Romano. *Query Difficulty, Robustness, and Selective Application of Query Expansion*, pages 127–137. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-24752-4. 108, 124

Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002. 116

Gianni Amati, Cornelis Joost, and Van Rijsbergen. Probabilistic models for information retrieval based on divergence from randomness. 2003. 26, 27

Ayan Bandyopadhyay, Kripabandhu Ghosh, Prasenjit Majumder, and Mandar Mitra. Query expansion for microblog retrieval. *International Journal of Web Science*, 1(4): 368–380, 2012. 22, 97

Amparo Elizabeth Cano Basave, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. Making sense of microposts (# msm2013) concept extraction challenge. In *# MSM*, pages 1–15, 2013. 52

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual* *international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 243–250, New York, NY, USA, 2008. ACM, ACM. ISBN 978-1-60558-164-4. 109

David Carmel and Elad Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1): 1–89, 2010. 73, 76, 108

C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1, 2012. 16, 82, 95, 96, 99

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 114

Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM, 2002. 17, 71

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Precision prediction based on ranked list coherence. *Information Retrieval*, 9(6):723–755, Dec 2006. ISSN 1573-7659. 108

Firas Damak, Karen Pinel-Sauvagnat, Mohand Boughanem, and Guillaume Cabanac. Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 914–919, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1656-9. 19

Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303. Association for Computational Linguistics, 2010. 26

M. Efron. Hashtag retrieval in a microblogging environment. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2010a. 22

Miles Efron. Linear time series models for term weighting in information retrieval. *Journal of the American Society for Information Science and Technology (JASIST)*, 6(7):1299–1312, 2010b. 20

Miles Efron, Jimmy Lin, Jiyin He, and Arjen de Vries. Temporal feedback for tweet search with non-parametric density estimation. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, pages 33–42, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. 97, 114

Paul Ferguson, Neil O'Hare, James Lanagan, Owen Phelan, and Kevin McCarthy. An investigation of term weighting approaches for microblog retrieval. In *Advances in Information Retrieval*, pages 552–555. Springer, 2012. 20, 28, 32, 33

George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987. 4, 93, 94

Jinhua Gao, Guoxin Cui, Shenghua Liu, Yue Liu, and Xueqi Cheng. Ictnet at microblog track in trec 2013. *TREC Microblog*, 2013. 19

Davide Feltoni Gurini and Fabio Gasparetti. Trec microblog 2012 track: Real-time algorithm for microblog ranking systems. *TREC Microblog*, 2012. 79, 81, 97

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145. 114

Zhongyuan Han, Xuwei Li, Muyun Yang, Haoliang Qi, Sheng Li, and Tiejun Zhao. Hit at trec 2012 microblog track. *TREC Microblog*, 2012. 19, 21, 96

Claudia Hauff. Predicting the effectiveness of queries and retrieval systems. University of Twente, 2010. x, 72, 74, 75

Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1419–1420. ACM, 2008. 71, 72, 75, 84

Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval*, pages 43–54. Springer, 2004. 71, 75

Ben He and Iadh Ounis. Combining fields for query expansion and adaptive query expansion. *Information processing & management*, 43(5):1294–1307, 2007. 109

Jiyin He, Martha Larson, and Maarten De Rijke. Using coherence-based measures to predict query difficulty. In *Advances in Information Retrieval*, pages 689–694. Springer, 2008. 72

D. Hiemstra. Using language models for information retrieval. *Thesis, University of Twente*, 2001. 26, 28

Djoerd Hiemstra and Arjen P De Vries. Relating the new language models of information retrieval to the traditional retrieval models. 2000. 35

S. Ishikawa, Y. Arakawa, S. Tagashira, and A. Fukuda. Hot topic detection in local areas using twitter and wikipedia. In *ARCS Workshops (ARCS), 2012*, pages 1 –5, 2012. 21

Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem. Irit at trec microblog track 2013. *TREC Microblog*, 2013. 19

D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 180–191. VLDB Endowment, 2004. 21

Yubin Kim, Reyyan Yeniterzi, and Jamie Callan. Overcoming vocabulary limitations in twitter microblogs. *TREC Microblog*, 2012. 19, 21, 96

J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003. 20

T. Lappas, M.R. Vieira, D. Gunopulos, and V.J. Tsotras. On the spatiotemporal burstiness of terms. *Proceedings of the VLDB Endowment*, 5(9):836–847, 2012. 21

C.H. Lau, Y.F. Li, and D. Tjondronegoro. Microblog retrieval using topical features and query expansion. *TREC Microblog*, 2011. 21, 96

Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. 97, 98, 114

C.H. Lee, C.H. Wu, and T.F. Chien. Burst: a dynamic term weighting scheme for mining microblogging messages. *Advances in Neural Networks–ISNN 2011*, pages 548–557, 2011. 21

X. Li and W.B. Croft. Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 469–475. ACM, 2003. 20

Y. Li, Z. Zhang, W. Lv, Q. Xie, Y. Lin, R. Xu, W. Xu, G. Chen, and J. Guo. Pris at trec2011 micro-blog track. *TREC Microblog*, 2011. 19, 21, 96

Christopher Manning, Hinrich Schtze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge University Press, 2008.

Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in Information Retrieval*, pages 362–367. Springer, 2011. 52, 57

D. Metzler and C. Cai. Usc/isi at trec 2011: Microblog track. In *TREC Microblog*, 2011. 19, 21, 96

Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–655, Montréal, Canada, June 2012. Association for Computational Linguistics. 21

Rinkesh Nagmoti, Ankur Teredesai, and Martine De Cock. Ranking approaches for microblog search. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 153–157. IEEE, 2010. 52

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 183–188. ACM, 2011. 20, 26, 28, 30, 31, 41, 49

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings SIGIR'06 Workshop (OSIR 2006)*, 2006. 42

I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the trec-2011 microblog track. In *TREC Microblog*, 2011. 4, 19

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *Advances in Information Retrieval*, pages 517–519. Springer, 2005. 29

S.E. Robertson. The probability ranking principle in ir. *Journal of documentation*, 33 (4):294–304, 1977. 14

Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009. 15, 27, 31

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995. 26

Jesus Rodriguez Perez, Teerapong Leelanupab, and Joemon M Jose. Cofox: A synchronous collaborative browser. In *Information Retrieval Technology*, pages 262–274. Springer, 2012a. 10

Jesus A. Rodriguez Perez and Joemon M. Jose. Predicting query performance in microblog retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, pages 1183–1186, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. 10, 113

Jesus A Rodriguez Perez, Stewart Whiting, and Joemon M Jose. Cofox: A visual collaborative browser. 2011. 11

Jesus A Rodriguez Perez, Andrew J McMinn, and Joemon M Jose. University of glasgow (uog_tw) at trec microblog 2012. *TREC Microblog*, 2012b. 11

Jesus A Rodriguez Perez, Andrew J McMinn, and Joemon M Jose. University of glasgow (uog_twteam) at trec microblog 2013. *TREC Microblog*, 2013a. 10, 19

Jesus A Rodriguez Perez, Yashar Moshfeghi, and Jose Joemon. On using inter-document relations for microblog retrieval. In *WWW 2013*, page p75. ACM, 2013b. 10

Jesus Alberto Rodriguez Perez and Joemon M. Jose. On microblog dimensionality and informativeness: Exploiting microblogs' structure and dimensions for ad-hoc retrieval. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 211–220, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3833-2. 10

Thomas Roelleke. Information retrieval models: Foundations and relationships. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5(3):1–163, 2013. 15

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988. ISSN 0306-4573. 16

D. Shan, W.X. Zhao, R. Chen, S. Baihan, H. Yan, and X. Li. Eventsearch: A system for event discovery and retrieval on multi-type historical data. KDD, 2012. 21

B. Sharifi, M.-A. Hutton, and J.K. Kalita. Experiments in microblog summarization. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 49–56, Aug 2010. 52

Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM, 1996. 15, 26

Shuangyong Song, Qiudan Li, and Hongyun Bao. Detecting dynamic association among twitter topics. In *Proceedings of the 21st international conference companion on*

World Wide Web, WWW '12 Companion, pages 605–606, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1230-1. 21

Ke Tao, Fabian Abel, Claudia Hauff, and Geert-Jan Houben. What makes a tweet relevant for a topic? *Making Sense of Microposts (# MSM2012)*, pages 49–56, 2012. 52

J. Teevan, D. Ramage, and M.R. Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011. 26, 28, 80

Sarvnaz Karimi Jie Yin Paul Thomas. Searching and filtering tweets: Csiro at the trec 2012 microblog track. *TREC Microblog*, 2012. 19

Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press, 2005. ISBN 0262220733. 13, 17

Jianshu Weng, Yuxia Yao, Erwin Leonardi, and Francis Lee. Event detection in twitter. In *ICWSM '11'*, 2011. 21

Stewart Whiting, Yashar Moshfeghi, and Joemon M. Jose. Exploring term temporality for pseudo-relevance feedback. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1245–1246, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. 21, 96

Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA, 1996. ACM. ISBN 0-89791-792-8. 17, 96

Zhen Yang, Guangyuan Zhang, Shuyong SI, Yingxu LAI, and Kefeng FAN. Bjut at trec 2013 microblog track. *TREC Microblog*, 2013. 19

Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow. Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR '05, pages 512–519, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5. 108, 109

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001. 26, 28

Jiayue Zhang, Sijia Chen, Yue Liu, Jie Yin, Qianqian Wang, Weiran Xu, and Jun Guo. Pris at 2012 microblog track. 2012. 22

Ying Zhao, Falk Scholer, and Yohannes Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval*, pages 52–64. Springer, 2008. 17, 72, 75, 76

Siming Zhu, Zhe Gao, Yajing Yuan, Hui Wang, and Guang Chen. Pris at 2013 microblog track. 2013. 19

Meriem Amina Zingla, Latiri Chiraz, and Yahya Slimani. Short query expansion for microblog retrieval. *Procedia Computer Science*, 96(Supplement C):225 – 234, 2016. ISSN 1877-0509. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES-2016. 22

Appendix A

QPP Predictors Correlation tables

This appendix contains the correlation results for the performance predictors specified in Chapter 5. Correlations of predictors are computed with respect to the evaluation measures obtained when retrieving microblog documents using DFRee and IDF models on the Tweets 11 and 12 collections.

Table A.1: Ranked list of correlations between predictors and evaluation measures for DFRee runs (Statistical significance: **$p < 0.01$ & *$p < 0.05$)

DFRee model correlations

Eval Measure	**Predictor/Feature**	**K.Tau**	**SP. Rho**	**Pearson**
P_10	post_http	0.163 **	0.206 **	0.213
P_10	post_firstScore	0.022	0.029	-0.006
P_10	post_ambiguity_min	0.075	0.092 *	0.076
P_10	post_ambiguity_max	-0.059	-0.091	0.020
P_10	post_ambiguity_mean	-0.012	-0.016	0.046
P_10	post_ambiguity_median	0.017	0.024	0.042
P_10	post_ambiguity_std	-0.148 *	-0.198	-0.045
P_10	post_tweetLength	-0.098	-0.130	-0.131
P_10	post_QTCov_mean	0.291 **	0.382 **	0.375
P_10	post_QTCov_median	0.305 **	0.382 **	0.373
P_10	post_QTCov_upper	0.325 **	0.404 **	0.392
P_10	post_QTCov_lower	0.266 **	0.336 **	0.312
P_10	post_QTCov_diff	-0.107	-0.134	-0.125
P_10	post_TTCov_mean	0.301 **	0.416 **	0.429
P_10	post_TTCov_median	0.365 **	0.456 **	0.441
P_10	post_TTCov_upper	0.264 **	0.355 **	0.374
P_10	post_TTCov_lower	0.253 *	0.303 **	0.298
P_10	post_TTCov_diff	0.028	0.036	-0.008
P_10	post_TTCov_max	0.084	0.102 **	0.081
P_10	post_TTCov_cond	0.256 **	0.337 **	0.356
P_10	post_QEAndQT_25	0.412 **	0.560 **	0.552

Continued on next page

Table A.1 – continued from previous page

DFRee model correlations

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
P_10	post_QEAndQT_50	0.424 **	0.576 **	0.561
P_10	post_QEAndQT_75	0.380 **	0.514 **	0.504
P_10	post_PMI_Mean	0.042	0.054 **	-0.022
P_10	post_PMI_Max	0.019	0.024 **	-0.025
P_10	post_NQC	0.147 **	0.209 **	0.197
P_10	post_WIG	0.127 *	0.180 *	0.113
P_10	post_TimeCH_lower	-0.212 **	-0.286 **	-0.236
P_10	post_TimeCH_median	-0.145 **	-0.199 *	-0.239
P_10	post_TimeCH_upper	-0.023	-0.033	-0.074
P_10	post_TimeCH_mean	-0.170 **	-0.233 **	-0.212
P_10	post_TimeCH_diff	0.192 **	0.269 **	0.198
P_10	post_HashTagCount_min	0.009	0.011 **	-0.003
P_10	post_HashTagCount_median	0.007	0.009 **	-0.005
P_10	post_HashTagCount_max	0.029	0.035 *	0.057
P_10	post_HashTagCount_mean	0.024	0.030 *	0.016
P_10	post_HashTagCount_diff	0.054	0.064 **	0.099
P_10	pre_idf_max	0.032	0.050	0.031
P_10	pre_SCS_std	0.016	0.023	0.023
P_10	pre_SCQ_min	0.072	0.107	0.139
P_10	pre_SCQ_max	0.120 *	0.177 *	0.159
P_10	pre_queryScope	0.006	0.014	-0.020
P_10	pre_posting_median	0.052	0.072	-0.067
P_10	pre_SCQ_sum	0.109 *	0.158 *	0.187
P_10	pre_posting_mean	0.070	0.097	0.055
P_10	pre_VAR_max	0.116	0.143 **	0.160
P_10	pre_SCS_min	-0.091	-0.130	-0.147
P_10	pre_VAR_sum	0.118	0.144 **	0.159
P_10	pre_posting_std	0.082	0.114	0.085
P_10	pre_idf_diff	0.048	0.070	0.053
P_10	pre_SCS_max	-0.086	-0.121	-0.144
P_10	pre_posting_max	0.066	0.095	0.104
P_10	pre_posting_diff	0.084	0.119	0.109
P_10	pre_posting_min	0.035	0.045	-0.048
P_10	pre_SCS_sum	-0.075	-0.103	-0.104
P_10	pre_idf_mean	0.031	0.041	0.018
P_10	pre_idf_min	0.010	0.017	-0.023
P_10	pre_idf_std	0.023	0.035	0.032
P_10	pre_SCQ_std	0.002	0.004	-0.035
P_10	pre_idf_median	0.036	0.050	0.033

Table A.2: Ranked list of correlations between predictors and evaluation measures for IDF runs (Statistical significance: $**p < 0.01$ & $*p < 0.05$)

DFR model correlations with P 30

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
P_30	post_http	0.104	0.146	0.102
P_30	post_firstScore	-0.022	-0.026	-0.112
P_30	post_ambiguity_min	0.101	0.130 **	0.148
P_30	post_ambiguity_max	0.002	0.000	0.092
P_30	post_ambiguity_mean	0.036	0.047	0.133
P_30	post_ambiguity_median	0.085	0.112 *	0.140
P_30	post_ambiguity_std	-0.087	-0.120	0.008
P_30	post_tweetLength	-0.068	-0.101	-0.086
P_30	post_QTCov_mean	0.279 **	0.393 **	0.396
P_30	post_QTCov_median	0.268 **	0.353 **	0.361
P_30	post_QTCov_upper	0.312 **	0.406 **	0.405
P_30	post_QTCov_lower	0.222 **	0.299 **	0.310
P_30	post_QTCov_diff	-0.032	-0.041	-0.068
P_30	post_TTCov_mean	0.261 **	0.373 **	0.433
P_30	post_TTCov_median	0.289 **	0.355 **	0.375
P_30	post_TTCov_upper	0.290 **	0.371 **	0.380
P_30	post_TTCov_lower	0.244	0.293 **	0.330
P_30	post_TTCov_diff	0.069	0.086 *	0.046
P_30	post_TTCov_max	0.088	0.108 **	0.082
P_30	post_TTCov_cond	0.268 **	0.343 **	0.324
P_30	post_QEAndQT_25	0.293 **	0.426 **	0.475
P_30	post_QEAndQT_50	0.314 **	0.452 **	0.508
P_30	post_QEAndQT_75	0.259 **	0.374 **	0.396

Continued on next page

Table A.2 – continued from previous page

IDF model correlations

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
P_30	post_PMI_Mean	0.103	0.141 *	0.114
P_30	post_PMI_Max	0.071	0.102 *	-0.016
P_30	post_NQC	0.184 **	0.267 **	0.226
P_30	post_WIG	0.063	0.086	0.013
P_30	post_TimeCH_lower	-0.173 **	-0.251 **	-0.268
P_30	post_TimeCH_median	-0.114 *	-0.167 *	-0.197
P_30	post_TimeCH_upper	-0.001	-0.001	-0.060
P_30	post_TimeCH_mean	-0.135 *	-0.195 *	-0.206
P_30	post_TimeCH_diff	0.105 *	0.152 *	0.120
P_30	post_HashTagCount_min	0.003	0.004 **	-0.009
P_30	post_HashTagCount_median	-0.010	-0.012	0.001
P_30	post_HashTagCount_max	0.089	0.111 **	0.104
P_30	post_HashTagCount_mean	0.066	0.088 *	0.064
P_30	post_HashTagCount_diff	0.094	0.116 **	0.107
P_30	pre_idf_max	-0.015	-0.017	-0.084
P_30	pre_SCS_std	-0.015	-0.023	-0.040
P_30	pre_SCQ_min	0.061	0.093	0.165
P_30	pre_SCQ_max	0.088	0.120	0.143
P_30	pre_queryScope	-0.024	-0.038	-0.092
P_30	pre_posting_median	0.052	0.079	-0.047
P_30	pre_SCQ_sum	0.106 *	0.152 *	0.213
P_30	pre_posting_mean	0.054	0.083	0.061
P_30	pre_VAR_max	0.066	0.085 **	0.084
P_30	pre_SCS_min	-0.065	-0.092	-0.146

Continued on next page

Table A.2 – continued from previous page

IDF model correlations

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
P_30	pre_VAR_sum	0.067	0.086 **	0.096
P_30	pre_posting_std	0.067	0.095	0.080
P_30	pre_idf_diff	0.025	0.037	-0.011
P_30	pre_SCS_max	-0.075	-0.113	-0.162
P_30	pre_posting_max	0.049	0.074	0.107
P_30	pre_posting_diff	0.069	0.098	0.109
P_30	pre_posting_min	0.054	0.070	0.015
P_30	pre_SCS_sum	-0.059	-0.089	-0.137
P_30	pre_idf_mean	-0.014	-0.028	-0.087
P_30	pre_idf_min	-0.013	-0.023	-0.083
P_30	pre_idf_std	-0.002	-0.004	-0.037
P_30	pre_SCQ_std	-0.036	-0.051	-0.106
P_30	pre_idf_median	-0.002	-0.011	-0.069

Table A.3: Ranked list of correlations between predictors and evaluation measures for IDF runs (Statistical significance: $**p < 0.01$ & $*p < 0.05$)

DFR model correlations with MAP

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
map	post_http	0.093	0.135	0.132
map	post_firstScore	0.176 **	0.254 **	0.187
map	post_ambiguity_min	-0.008	-0.013	0.009
map	post_ambiguity_max	-0.069	-0.105	-0.004
map	post_ambiguity_mean	-0.059	-0.092	0.011

Continued on next page

Table A.3 – continued from previous page

IDF model correlations

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
map	post_ambiguity_median	-0.035	-0.052	0.023
map	post_ambiguity_std	-0.061	-0.087	-0.023
map	post_tweetLength	-0.058	-0.077	-0.042
map	post_QTCov_mean	0.062	0.094	0.062
map	post_QTCov_median	0.049	0.067	0.035
map	post_QTCov_upper	0.092	0.124 *	0.070
map	post_QTCov_lower	0.027	0.036	0.038
map	post_QTCov_diff	0.053	0.075	0.078
map	post_TTCov_mean	0.302 **	0.447 **	0.403
map	post_TTCov_median	0.253 **	0.312 **	0.274
map	post_TTCov_upper	0.356 **	0.463 **	0.434
map	post_TTCov_lower	0.178	0.218 **	0.197
map	post_TTCov_diff	0.080	0.104 **	0.107
map	post_TTCov_max	0.087	0.113 **	0.120
map	post_TTCov_cond	0.362 **	0.460 **	0.422
map	post_QEAndQT_25	0.117 *	0.174 *	0.169
map	post_QEAndQT_50	0.144 **	0.206 **	0.209
map	post_QEAndQT_75	0.100	0.152 *	0.124
map	post_PMI_Mean	0.076	0.111	0.096
map	post_PMI_Max	0.129 *	0.174 **	0.072
map	post_NQC	0.245 **	0.360 **	0.279
map	post_WIG	0.146 **	0.214 **	0.137
map	post_TimeCH_lower	-0.202 **	-0.300 **	-0.273
map	post_TimeCH_median	-0.200 **	-0.291 **	-0.310

Continued on next page

Table A.3 – continued from previous page

IDF model correlations

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
map	post_TimeCH_upper	-0.122 *	-0.188 *	-0.231
map	post_TimeCH_mean	-0.197 **	-0.288 **	-0.281
map	post_TimeCH_diff	0.030	0.046	-0.004
map	post_HashTagCount_min	-0.048	-0.060	-0.069
map	post_HashTagCount_median	-0.045	-0.056	-0.031
map	post_HashTagCount_max	0.070	0.088 *	0.093
map	post_HashTagCount_mean	0.042	0.056	0.019
map	post_HashTagCount_diff	0.085	0.108 **	0.104
map	pre_idf_max	0.139 **	0.203 **	0.163
map	pre_SCS_std	0.056	0.078	0.053
map	pre_SCQ_min	-0.094	-0.142	-0.089
map	pre_SCQ_max	-0.022	-0.038	-0.032
map	pre_queryScope	0.016	0.024	-0.047
map	pre_posting_median	-0.104 *	-0.150	-0.130
map	pre_SCQ_sum	0.094	0.138	0.254
map	pre_posting_mean	-0.080	-0.112	-0.019
map	pre_VAR_max	0.104	0.141 **	0.069
map	pre_SCS_min	-0.009	-0.011	-0.116
map	pre_VAR_sum	0.104	0.139 **	0.079
map	pre_posting_std	-0.016	-0.027	0.029
map	pre_idf_diff	0.140 **	0.209 **	0.200
map	pre_SCS_max	0.009	0.010	-0.094
map	pre_posting_max	-0.060	-0.090	0.059
map	pre_posting_diff	-0.011	-0.021	0.067

Continued on next page

Table A.3 – continued from previous page

IDF model correlations

Eval Measure	Predictor/Feature	K.Tau	SP. Rho	Pearson
map	pre_posting_min	-0.133 *	-0.193 *	-0.128
map	pre_SCS_sum	0.084	0.120	0.058
map	pre_idf_mean	0.097	0.141	0.080
map	pre_idf_min	0.025	0.041	-0.034
map	pre_idf_std	0.105 *	0.159 *	0.152
map	pre_SCQ_std	0.052	0.073	0.056
map	pre_idf_median	0.100	0.145	0.074

Appendix B

AQE Parameter Exploration Tables

This appendix contains the parameter optimisation tables utilised to fine-tune the AQE methods discussed in Chapter 6. There are three tables corresponding to each of the baselines considered (IDF, DFR, BM25). In the tables we show the results obtained for each AQE approach in terms of Precision and Map metrics and different configurations of MaxTerms, and MaxDocs. MaxTerms, refers to the maximum number of terms to be accepted from the pseudo relevant set, whereas MaxDocs, refers to the maximum number of documents to consider to compose the pseudo relevant set.

Table B.1: This table shows the results obtained for each considered configuration of AQE approached on the training set (First 90 topics out of all 225 topics available)

BM25: AQE optimisation table.

AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_IDF	1	5	0.4899	0.4416	0.4157	0.4	0.3592	0.2539
AQE_IDF	1	10	0.5101	0.4494	0.4217	0.4017	0.3644	0.264
AQE_IDF	1	20	0.5079	0.4551	0.4225	0.4084	0.3629	0.2532
AQE_IDF	1	30	0.4966	0.4584	0.4225	0.4101	0.3697	0.2535
AQE_IDF	3	5	0.4652	0.4382	0.4135	0.3938	0.3539	0.2578
AQE_IDF	3	10	0.5213	0.4528	0.4285	0.4034	0.3625	0.2714
AQE_IDF	3	20	0.5191	0.4944	0.4704	0.4354	0.3899	0.2691
AQE_IDF	3	30	0.5258	0.4831	0.4659	0.4365	0.3944	0.2712
AQE_IDF	5	5	0.5034	0.4596	0.4172	0.3949	0.3558	0.2572
AQE_IDF	5	10	0.5146	0.4697	0.4367	0.3989	0.3573	0.2677
AQE_IDF	5	20	0.5146	0.4753	0.4547	0.427	0.3835	0.2629
AQE_IDF	5	30	0.5393	0.4978	0.4659	0.436	0.3978	0.2786
AQE_IDF_Lin	1	5	0.4944	0.4337	0.4045	0.3826	0.3479	0.2477
AQE_IDF_Lin	1	10	0.5101	0.4427	0.4127	0.3938	0.3566	0.2555
AQE_IDF_Lin	1	20	0.5056	0.4506	0.4225	0.4101	0.3633	0.2652

Continued on next page

Table B.1 – continued from previous page

BM25: AQE optimisation table.

AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_IDF_Lin	1	30	0.4966	0.4551	0.427	0.4146	0.3693	0.2609
AQE_IDF_Lin	3	5	0.4966	0.4629	0.4262	0.3955	0.3487	0.2592
AQE_IDF_Lin	3	10	0.5124	0.4674	0.4442	0.4191	0.3712	0.2766
AQE_IDF_Lin	3	20	0.5191	0.4742	0.4472	0.4118	0.3712	0.2796
AQE_IDF_Lin	3	30	0.5461	0.5034	0.4779	0.4433	0.394	0.2843
AQE_IDF_Lin	5	5	0.4989	0.4472	0.4165	0.3854	0.3431	0.2563
AQE_IDF_Lin	5	10	0.5258	0.4652	0.433	0.3994	0.3596	0.2686
AQE_IDF_Lin	5	20	0.5461	0.4899	0.4509	0.4275	0.3816	0.2837
AQE_IDF_Lin	5	30	0.5281	0.4944	0.4637	0.4326	0.3869	0.2773
AQE_IDF_Log	1	5	0.5034	0.4393	0.4097	0.3888	0.3491	0.2494
AQE_IDF_Log	1	10	0.5124	0.4438	0.4165	0.3955	0.3596	0.2569
AQE_IDF_Log	1	20	0.4989	0.4472	0.415	0.3966	0.3562	0.2549
AQE_IDF_Log	1	30	0.5124	0.4607	0.4225	0.4011	0.3607	0.2583
AQE_IDF_Log	3	5	0.4742	0.436	0.412	0.3837	0.3386	0.2532
AQE_IDF_Log	3	10	0.5034	0.4596	0.4345	0.4022	0.3558	0.2644
AQE_IDF_Log	3	20	0.5146	0.4663	0.4449	0.414	0.3682	0.267
AQE_IDF_Log	3	30	0.5146	0.4775	0.4539	0.4242	0.379	0.2758
AQE_IDF_Log	5	5	0.4831	0.4382	0.4157	0.3876	0.3457	0.2574
AQE_IDF_Log	5	10	0.5101	0.4494	0.4225	0.3893	0.3532	0.2662
AQE_IDF_Log	5	20	0.5416	0.4742	0.4517	0.4191	0.3764	0.2711
AQE_IDF_Log	5	30	0.5461	0.4854	0.4577	0.436	0.3963	0.2772
AQE_ROC	1	5	0.5011	0.4472	0.4232	0.4	0.3584	0.2515
AQE_ROC	1	10	0.5034	0.4472	0.409	0.3938	0.3551	0.2504
AQE_ROC	1	20	0.4966	0.4438	0.4045	0.3904	0.3524	0.2457
AQE_ROC	1	30	0.4966	0.4483	0.4172	0.4	0.3629	0.2526
AQE_ROC	3	5	0.5056	0.4607	0.4292	0.3961	0.3554	0.2512
AQE_ROC	3	10	0.5056	0.4652	0.4315	0.3955	0.3517	0.2528
AQE_ROC	3	20	0.5236	0.4854	0.4554	0.4213	0.3779	0.2561
AQE_ROC	3	30	0.5191	0.4764	0.4599	0.4275	0.3794	0.2601
AQE_ROC	5	5	0.5146	0.464	0.427	0.3994	0.3528	0.2514
AQE_ROC	5	10	0.5303	0.4798	0.4449	0.4157	0.3667	0.2566
AQE_ROC	5	20	0.5191	0.4831	0.4509	0.4264	0.3749	0.2543
AQE_ROC	5	30	0.5146	0.4787	0.4569	0.4247	0.3861	0.2571
AQE_ROC_Lin	1	5	0.4944	0.4393	0.4045	0.3826	0.3427	0.2411
AQE_ROC_Lin	1	10	0.5056	0.4573	0.4165	0.4	0.3577	0.2572
AQE_ROC_Lin	1	20	0.4966	0.4483	0.4015	0.3854	0.3479	0.2485
AQE_ROC_Lin	1	30	0.5011	0.4584	0.421	0.4034	0.3562	0.2473
AQE_ROC_Lin	3	5	0.4989	0.4472	0.4157	0.386	0.3401	0.2469
AQE_ROC_Lin	3	10	0.5169	0.464	0.4345	0.4045	0.3637	0.2611
AQE_ROC_Lin	3	20	0.5551	0.5	0.4652	0.4258	0.3813	0.2703
AQE_ROC_Lin	3	30	0.5506	0.5112	0.4787	0.4382	0.3895	0.2725
AQE_ROC_Lin	5	5	0.4787	0.4281	0.4015	0.377	0.3333	0.2434
AQE_ROC_Lin	5	10	0.5124	0.464	0.4292	0.3983	0.3569	0.2574

Continued on next page

Table B.1 – continued from previous page

BM25: AQE optimisation table.								
AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_ROC_Lin	5	20	0.5326	0.5011	0.4659	0.432	0.385	0.2777
AQE_ROC_Lin	5	30	0.5573	0.5101	0.4697	0.4382	0.3873	0.276
AQE_ROC_Log	1	5	0.4876	0.4371	0.4015	0.3803	0.3401	0.2422
AQE_ROC_Log	1	10	0.5079	0.4517	0.4082	0.3882	0.3517	0.2519
AQE_ROC_Log	1	20	0.4989	0.4382	0.3978	0.3792	0.3401	0.2465
AQE_ROC_Log	1	30	0.5236	0.4584	0.4142	0.3978	0.3547	0.2524
AQE_ROC_Log	3	5	0.4876	0.4461	0.4172	0.3882	0.3416	0.2475
AQE_ROC_Log	3	10	0.5169	0.4517	0.421	0.3966	0.3573	0.2548
AQE_ROC_Log	3	20	0.5438	0.4921	0.4569	0.427	0.3798	0.2657
AQE_ROC_Log	3	30	0.5618	0.5101	0.4764	0.4466	0.3925	0.2793
AQE_ROC_Log	5	5	0.4876	0.4326	0.409	0.3848	0.3356	0.2444
AQE_ROC_Log	5	10	0.5213	0.4551	0.4255	0.3961	0.3532	0.2572
AQE_ROC_Log	5	20	0.5258	0.4719	0.4375	0.4197	0.3779	0.2594
AQE_ROC_Log	5	30	0.5438	0.4944	0.4539	0.4343	0.3955	0.275
RM3	1	5	0.5011	0.4461	0.4255	0.4045	0.3663	0.2565
RM3	1	10	0.5191	0.4517	0.4255	0.4034	0.37	0.2614
RM3	1	20	0.5191	0.4652	0.4315	0.4129	0.3693	0.2641
RM3	1	30	0.5191	0.473	0.4397	0.4242	0.3801	0.2681
RM3	3	5	0.5169	0.4551	0.4285	0.4096	0.3727	0.2695
RM3	3	10	0.5326	0.473	0.4524	0.4264	0.3809	0.2818
RM3	3	20	0.5506	0.5022	0.4801	0.4494	0.3996	0.2763
RM3	3	30	0.5551	0.5045	0.4869	0.4556	0.4097	0.2832
RM3	5	5	0.5303	0.464	0.4292	0.4011	0.3603	0.2702
RM3	5	10	0.5348	0.4753	0.4427	0.4067	0.3633	0.2792
RM3	5	20	0.5461	0.4876	0.4584	0.4281	0.382	0.2648
RM3	5	30	0.5618	0.5022	0.4734	0.4455	0.4071	0.2817

Table B.2: This table shows the results obtained for each considered configuration of AQE approached on the training set (First 90 topics out of all 225 topics available)

IDF: AQE optimisation table.								
AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_IDF	1	5	0.6311	0.58	0.5615	0.5122	0.4622	0.3845
AQE_IDF	1	10	0.6356	0.58	0.5556	0.5056	0.4563	0.3815
AQE_IDF	1	20	0.6578	0.5844	0.5526	0.5011	0.46	0.3917
AQE_IDF	1	30	0.6622	0.58	0.5526	0.5022	0.4607	0.3922
AQE_IDF	3	5	0.6578	0.5778	0.5393	0.5011	0.4607	0.3727
AQE_IDF	3	10	0.6489	0.5867	0.5556	0.5022	0.457	0.3705
AQE_IDF	3	20	0.6489	0.5889	0.5541	0.5078	0.4563	0.3719
AQE_IDF	3	30	0.6489	0.5844	0.5511	0.51	0.4563	0.3667
AQE_IDF	5	5	0.6533	0.5889	0.5481	0.5067	0.443	0.3518

Continued on next page

Table B.2 – continued from previous page

IDF: AQE optimisation table.

AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_IDF	5	10	0.6311	0.5822	0.5496	0.5067	0.4452	0.356
AQE_IDF	5	20	0.6267	0.5733	0.5378	0.5067	0.4474	0.3553
AQE_IDF	5	30	0.6044	0.5556	0.5215	0.49	0.4378	0.35
AQE_IDF_Lin	1	5	0.6489	0.5867	0.5615	0.5189	0.4719	0.4063
AQE_IDF_Lin	1	10	0.6889	0.62	0.5793	0.5356	0.4948	0.4248
AQE_IDF_Lin	1	20	0.6622	0.5978	0.563	0.5222	0.4815	0.414
AQE_IDF_Lin	1	30	0.68	0.6156	0.5807	0.5422	0.4941	0.4198
AQE_IDF_Lin	3	5	0.6444	0.58	0.5481	0.4978	0.4585	0.4125
AQE_IDF_Lin	3	10	0.68	0.6156	0.5733	0.5211	0.4726	0.4127
AQE_IDF_Lin	3	20	0.64	0.5956	0.5644	0.5233	0.4733	0.4114
AQE_IDF_Lin	3	30	0.6844	0.6244	0.5926	0.5344	0.4859	0.4177
AQE_IDF_Lin	5	5	0.6444	0.5644	0.5363	0.4933	0.4393	0.4001
AQE_IDF_Lin	5	10	0.6711	0.6133	0.5674	0.5122	0.463	0.4148
AQE_IDF_Lin	5	20	0.6444	0.6022	0.5719	0.5344	0.46	0.3912
AQE_IDF_Lin	5	30	0.6578	0.6133	0.5822	0.5478	0.4867	0.4052
AQE_IDF_Log	1	5	0.6356	0.5689	0.5511	0.5144	0.4681	0.4018
AQE_IDF_Log	1	10	0.6533	0.5889	0.5704	0.53	0.4896	0.4139
AQE_IDF_Log	1	20	0.6711	0.5956	0.5644	0.5233	0.483	0.4181
AQE_IDF_Log	1	30	0.6933	0.6089	0.5733	0.5344	0.4844	0.4157
AQE_IDF_Log	3	5	0.6311	0.56	0.5378	0.49	0.4607	0.4089
AQE_IDF_Log	3	10	0.6489	0.6067	0.5704	0.5211	0.4726	0.4246
AQE_IDF_Log	3	20	0.6711	0.6178	0.5807	0.5356	0.4837	0.4308
AQE_IDF_Log	3	30	0.6756	0.6133	0.5837	0.5378	0.4948	0.427
AQE_IDF_Log	5	5	0.6267	0.5556	0.5215	0.4833	0.4474	0.3929
AQE_IDF_Log	5	10	0.6489	0.5956	0.557	0.5133	0.46	0.4101
AQE_IDF_Log	5	20	0.64	0.5867	0.5511	0.5178	0.4652	0.407
AQE_IDF_Log	5	30	0.6756	0.6156	0.563	0.5422	0.4859	0.4101
AQE_ROC	1	5	0.6844	0.6378	0.5926	0.5567	0.4956	0.4274
AQE_ROC	1	10	0.6844	0.6356	0.5956	0.5467	0.5	0.4285
AQE_ROC	1	20	0.6756	0.6289	0.5807	0.5389	0.5015	0.4266
AQE_ROC	1	30	0.6933	0.6356	0.5837	0.5389	0.4911	0.4325
AQE_ROC	3	5	0.6667	0.6222	0.5526	0.5189	0.4578	0.385
AQE_ROC	3	10	0.6756	0.6378	0.5778	0.5222	0.4667	0.3887
AQE_ROC	3	20	0.6533	0.6178	0.5837	0.5278	0.4674	0.3823
AQE_ROC	3	30	0.6667	0.6178	0.5881	0.5322	0.4689	0.3856
AQE_ROC	5	5	0.6756	0.6156	0.5704	0.5122	0.4563	0.3699
AQE_ROC	5	10	0.6756	0.6067	0.5674	0.5167	0.4622	0.3812
AQE_ROC	5	20	0.6533	0.5867	0.5526	0.5056	0.4496	0.3734
AQE_ROC	5	30	0.6533	0.5622	0.5244	0.4989	0.4481	0.3636
AQE_ROC_Lin	1	5	0.64	0.5956	0.5659	0.5289	0.4726	0.4086
AQE_ROC_Lin	1	10	0.68	0.6133	0.5748	0.5311	0.4793	0.4158
AQE_ROC_Lin	1	20	0.6711	0.6044	0.5659	0.53	0.4822	0.4135
AQE_ROC_Lin	1	30	0.6889	0.62	0.5793	0.5433	0.4926	0.4179

Continued on next page

Table B.2 – continued from previous page

IDF: AQE optimisation table.

AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_ROC_Lin	3	5	0.6444	0.5733	0.5304	0.4867	0.4356	0.392
AQE_ROC_Lin	3	10	0.6933	0.6133	0.5659	0.5078	0.4593	0.3983
AQE_ROC_Lin	3	20	0.6267	0.6067	0.5556	0.5033	0.4526	0.3773
AQE_ROC_Lin	3	30	0.6622	0.6133	0.5659	0.5189	0.4637	0.3807
AQE_ROC_Lin	5	5	0.6267	0.5511	0.5096	0.4644	0.4259	0.3731
AQE_ROC_Lin	5	10	0.6889	0.6	0.5511	0.5056	0.4519	0.386
AQE_ROC_Lin	5	20	0.6489	0.6178	0.5763	0.5311	0.4607	0.3777
AQE_ROC_Lin	5	30	0.6533	0.6156	0.5704	0.5289	0.4681	0.3738
AQE_ROC_Log	1	5	0.6533	0.6	0.5496	0.5089	0.4711	0.4064
AQE_ROC_Log	1	10	0.6711	0.6067	0.5615	0.5233	0.4867	0.4118
AQE_ROC_Log	1	20	0.6667	0.5933	0.56	0.5267	0.4926	0.4148
AQE_ROC_Log	1	30	0.6667	0.5956	0.5585	0.5289	0.4941	0.4157
AQE_ROC_Log	3	5	0.6844	0.5911	0.5378	0.4922	0.4481	0.3876
AQE_ROC_Log	3	10	0.6667	0.5844	0.5378	0.4933	0.4407	0.3772
AQE_ROC_Log	3	20	0.6533	0.5733	0.5348	0.4811	0.4393	0.3776
AQE_ROC_Log	3	30	0.6622	0.58	0.5333	0.4856	0.4393	0.3804
AQE_ROC_Log	5	5	0.64	0.5578	0.5081	0.4678	0.4133	0.3637
AQE_ROC_Log	5	10	0.6178	0.5422	0.5037	0.4656	0.4044	0.3588
AQE_ROC_Log	5	20	0.6267	0.54	0.5037	0.4644	0.4052	0.3536
AQE_ROC_Log	5	30	0.6267	0.54	0.5111	0.4656	0.4059	0.3526
RM3	1	5	0.6444	0.5778	0.5541	0.5011	0.4504	0.4002
RM3	1	10	0.6667	0.5933	0.5615	0.5222	0.48	0.4137
RM3	1	20	0.6933	0.6178	0.5778	0.5344	0.4889	0.4205
RM3	1	30	0.6933	0.6178	0.5733	0.5344	0.4881	0.418
RM3	3	5	0.68	0.6089	0.5674	0.5078	0.4674	0.4187
RM3	3	10	0.6489	0.6044	0.5704	0.5167	0.4711	0.4028
RM3	3	20	0.6711	0.6267	0.5956	0.5356	0.4822	0.4113
RM3	3	30	0.6489	0.5978	0.5659	0.5144	0.4778	0.3944
RM3	5	5	0.68	0.58	0.5585	0.5267	0.4793	0.4105
RM3	5	10	0.6356	0.6022	0.557	0.5211	0.4726	0.3895
RM3	5	20	0.6844	0.6244	0.5881	0.55	0.4807	0.4031
RM3	5	30	0.6489	0.5978	0.5644	0.5244	0.4822	0.3878

Table B.3: This table shows the results obtained for each considered configuration of AQE approached on the training set (First 90 topics out of all 225 topics available)

DFR: AQE optimisation table.

AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_IDF	1	5	0.6044	0.5689	0.5437	0.5156	0.4844	0.4043
AQE_IDF	1	10	0.6133	0.5733	0.5437	0.5156	0.4785	0.4021
AQE_IDF	1	20	0.6133	0.5711	0.5481	0.5178	0.4793	0.4082

Continued on next page

Table B.3 – continued from previous page

DFR: AQE optimisation table.

AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_IDF	1	30	0.6178	0.5644	0.5526	0.5267	0.4844	0.4108
AQE_IDF	3	5	0.6044	0.5578	0.5556	0.5244	0.4644	0.3957
AQE_IDF	3	10	0.6089	0.56	0.5541	0.52	0.4622	0.396
AQE_IDF	3	20	0.6	0.5689	0.5422	0.51	0.4696	0.4077
AQE_IDF	3	30	0.6178	0.5644	0.5407	0.5067	0.4563	0.401
AQE_IDF	5	5	0.6089	0.5711	0.5467	0.51	0.46	0.3705
AQE_IDF	5	10	0.5733	0.5667	0.5348	0.5044	0.457	0.3945
AQE_IDF	5	20	0.5511	0.5289	0.5126	0.4856	0.4496	0.3907
AQE_IDF	5	30	0.5378	0.5156	0.4978	0.4889	0.4496	0.3876
AQE_IDF_Lin	1	5	0.6133	0.5911	0.5778	0.5522	0.4985	0.4099
AQE_IDF_Lin	1	10	0.6444	0.6022	0.5704	0.5489	0.4978	0.4134
AQE_IDF_Lin	1	20	0.6667	0.6111	0.5778	0.5533	0.5044	0.4224
AQE_IDF_Lin	1	30	0.6756	0.6044	0.5719	0.5467	0.4941	0.4204
AQE_IDF_Lin	3	5	0.6444	0.5711	0.5526	0.5178	0.4659	0.4007
AQE_IDF_Lin	3	10	0.6311	0.5667	0.5244	0.4878	0.4526	0.3867
AQE_IDF_Lin	3	20	0.6	0.5711	0.5348	0.4944	0.4511	0.3892
AQE_IDF_Lin	3	30	0.6444	0.6222	0.5807	0.54	0.4756	0.416
AQE_IDF_Lin	5	5	0.6533	0.5711	0.5304	0.5078	0.4556	0.3763
AQE_IDF_Lin	5	10	0.6622	0.6067	0.5511	0.5122	0.4637	0.3836
AQE_IDF_Lin	5	20	0.6711	0.6156	0.6015	0.5611	0.4926	0.4047
AQE_IDF_Lin	5	30	0.64	0.6133	0.5778	0.5433	0.4822	0.4005
AQE_IDF_Log	1	10	0.6267	0.6044	0.5852	0.5633	0.5119	0.4155
AQE_IDF_Log	1	20	0.64	0.5756	0.5541	0.5389	0.4904	0.4018
AQE_IDF_Log	1	30	0.6356	0.5756	0.5437	0.5267	0.4793	0.4027
AQE_IDF_Log	3	5	0.64	0.5733	0.5541	0.5089	0.4615	0.3949
AQE_IDF_Log	3	10	0.6533	0.5911	0.5526	0.51	0.4748	0.404
AQE_IDF_Log	3	20	0.6533	0.5911	0.557	0.5167	0.4704	0.408
AQE_IDF_Log	3	30	0.6444	0.5933	0.5615	0.5244	0.4793	0.4132
AQE_IDF_Log	5	5	0.6667	0.5689	0.5467	0.5144	0.4563	0.3787
AQE_IDF_Log	5	10	0.6667	0.5889	0.5407	0.5044	0.4607	0.3941
AQE_IDF_Log	5	20	0.6622	0.6022	0.5659	0.5344	0.4844	0.4081
AQE_IDF_Log	5	30	0.6667	0.5956	0.5659	0.5278	0.4704	0.4045
AQE_ROC	1	5	0.6222	0.5867	0.5704	0.5411	0.4911	0.4168
AQE_ROC	1	10	0.6178	0.58	0.56	0.5322	0.4911	0.4152
AQE_ROC	1	20	0.6311	0.5911	0.5719	0.5422	0.4993	0.4218
AQE_ROC	1	30	0.6267	0.5956	0.5733	0.5456	0.5007	0.4225
AQE_ROC	3	5	0.6267	0.5911	0.5822	0.5422	0.4652	0.3972
AQE_ROC	3	10	0.6444	0.5956	0.5807	0.5311	0.4615	0.4015
AQE_ROC	3	20	0.6267	0.5911	0.56	0.5178	0.4541	0.4003
AQE_ROC	3	30	0.6133	0.5867	0.5674	0.5178	0.4474	0.3854
AQE_ROC	5	5	0.6089	0.5778	0.5556	0.5156	0.4548	0.3741
AQE_ROC	5	10	0.5911	0.5667	0.5422	0.51	0.4548	0.3715
AQE_ROC	5	20	0.5822	0.5533	0.5304	0.5078	0.4556	0.3729

Continued on next page

Table B.3 – continued from previous page

DFR: AQE optimisation table.

AQE Approach	MaxTerms	MaxDocs	P@5	P@10	P@15	P@20	P@30	Map
AQE_ROC	5	30	0.5911	0.5644	0.5215	0.5022	0.4533	0.3696
AQE_ROC_Lin	1	5	0.6267	0.5978	0.5748	0.5467	0.4874	0.4133
AQE_ROC_Lin	1	10	0.6756	0.6156	0.5615	0.5333	0.4822	0.415
AQE_ROC_Lin	1	20	0.6756	0.6022	0.5733	0.5489	0.4904	0.4113
AQE_ROC_Lin	1	30	0.6711	0.6	0.5748	0.5544	0.4911	0.4019
AQE_ROC_Lin	3	5	0.6356	0.5822	0.5467	0.5044	0.4622	0.3924
AQE_ROC_Lin	3	10	0.6311	0.5689	0.523	0.4944	0.4622	0.3897
AQE_ROC_Lin	3	20	0.6578	0.6089	0.5763	0.5178	0.4622	0.3999
AQE_ROC_Lin	3	30	0.6578	0.6089	0.5941	0.5467	0.4896	0.4121
AQE_ROC_Lin	5	5	0.6356	0.5644	0.5304	0.4922	0.4437	0.3696
AQE_ROC_Lin	5	10	0.6844	0.6067	0.5748	0.5267	0.4674	0.3992
AQE_ROC_Lin	5	20	0.6533	0.6067	0.5911	0.54	0.4748	0.3972
AQE_ROC_Lin	5	30	0.6533	0.6111	0.5793	0.5422	0.4733	0.3942
AQE_ROC_Log	1	5	0.6267	0.6022	0.5526	0.5211	0.4674	0.4037
AQE_ROC_Log	1	10	0.6444	0.5978	0.5556	0.5222	0.4704	0.4088
AQE_ROC_Log	1	20	0.6667	0.6022	0.5659	0.5378	0.4859	0.4181
AQE_ROC_Log	1	30	0.6622	0.6	0.5644	0.5389	0.4867	0.4185
AQE_ROC_Log	3	5	0.6622	0.5978	0.5704	0.5289	0.4748	0.4069
AQE_ROC_Log	3	10	0.6711	0.6178	0.5896	0.5433	0.4733	0.4168
AQE_ROC_Log	3	20	0.6711	0.6244	0.5867	0.5367	0.4756	0.4199
AQE_ROC_Log	3	30	0.6622	0.6111	0.5689	0.5244	0.4711	0.4183
AQE_ROC_Log	5	5	0.6578	0.5844	0.5511	0.5122	0.4637	0.4023
AQE_ROC_Log	5	10	0.6444	0.58	0.5481	0.5144	0.4593	0.4009
AQE_ROC_Log	5	20	0.6444	0.5778	0.5467	0.5122	0.4615	0.401
AQE_ROC_Log	5	30	0.6444	0.5711	0.5407	0.5133	0.4622	0.4005
RM3	1	5	0.6133	0.5956	0.5719	0.5444	0.4941	0.4089
RM3	1	10	0.6622	0.6044	0.5733	0.5533	0.4993	0.4202
RM3	1	20	0.6889	0.6133	0.5822	0.55	0.4941	0.4225
RM3	1	30	0.6711	0.6	0.5748	0.5467	0.4978	0.4212
RM3	3	5	0.6311	0.5778	0.5407	0.4989	0.4607	0.3909
RM3	3	10	0.6533	0.5956	0.5496	0.5022	0.4511	0.4009
RM3	3	20	0.6489	0.6133	0.5793	0.5389	0.4756	0.4054
RM3	3	30	0.6667	0.6044	0.5704	0.5367	0.4807	0.4103
RM3	5	5	0.6267	0.5756	0.5304	0.5011	0.4452	0.3767
RM3	5	10	0.7022	0.6222	0.5733	0.5267	0.4556	0.4044
RM3	5	20	0.6844	0.62	0.6015	0.5567	0.4844	0.4118
RM3	5	30	0.6356	0.6	0.563	0.5267	0.4659	0.3901