Microblogs Structure. Challenges and Opportunities

Jesus Alberto Rodriguez Perez, The University of Glasgow
Joemon M. Jose, The University of Glasgow

In recent years, microblog services such as Twitter have gained increasing popularity, leading to active research on how to effectively exploit its content. Microblog documents such as 'tweets' differ in morphology to more traditional documents such as web pages. Particularly, tweets are considerably shorter (140 characters) than web documents and contain contextual tags regarding the topic (hashtags), intended audience (mentions) as well as links to external content (URLs). Unfortunately, state of the art retrieval models perform rather poorly in capturing the relevance of microblogs, due to the previously unforeseen conditions in which they operate.

In this work, first we focus on investigating the problems that state of the art retrieval models suffer when handling microblogs, then we provide a number of solutions to adapt to the new medium. Initially we simulate the behaviour of such retrieval models under microblog retrieval conditions. Based on our findings we devised a retrieval model, namely **MBRM**, which significantly outperforms the state of the retrieval models in the microblog context.

Furthermore we look at microblog documents as a high-dimensional entity and study the structural differences between those documents which are deemed relevant against those non-relevant. Moreover we leverage these statistical differences in experiments to enhance the behaviour of retrieval models. Additionally we study the interactions between the different dimensions in terms of their order within the documents by modelling relevant and non-relevant tweets as state machines. These state machines are then utilised to produce scores which in turn are used for re-ranking. Our evaluation results show statistically significant improvements over the baseline in terms of precision at different cut-off points for both approaches. These results confirm that the relative presence of the different dimensions within a document and their ordering are connected with the relevance of microblogs.

Categories and Subject Descriptors: C.2.2 [**Computer-Communication Networks**]: Network Protocols

General Terms: Microblog, State machines, Classification

Additional Key Words and Phrases: Information Retrieval, Structural Models, Document Dimentions

ACM Reference Format:
Jesus Rodriguez et al., 2017. Microblogs Structure. Challenges and Opportunities. *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 39 (March 2010), 32 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Microblogs have grown in popularity in recent years, gradually transforming the way we find out about the latest events and communicate. Twitter is the most prominent service [1], as it is used by millions, posting over 340 million tweets every day[2]. Microblog services are used for various purposes including: (i) self promotion, (ii) advertising, (iii) real-time news broadcasting, (iv) social discussions etc. The most important aspect of

[1] https://twitter.com/

[2] http://blog.twitter.com/2012/03/twitter-turns-six.html

DOI:http://dx.doi.org/10.1145/0000000.0000000

Twitter is that it provides unique insight into real-time events, such as first hand reports of events as they are developing, along with the opinion of those discussing them. This information makes Twitter a uniquely valuable media source, which led to obtaining much attention by research and industrial communities.

Retrieving documents in Twitter can be extremely challenging because of their morphology. The content is limited to 140 characters per messages (known as *Tweets*). This constraint leads to varied linguistic quality [Teevan et al. 2011] due to colloquialisms and users efforts to fit their content within the limitations. More importantly tweets pose new challenges for which state of the art retrieval models were not designed for[3].

Whilst few recent works have identified some features as possibly being detrimental in microblog ad-hoc retrieval [Naveed et al. 2011], no study has been carried out to determine the concrete effect of the different features on state of the art retrieval models. Therefore we are set to investigate the connection of the structure of microblog documents with their relevance during an ad-hoc search task. This whole work revolves around the following main question:

> **What are the reasons behind the underperformance of state of the art retrieval models in the context of microblogs? And what can we do about it?**

To this end, firstly we observe the performance of state of the art retrieval models in the context of Twitter corpora selecting the best retrieval model as a baseline. Then we perform a series of experiments which simulate the behaviour of a number of state of the art retrieval models in order to identify possible shortcomings in their design with respect to microblog documents. This initial experiment is completed with the creation of a retrieval model which takes into account all previous findings, namely MBRM. MBRM demonstrates that the scope hypotheses still holds within microblog documents, and that microblog document statistics can be leveraged to significantly improve ad-hoc retrieval performance.

Secondly, we study the behaviour of inherent features of microblog documents and evaluate their suitability for enhancing the behaviour of state of the art retrieval models. Moreover, we demonstrate which microblog specific features are most indicative of the relevance of microblogs by reporting statistically significantly improved retrieval performance for ad-hoc search when taking them into account.

Finally, we extend our analysis by considering the ordering of the different component that make up microblog documents. In order to do so, we encode the structure of observed relevant and non-relevant documents into state machines. These are in turn used to produce scores for re-ranking. We utilise the 2013 microblog collection to construct such state machines, and we test on the 2011 and 2012 microblog collections combined. Our results show statistically significantly improved results over the selected baseline, demonstrating the connection of microblog structure with relevance.

This work will be driven by the following research questions:

RQ1. Are there structural differences between relevant and non-relevant microblog documents? Can we exploit them for ad-hoc retrieval?
RQ1.1 Does document length have any connection with the relevance of microblogs?
RQ1.2 Does term frequency of query terms relate to the relevance of microblogs?
RQ1.3 Can we adapt state of the art retrieval models to better handle microblogs?

[3] 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]

RQ1.4 Can we devise a retrieval model to better capture the relevance of microblogs?
RQ2. Can microblog features be exploited to help retrieval models better capture relevance than current retrieval models?
RQ3. Is the order of the different elements in a microblog document connected with relevance? Can it be utilised for ad-hoc retrieval?

The rest of the chapter is organised as follows. First, we cover relevant literature regarding microblog searches and introduce the concepts utilised throughout this work (Section 2). Section 3 sets the evaluation environment in which our experimentation is carried out, giving way to our main analysis (Section 6). Finally Section 7 concludes the work and points future research directions.

2. BACKGROUND

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

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 4, and their behaviour described individually against microblog conditions. However we first introduce some basic background to ease the understanding of the following sections.

2.1.1. 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.

2.1.2. Document length normalization. [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 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 adding more relevant information to the topic and occupying more space. BM25 applies a soft normalisation that takes into account both cases.

2.2. Retrieval of Microblogs is Hard

Retrieval models are designed to rely on term frequency and document length as the variables to quantify whether a document is more important than other. From a very

Table I: 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

simplified perspective, a retrieval model will give more importance to a document that contains query terms more frequently than another document (Assuming similar document lengths). Likewise, when query terms appear the same number of times, a document will be deemed less or more informative based on the document lengths.

However, as stated before, microblog documents are limited in length to 140 characters in the case of Twitter. This limitation obviously challenges the abovementioned assumptions, which unfortunately form the basis of the workings of most retrieval models in a way or another.

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.

2.3. TREC Microblog Retrieval Tracks

TREC organised a number of 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 gathered in 2013 and was used for both the 2013 and 2014 microblog tracks. Finally the number of topics varied between 50 and 60, but are 225 in total.

The summary results for each of the tracks are presented in Table I 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 [Kim et al. 2012; Aboulnaga et al. 2012; Han 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 [Siming Zhu 2013; Gao et al. 2013], query expansion [Pre-Processing 2013; Perez 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.

Work by [Thomas 2012] studied the effects that preprocessing had on retrieval performance. Their findings showed that the best performance was achieved 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*.

2.4. Making Sense Of Microposts

The MSM workshop [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, Ke et al. [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. In our work, we pay attention to the same features, but from a different angle, by looking how much space relative to the total characters in the document is dedicated to each of the microblogs elements.

2.5. Other Microblog retrieval features

Work by [Massoudi et al. 2011] explored the use of other features to improve ad-hoc retrieval. 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 enhance ad-hoc retrieval performance. While all these works attempt to exploit some microblog features or augment them with external resources, they do not try 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 use those features that seem beneficial to improve the behaviour of a state of the art retrieval model.

2.6. Understanding Microblogs

We believe that 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. We believe it is important to explore, and properly assess the interaction of such features. Better understanding could lead to better performance of retrieval models, or new models altogether, which are the starting point for many techniques commonly used in microblog retrieval (E.g. Automatic Query Expansion).

3. EXPERIMENTAL SETTING

Datasets. In this evaluation we have used the four collections 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. However, the 2013 and 2014 relevance assessments are statiscally comparable to the 2012 track. Moreover, the ratio of documents $\frac{relevant}{non-relevant}$ is much higher for the 2013, which can result in generally better retrieval performance than previous tracks by default. The 2014 on the other hand is closer in this ratio to the 2012 collection. In fact it has a considerably lower number of relevant documents per topic.

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 II for an extended overview of these collections.

Table II: 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

Table III: 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**

Evaluation measures. We pay attention to precision at different ranks, with a maximum cut-off point at rank 100. 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 [4].

Baseline selection. Table III contains evaluation results for the considered state of the art retrieval models when applied to Twitter corpora from the 2011, 2012 and 2013 Trec microblog collections. The models considered in this evaluation are TF-IDF (IDF)[5], 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.

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 promoted over shorter ones should not be normalised w.r.t their length.

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 in their work 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 is just how it is interpreted in this particular case what makes it harmful.

Intuition tell us that document length normalization as we know it does not interact well with the limitations characterised by microblogs. The **Verbosity** and **Scope** hypotheses seem not to model the behaviour of users publishing microblogs. Microblog users generally have the challenge of fitting 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.

To aid in developing our understanding of the behaviour of retrieval models we formalise their composition. To this end we have compiled Table IV to show 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

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

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

statistics (I.e. Collection feature) or the document itself (Document feature). The second row contains acronyms for each of the features, which are expanded as:

- **AverageDocumentLength (ADL):** This is the average document length, in number of tokens, for the whole collection.
- **DocumentLength (DL):** This is the document length, in number of tokens, for the document being scored.
- **NumberOfDocuments (ND):** Total number of documents in the collection.
- **DocumentFrequency (DF):** Number of documents in which the term appears (I.e. A term's posting list size).
- **NumberOfTokens (NT):** Number of different tokens in the collection.
- **CollectionTermFrequency (CTF):** Frequency of a term in the whole collection. (I.e. Total number of occurences of a term in the collection)
- **TermFrequency (TF):** Frequency of the term in the document being evaluated.

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

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

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 NumberOfTokens (NT), CollectionTermFrequency (CTF), TermFrequency (TF) and Document Length (DL).

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 IV 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:

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

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

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

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 is 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 3.

$$\begin{aligned}
\text{BM25}(D,Q) &= \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{\text{avgdl}})} \\
&= \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(0+1)}{f(q_i,D)+0\cdot(1-0+0\cdot\frac{|D|}{\text{avgdl}})} \\
&= \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)}{f(q_i,D)} \\
&= \sum_{i=1}^{n} \text{IDF}(q_i) \qquad (3)
\end{aligned}$$

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[6].

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 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 1 shows the possible BM25 scores for a range of Term Frequency (TF) and Doc. Length (DL) values.[7]. 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

[6] We later demonstrate that **scope** does hold, but not **verbosity**

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

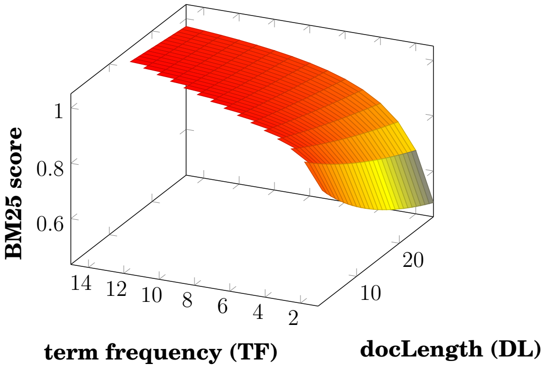


Fig. 1: Term Frequency (TF) vs, Doc. Length (DL)

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.

4.2. The Hiemstra's Language Model (HLM) Case

Table IV 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 III 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:

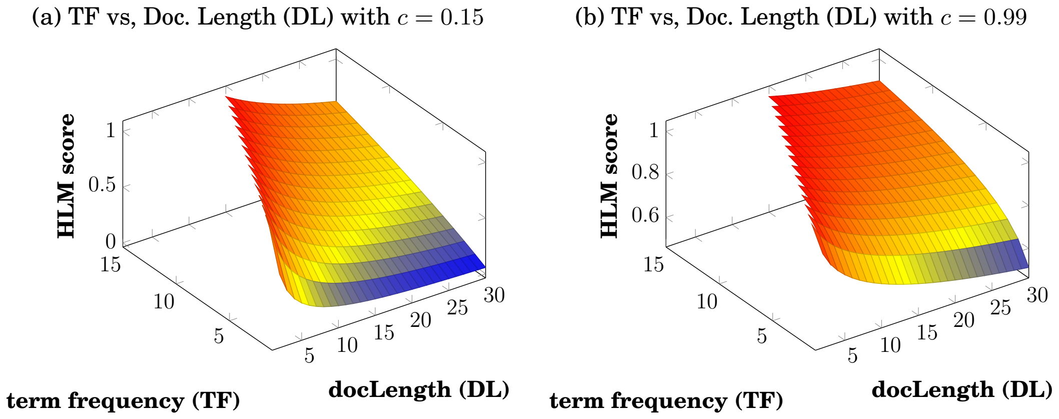
$$\text{HLM}(D,Q) = \sum_{i=1}^{n} \log_2 \left[1 + \frac{c \cdot f(q_i, D) \cdot ntoks}{(1-c) \cdot f(q_i, C) \cdot |D|} \right] \tag{4}$$

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.

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 2a shows a plot of the possible scores produced by HLM in its default configuration ($c = 0.15$)[8]. 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.

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



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

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

Fig. 2: HLM analysis

Furthermore, we can observe in Formula 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)[9].

Table III 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 de differences in the scoring, should yield improved retrieval performance.

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 V 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 only offsetting 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 in-

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

Table V: 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**

crementing 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 V 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 2a and 2b which show scores produced by HLM w.r.t. TF and DL with different values of c. Figure 2a sets $c = 0.15$ whereas Figure 2b sets $c = 0.99$. It is easily observed how Figure 2a shows more differences across the spectrum of scores with respect to TF and DL than Figure 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 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.

4.3. The DLM Case

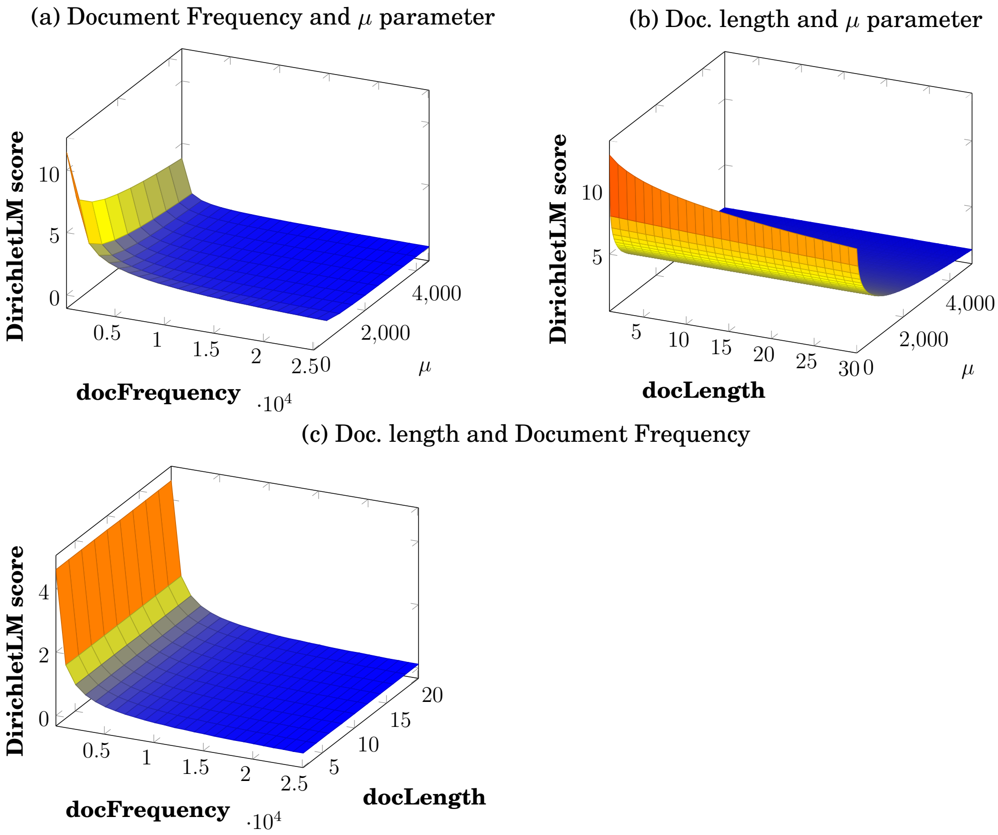
Dirichlet Smoothed language model (DLM), was the baseline retrieval model for the 2013 and 2014 instances of the microblog track. DLM was used within the "Microblog track as a service" client which managed a Lucene index in its core. DLM has a smoothing parameter named μ, which was set to 2500 by default during the 2013 and 2014 microblog tracks. Moreover, DLM scores are produced [10] by the following equation:

$$\text{DLM}(D,Q) = \sum_{i=1}^{n} \log_2 \left[1 + \frac{f(q_i,D)}{\mu \cdot \frac{f(q_i,C)}{ntoks}} \right] + \log_2 \left[\frac{\mu}{|D| + \mu} \right] \tag{5}$$

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 3a and 3b show DLM scores in terms of the μ parameter, w.r.t. document frequency and document length respectively. Figure 3c on the other hand demonstrates the relation between document frequency and document length.

[10] 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

Fig. 3: DLM evaluation figures

As we can observe from Equation 4.3 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 3a. Similarly, we can observe in Figure 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 3c, contrary to other retrieval models.

This could be a desired feature for microblog retrieval, however let us look at the performance achieved for a range of μ values in Table VI. 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.

Table VI: 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

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

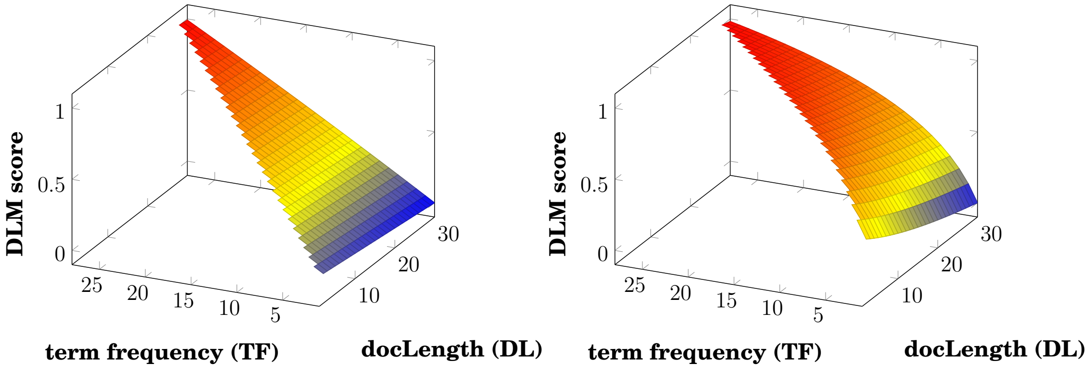
μ	***dTF***	***dDL***	***P@30***
20			0.4241
20	20		0.4558
20		20	0.3901
20	20	20	0.4547
2500			0.2988
2500	20		0.4468
2500		20	0.2892
2500	20	20	0.4466

The results of the evaluation are presented in Table VII. 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 4a and 4b. Figures 4a and 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 VII in connection with Figures 4a and 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 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 VII 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 4a shows an almost linear progression of DLM scores with respect to TF, whereas Figure 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 considera-



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

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

Fig. 4: Evaluating DLM's behaviour

tion 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 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 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.

4.4. The DFRee Case

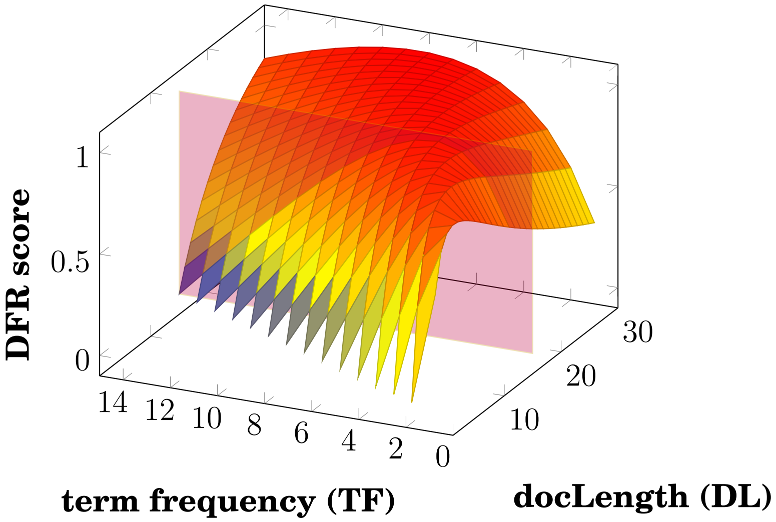
DFRee[11] is a Divergence From Randomness model implemented in the Terrier IR platform [Ounis et al. 2006]. DFRee has been designed as a parameter-free model and adheres to the following implementation:

$$prior = \frac{f(q_i, D)}{|D|}, posterior = \frac{f(q_i, D) + 1}{|D| + 1} \tag{6}$$

$$InvPriorColl = \frac{ntoks}{f(q_i, C)}, norm = f(q_i, D) * log_2 \frac{posterior}{prior} \tag{7}$$

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

Fig. 5: Evaluating DFR's behaviour: Doc. length (DL) and Term Frequency (TF)



$$
\begin{aligned}
DFRee(q_i, D, C) = norm * [\\
&f(q_i, D) * (-log_2(prior * InvPriorColl))\\
&+ (f(q_i, D) + 1) * log_2(posterior * InvPriorColl)\\
&+ 0.5 * log_2(posterior/prior)], \quad (8)
\end{aligned}
$$

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 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 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 III).

We believe that the *saturation point* observed in Figure 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 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.

5. TOWARDS A MICROBLOG RETRIEVAL MODEL

In this section firstly we further extend our experimentation on the above-mentioned retrieval models. Secondly we gather all our findings to produce a retrieval model specifically tailored to microblog retrieval.

5.1. Score differences and Harmonisation

So far we have introduced a set of representative retrieval models, and discussed how they behave when facing microblog-like conditions. We have mainly done so by simulating the scores produced by each model, when fixing all parameters except TF and DL which are the variables to be considered. In the above-mentioned experiments, we have observed a very interesting relationship between TF, DL and the score given to the terms. In most retrieval models performance seems to increase when we overestimate the values of TF and DL, thus forcing the models into an area of values where the score differences with respect to TF and DL are much lower.

Table VIII holds a summary of the results for all retrieval models in their various configurations with respect to Precision@30. Additionally the third column holds the standard deviation of the simulated scores produced by the retrieval models. As it can be easily observed that the possible document scores are much closer together for those configurations that improve a retrieval model's performance. In fact there seems to be a strong negative statistical correlation between the standard deviation and the performance achieved by the retrieval models.

In other words, reducing the standard deviation of the possible scores for most of the retrieval models is connected with significantly better performance. These observations motivate the following hypothesis:

Table VIII: Behaviour when harmonising score differences.

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
DFRee	NA	0.2268	0.4614
PEARSON	-0.70 or -0.58		
KTau	-0.66 or -0.5555		

Table IX: 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%

"The range of scores produced by retrieval models when ranking microblogs can be unfairly different due to the retrieval model's behaviour with respect to scarce TF and DL values".

If this hypothesis is true, we should be able to achieve similar results using a different technique to reduce the standard deviation of the scores produced by the different retrieval models. To this end we produced the results in Table IX. This Table holds performance metrics for all retrieval models with their standard configurations, however each of the scores computed for each document have been normalised using a logarithm base 2.

As an example, the formulation for HLM would look as follows:

$$\mathrm{HLM}(D, Q) = \sum_{i=1}^{n} log_2 \left[\log_2 \left[1 + \frac{c \cdot f(q_i, D) \cdot ntoks}{(1-c) \cdot f(q_i, C) \cdot |D|} \right] \right] \tag{9}$$

As we can observe in Table IX the results for DLM, HLM and BM25 are significantly better than standard (Table VIII), whereas DFRee performs marginally worse than its default form and IDF remains unaffected.

We can conclude that based on this evidence most retrieval models are not prepared to effectively capture the relevance of microblogs. The verbose and scope hypotheses, which serves as inspiration to most retrieval models, do not hold for microblog documents. Additionally, the main reason points to their over-sensitiveness to low values of term frequency and document length. This sensitiveness often produces a high degree of score differences amongst the ranked documents which ultimately negatively affects performance.

Table X: 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

5.2. MBRM: A Microblogs Retrieval Model

In previous sections, we have presented a number of problems faced by retrieval models when dealing with microblogs. We have shown empirical evidence of their existence by improving the performance of state of the art retrieval models. However we can further investigate these issues by devising a retrieval model, which relies on what we have learnt so far about microblogs. To this end, we would like to introduce a "MicroBlogs Retrieval Model", namely MBRM.

MBRM is made up of two main components which deal with document based information attached 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. In fact we came up with a formulation to represent the behaviour we observed as being the best for microblog retrieval. The first component deals with the document length and is given by the following logistic distribution:

$$DLComp(DL) = \frac{c_1}{1 + a_1 \mathrm{e}^{-b_1 DL}} \tag{10}$$

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:

$$TFComp(TF) = a_2 e^{-\frac{(TF-b_2)^2}{2c_2^2}} \tag{11}$$

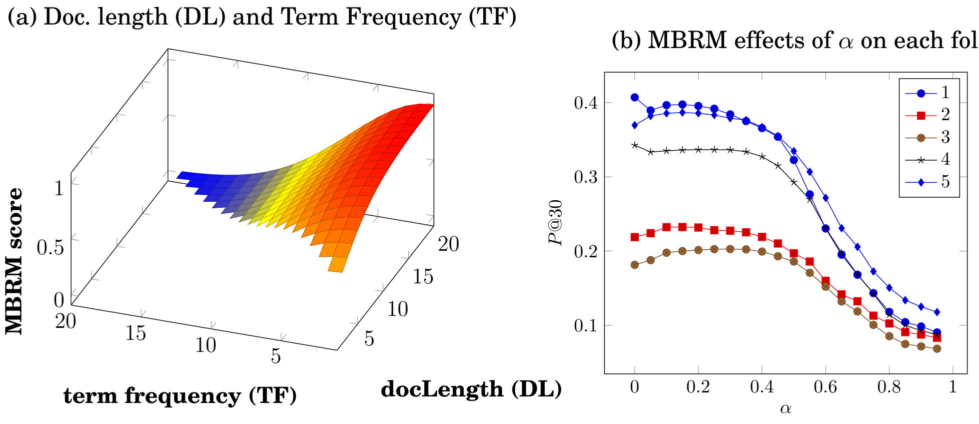
where a_2, b_2 and c_2 are parameters similar parameters to those found in the previous function. These functions were chosen as they offer good control over the curves, and their values can be bound between 1 and 0 so we do not need to normalise values when combining them. The final formulation for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

$$MBRM(D,Q) = \sum_{i=1}^{|Q|} (1-\alpha) * \mathbf{IDF}(q_i) + \alpha * DLComp(|D|) * TFComp(q_i) \tag{12}$$

which can be also expressed as:

$$MBRM(D,Q) = \sum_{i=1}^{|Q|} (1-\alpha) * \mathbf{IDF}(q_i) + \alpha * \left(\frac{c_1}{1 + a_1 \mathrm{e}^{-b_1 DL(|D|)}} \right) * \left(a_2 e^{-\frac{(TF(q_i)-b_2)^2}{2c_2^2}} \right) \tag{13}$$

Figure 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 Table X. In Figure 6a we can observe how the values obtained on the TF axis decrease slowly


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

(b) MBRM effects of α on each fold.

Fig. 6: MBRM: A Microblog Retrieval Model

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 in which the highest importance is given to TF when near the average 2 and then it is reduced as it increases. High TF values are most likely than not associated with spam or unimportant documents, since actual users struggling to fit their content in the 140 characters limit are unlikely to repeat words. Although this is not always the case, thus the slow descent for low values of TF.

Now we pay attention to the values in terms of the DL axis. We can observe that they increase in a soft slope as we traverse DL. Unlike $DFRee$, the slope is always incremental. The idea behind is that no document is significantly more important when it contains a single word more. However it should be more important since the more terms in a document the more comprehensive the document should be, particularly in terms of the amount of information encoded in it, regardless of the character limitation.

Figure 6b shows the results of parameter optimisation by means of a 5-fold cross-validation. The whole set of topics is subdivided into 5 groups where 4 are used for training and one for testing. The roles of each group are alternated resulting in 5 different experiments. It can very easily be observed that the most optimal values for the mixing parameter α are near 0.20.

Table XI shows the performance results obtained for MBRM in terms of Precision at different levels with respect to 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. In other words, the authors of the documents will attempt to encode as much information as possible even with the obvious document limitations. This contradicts our findings in Subsection 4.1 however we believe that in the particular case of BM25, document length has a much more aggressive effect on the scores, thus resulting in a misleading behaviour.

The verbose hypotheses however seems not to hold, as authors are very careful to come up with specific words to effectively encode their message. 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. And this is what is ultimately captured by our MBRM retrieval model.

Final note, the previous experiment where we apply a logarithmic function to the scores of the retrieval models, reduce the effect of the possible values of DL. This can also be interpreted as being closer to the behaviour of MBRM where increasing DL also increases the score, which provides the best experimental results.

6. UNDERSTANDING MICROBLOG DOCUMENTS

In this Section we first study the structure of microblog documents in order to define a hypotheses that captures their relevance. Subsequently, we test our hypotheses through the implementation of a number of approaches that capture microblogs' structure and their evaluation with respect to our DFR baseline. Additionally, we evaluate the relation of the order of the different dimensions within the microblogs, and determine how to utilise this evidence for ad-hoc retrieval.

6.1. Informativeness of Microblogs

For web and similar documents, relevance is modelled by the inclusion of statistical measures extracted both from the collection as a whole, and the documents themselves. Most retrieval models take into consideration document based statistics, such as document length and term frequency, in an attempt to capture the relevance of the documents according to the scope and verbosity hypotheses (or similar assumptions). For the purposes of this work, we can think of each retrieval model as a delicate relationship "$\boxed{?}$" between document length $|D|$ and term frequency $P(q \cap D|Q)$ amongst other components. We pay attention to those components as they are most likely affected by the structure of microblogs.

$$P(I|Q, D) = |D| \boxed{?} P(q \cap D|Q) \tag{14}$$

Microblog documents are however very short as they have a fixed maximum size. Additionally, authors tend to optimise their content to fit within the character limits and constraints set by the platform, leading to a more or less constant document length (~ 15 terms in the case of Twitter). Moreover, due to these limitations, the value of term frequencies revolve around ~ 1.5. Thus in-document statistical information is limited.

Both the **scope** and **verbosity** hypotheses are defined within the assumption that authors may write as much as they desire. As a result it is logical to assume that when this condition is broken unexpected behaviour may follow. Fortunately, microblog documents contain other inherent features which encode extra information in the same message following an organic community-agreed vocabulary. In our work we draw inspiration from the ideas behind the scope and verbosity hypotheses and we are set to

Table XI: 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***

describe a new hypotheses tailored to microblog retrieval, which highlights and relies on characteristics of microblog documents' structure.

Firstly, we assume that microblog documents **(D)** are 4-dimensional entities comprised of **Text** $T(D)$**;** a **URL** $U(D)$ (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 the amount of space in a microblog document dedicated to each of the dimensions may have a connection with how likely it is to be relevant to the searcher. Having these characteristics in mind, we define **(H1) Microblog Informativeness** (MI) as the probability for a Microblog document D being informative given a query $P(MI|Q, D)$, which depends on an optimal unobserved combination "$\boxed{?}$" of the aforementioned dimensions:

$$P(MI|Q, D) = T(D) \boxed{?} U(D) \boxed{?} \#(D) \boxed{?} @(D)\boxed{?} P(q \cap D|Q) \tag{15}$$

where $T(D)$, $U(D)$, $\#(D)$ and $@(D)$ are the ratios in terms of number of characters spent in the document for each of the dimensions considered [12]. For example, the ratio for the text dimension $T(D)$ is given by:

$$T(D) = \frac{\#ofCharsforTextDimension}{Total\#ofChars}, \tag{16}$$

In order to test our hypotheses and learn about what are the most prominent characteristics that make up relevant microblog documents, we analyse retrieval runs produced by the state of the art baseline DFR because it is the best performing model as shown in Table III. We use the documents in the runs instead of all documents in the relevance judgements in order to analyse the documents that are most likely to contain query terms and find differences amongst those documents.

We take into consideration the TREC Microblog topics 1 to 110 so that we can confirm our findings through an evaluation on the newer 111 to 170 topics which belong to TREC's 2013 iteration of the microblog search task.

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

First we look at "DocLength". As we can observe in Tables XII(a...e), the differences between relevant and non-relevant documents are not significant. Furthermore, we can see 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. It is evident that the behaviour of this feature is unstable, and the differences between both groups of documents change wildly depending on the cut-off point, contradicting each other.

Based on this observation we can conclude that the document length feature, popular amongst retrieval models, is ineffective in estimating the relevance of a microblog document. Therefore, this helps us to confirm that the scope and verbosity hypothesis

[12] URL's are automatically shortened by Twitter, thus their length is constant.

Table XII: Ratio of each dimension for relevant (Rel) and non-relevant (Non-Rel) documents at different cutoffs.

(a) Cutoff @ 10

	Rel	Non-Rel
Hash	1.960	1.619
Ment	2.750	2.444
Urls	17.32	14.16 *
Text	77.95	81.77 *
DocLength	97.47	100.2

(b) Cutoff @ 20

	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

	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

	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

	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

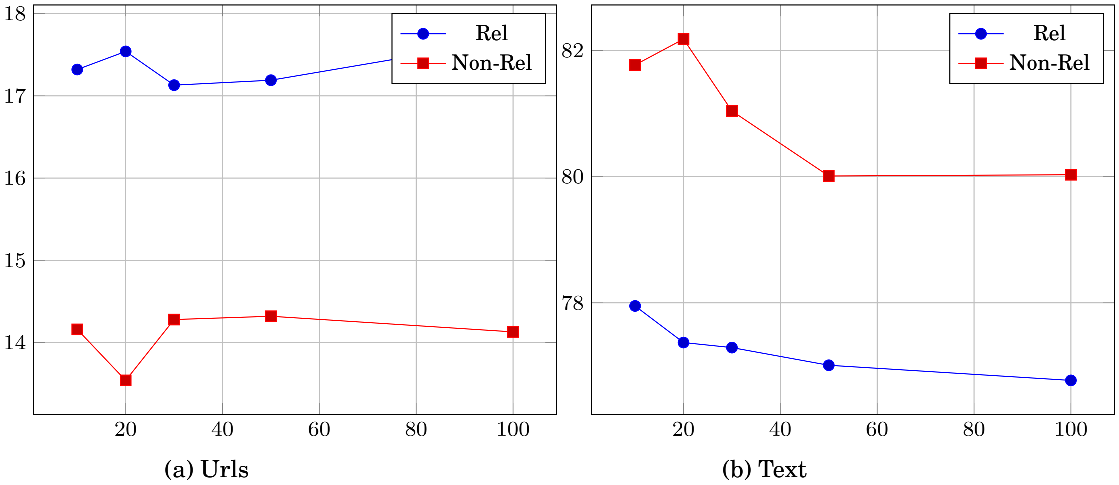
do not hold for microblog documents, as differences should have followed a more clear trend if the hypotheses were true. (i.e. One relevance group should have remained higher than the other for all cut-off cases.). Therefore we can confirm that in the case of microblog documents, longer (or shorter) does not have a connection with a document being relevant.

Secondly, we look at the **Urls** and **Text** dimensions of microblog documents in Figure 7. In the case of **Urls**, this dimension tends to be significantly larger on relevant documents than in their non-relevant counterparts. This is in line with previous works suggesting that the presence of URL's increases the likelihood for a document to be relevant [Massoudi et al. 2011]. Figure 7a shows the changes in space dedicated to the URL dimension as we go down the results list. An interesting behaviour that can be observed is that, relevant documents behave in exactly the opposite way to non-relevant documents. As we traverse the results list the space for the URL's in relevant documents increases whereas, it slowly decreases for non-relevant documents.

The **Text** dimension on the other hand, is significantly smaller for relevant documents, across all cut-offs. However, as observed in Figure 7b, the behaviour as we traverse the list towards lower cut-off 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 in the differences of both the **Urls** and **Text** dimensions make them especially interesting feature candidates to be studied, and possibly employed to improve the behaviour of retrieval systems.

Figure 9 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. Additionally, relevant documents seem to have a higher portion of the content dedicated to this dimension than non-relevant documents. This behaviour can be observed in Figure 9a, as relevant documents seem to dedicate more space for hashtags regardless of the cut-off chosen. Another observation that can be made, is that as we traverse the result list, the presence of hashtags become more pronounced for both relevant and



(a) Urls

(b) Text

Fig. 7: Rate (%) of space dedicated to Urls and Text in Relevant and Non-Relevant documents at different cut-off points.

non-relevant documents, thus the increased (or decreased) presence of hashtags does not serve as a discriminative factor in microblog ranking.

Finally, we observe the behaviour of the **Mention** dimension in Figure 9b. For the three first cut-offs @10; @20 and @30, relevant documents seem to spend more space in defining an 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. This makes sense if we assume that many non-relevant documents may be conversational in nature, instead of introducing facts interesting to a wider audience. In fact the differences in terms of the space dedicated to the **Mentions** dimension is only significant once we are much lower in the ranking at the @100 cut-off.

One could argue that our conclusions may be biased since the result lists are produced with respect to the retrieval model inherent features (e.g. document length). However, we can see that the differences in the observations between relevant and non-relevant documents for the good dimensions (Urls, Text and Hash) are relatively constant, thus independent from the rank for our purposes.

6.2. 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 8). More specifically, relevant documents tend to use less space for text, and more space to contain the URLs, and hashtag dimensions than non-relevant documents. An important note is that 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 7b. Thus we model informativeness in terms of the retrieval model score $P(q \cap D|Q)$ for document D given query Q and its Text dimension as:

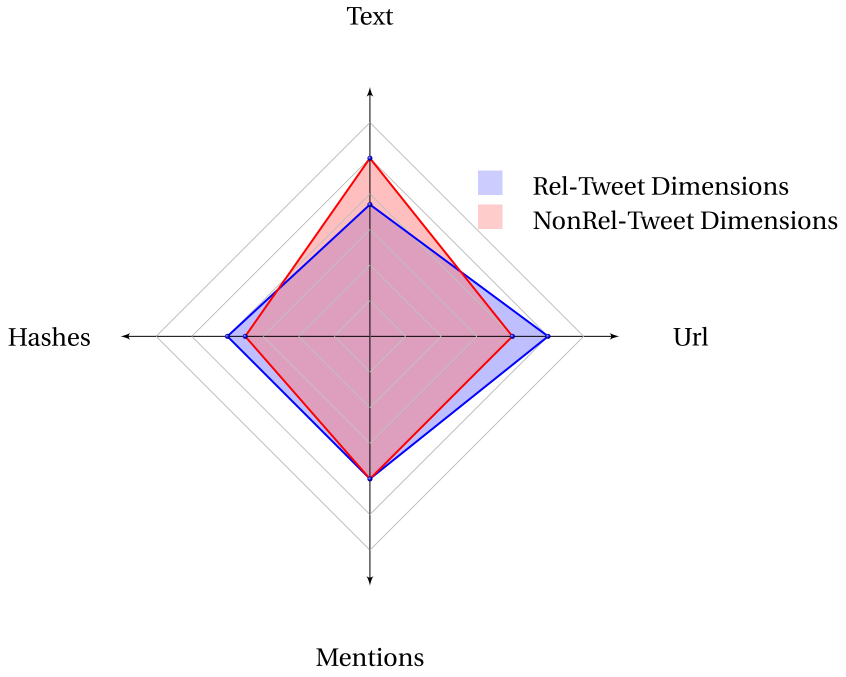


Fig. 8: Dimensional differences between relevant and non-relevant documents. Statistically significant differences are exaggerated for easier visualization.

$$P(MI|D,Q) = P(q \cap D|Q) + \lambda[1 - |T(D) - 0.76|], \tag{17}$$

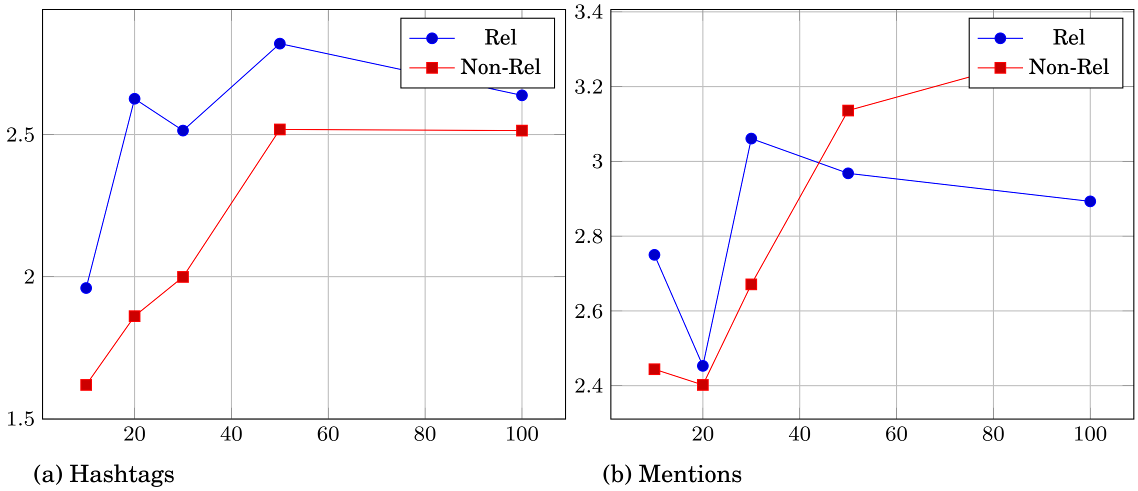
where we give a lower score to those documents diverging from the optimal text dimension rate 0.76[13]. We test this formulation using DFR to produce the $P(q \cap D|Q)$ score over the microblog 2013 collection, which was **not** used in producing the analysis results in the previous section. We retrieve the first 500 documents using DFR and re-rank them using our first model (Equation 17) with λ set to 1. The results are shown in the RR-text[14] row within Table XIII. As we can observe, the performance of DFR 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 URL dimension expressed as a rate with the score of the retrieval model as follows:

$$P(MI|D,Q) = P(q \cap D|Q) + \omega U(D), \tag{18}$$

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

[14] “RR-” stands for “Re Ranking”, and precedes the features utilised in the operation



(a) Hashtags

(b) Mentions

Fig. 9: Rate (%) of space dedicated to HashTags and Mentions in Relevant and Non-Relevant documents at different cut-off points.

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

$$P(MI|D,Q) = P(q \cap D|Q) + \lambda[1 - |T(D) - 0.76|] + \omega U(D), \tag{19}$$

The results for this combination are shown in Table XIII as row RR-text-url. Further improvements with respect to previous approaches are introduced at all cut-offs except P@10, where RR-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:

$$\begin{aligned} P(MI|D,Q) = P(q \cap D|Q) + \lambda[1 - |T(D) - 0.76|] \\ +\omega U(D) + \gamma\#(D), \end{aligned} \tag{20}$$

$$\begin{aligned} P(MI|D,Q) = P(q \cap D|Q) + \lambda[1 - |T(D) - 0.76|] \\ +\omega U(D) + \gamma\#(D) + \delta@(D), \end{aligned} \tag{21}$$

where the free parameters are set to 1[15].

[15] Parameter optimisation would be beneficial in the future, although it was not necessary to evaluate the hypothese of this work

Table XIII: Experimental results when considering different dimensions, using the 2013 TREC Microblog collection ($*p < 0.05$ over the DFR baseline).

Model	P@5	P@10	P@15	P@20	P@30
DFR	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

The results for both models (Equations 20 and 21) are shown in Table XIII as RR-text-url-hash and RR-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 RR-text-url in terms of P@30. The addition of the mentions component in RR-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 7a, 7b, 9a and 9b and Table XIII we can see how the dimensions that showed constant differences across all cut-offs 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 Figure9b follows an erratic behaviour (For earlier cut-offs more space is dedicated to the mentions in relevant documents, and then after the cut-off 40 is the opposite case).

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 which captures Microblog characteristics in the shape of a model given by Equation 20. 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 searcher. In the following subsection, we expand our analysis by taking into consideration the order of the dimensions.

6.3. 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 [16] which are meant to assist in writing the perfect tweet to grab the attention of readers.

6.3.1. State Machine Structure. To properly model such interactions is no simple task. 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 10 shows an example, denoting

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

how state 1, can transition to other states, such as state 2, with the probabilities stated above the arrows [17].

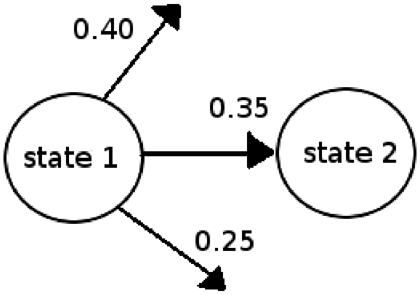


Fig. 10: State machine example.

Figures 11a and 11b 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**.

While both figures look very similar, there are some 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.

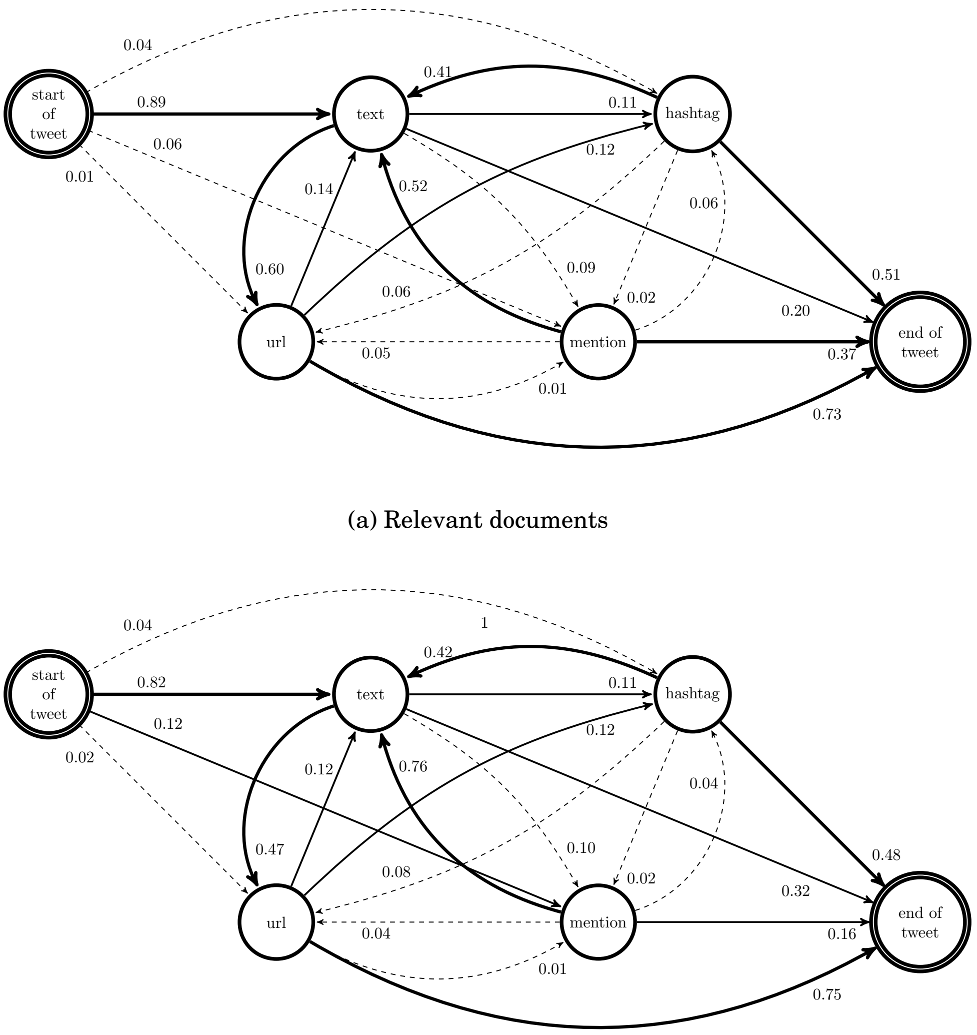
In order to test whether we can use this evidence for producing better rankings, we devised our **"State"** approach. The State approach is a re-ranking method that linearly combines the score given by any retrieval method with the aggregation of probabilities from start to end nodes w.r.t a microblog's 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 11a) 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:

$$\begin{aligned} State(D, Q) = (1 - \alpha)P(q \subset D|Q) \\ +\alpha * (R\_Score(D) - NR\_Score(D)), \end{aligned} \tag{22}$$

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



(a) Relevant documents

(b) Non-Relevant documents

Fig. 11: Tweet automatas for the 2013 collection

1

where $R\_Score(D)$ and $NR\_Score(D)$ are the scores computed by traversing the automatas in Figures 11a and 11b respectively and α is a weighting factor which balances the linear combination. 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.

Table XIV: 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
Baseline	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

Table XIV shows the retrieval results for our re-ranking approach over the 2011 and 2012 collections. P@5 to P@30 represents 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 to the very first few documents, which is a highly desirable product 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.

6.4. 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 dimensions in the transition and its associated probability based on the observed data. The model itself should be updated from time to time to accommodate any shifting in the structuring and style of micro-bloggers.

7. CONCLUSIONS

In this work, we verified whether the scope and verbosity hypotheses still hold for microblog document retrieval. We hypothesise that, since microblog documents have a fixed maximum size, the scope and verbosity hypotheses do not hold, as they assume the author of the document is able to produce documents of any length. Furthermore we showed that there are no statistical differences in document length between relevant and non-relevant documents, therefore supporting our hypotheses.

This finding highlights the need for alternative ways to capture relevance in microblog documents. Firstly we redefine 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.

Moreover, we proposed the notion of "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 strong relationship with the structure of the document in terms of how much space is dedicated to each dimension.

Secondly, we propose a technique which re-weights the retrieval score of microblog documents based on how much the space dedicated to each dimension diverges from the optimal. By doing so, we were able to significantly improve the behaviour of a state of the art retrieval model in the context of 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 paper for other purposes, such as Automatic Query Expansion.

REFERENCES

Younos Aboulnaga, Charles L. A. Clarke, and David R. Cheriton. 2012. Frequent Itemset Mining for Query Expansion in Microblog Ad-hoc Search. (2012).

Gianni Amati, Giuseppe Amodeo, Marco Bianchi, Giuseppe Marcone, Fondazione Ugo Bordoni, Carlo Gaibisso, Giorgio Gambosi, Alessandro Celi, Cesidio Di Nicola, and Michele Flammini. 2011. FUB, IASI-CNR, UNIVAQ at TREC 2011 Microblog Track.. In *TREC*.

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

Amparo Elizabeth Cano Basave, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2013. Making Sense of Microposts (# MSM2013) Concept Extraction Challenge.. In # *MSM*. 1–15.

Firas Damak, Karen Pinel-Sauvagnat, Mohand Boughanem, and Guillaume Cabanac. 2013. Effectiveness of State-of-the-art Features for Microblog Search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13)*. ACM, New York, NY, USA, 914–919. DOI:http://dx.doi.org/10.1145/2480362.2480537

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

Jinhua Gao, Guoxin Cui, Shenghua Liu, Yue Liu, and Xueqi Cheng. 2013. ICTNET at Microblog Track in TREC 2013. (2013).

Zhongyuan Han, Xuwei Li, Muyun Yang, Haoliang Qi, Sheng Li, and Tiejun Zhao. 2012. HIT at TREC 2012 Microblog Track. *TREC Microblog 2012* (2012).

D. Hiemstra. 2001. Using Language Models for Information Retrieval. (2001). http://purl.org/utwente/36473

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

Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem. 2013. IRIT at TREC Microblog Track 2013. (2013).

Yubin Kim, Reyyan Yeniterzi, and Jamie Callan. 2012. Overcoming Vocabulary Limitations in Twitter Microblogs. *TREC Microblog 2012* (2012).

Y. Li, Z. Zhang, W. Lv, Q. Xie, Y. Lin, R. Xu, W. Xu, G. Chen, and J. Guo. 2011. PRIS at TREC2011 Micro-blog Track. (2011).

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

D. Metzler and C. Cai. 2011. Usc/isi at trec 2011: Microblog track. In *Proceedings of the Text REtrieval Conference (TREC 2011)*.

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

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

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

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

Jesus A Rodriguez Perez, Andrew J McMinn, and Joemon M Jose. 2013. University of Glasgow (UoG_-TwTeam) at TREC Microblog. (2013).

B Pre-Processing. 2013. BJUT at TREC 2013 Microblog Track. (2013).

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

Thomas Roelleke. 2013. Information Retrieval Models: Foundations and Relationships. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 5, 3 (2013), 1–163.

B. Sharifi, M.-A. Hutton, and J.K. Kalita. 2010. Experiments in Microblog Summarization. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. 49–56. DOI:http://dx.doi.org/10.1109/SocialCom.2010.17

Yajing Yuan Hui Wang Guang Chen Siming Zhu, Zhe Gao. 2013. PRIS at 2013 Microblog Track. (2013).

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

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

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

Sarvnaz Karimi Jie Yin Paul Thomas. 2012. Searching and Filtering Tweets: CSIRO at the TREC 2012 Microblog Track. (2012).

Chengxiang Zhai and John Lafferty. 2001. 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*. ACM, 334–342.

Received February 2007; revised March 2009; accepted June 2009