Hypergeometric Language Models for Republished Article Finding

Manos Tsagkias
ISLA, University of Amsterdam
e.tsagkias@uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

Wouter Weerkamp
ISLA, University of Amsterdam
w.weerkamp@uva.nl

ABSTRACT

Republished article finding is the task of identifying instances of articles that have been published in one source and republished more or less verbatim in another source, which is often a social media source. We address this task as an ad hoc retrieval problem, using the source article as a query. Our approach is based on language modeling. We revisit the assumptions underlying the unigram language model taking into account the fact that in our setup queries are as long as complete news articles. We argue that in this case, the underlying generative assumption of sampling words from a document with replacement, i.e., the multinomial modeling of documents, produces less accurate query likelihood estimates.

To make up for this discrepancy, we consider distributions that emerge from sampling without replacement: the central and non-central hypergeometric distributions. We present two retrieval models that build on top of these distributions: a log odds model and a bayesian model where document parameters are estimated using the Dirichlet compound multinomial distribution.

We analyse the behavior of our new models using a corpus of news articles and blog posts and find that for the task of republished article finding, where we deal with queries whose length approaches the length of the documents to be retrieved, models based on distributions associated with sampling without replacement outperform traditional models based on multinomial distributions.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Experiment, Theory

Keywords

Language models, Hypergeometric, Multinomial, Linking, Online news, Social Media

1. INTRODUCTION

Republished article finding (RAF) is the task of identifying instances of articles that have been published in one source and republished more or less verbatim in another. A common instance of the phenomenon occurs with news articles that are being republished by bloggers. The RAF task is important for a number of stakeholders. Publishers of news content are a prime example. For us, the motivation for considering the RAF task comes from the area of online reputation management.

Over the past decade, the web has come to play an increasingly important role in the overall communication strategy of organisations. It continues to offer new opportunities for organisations to directly interact with their customers or audience but it also contains possible threats as online conversations are impossible to control, while the potential impact on an organisation's reputation may be deep and long-lasting. Online reputation management (ORM) is aimed at monitoring the online reputation of an organization, brand or person, by mining news, social media and search engine result pages.

A key aspect of ORM is early detection of news topics that may end up harming the reputation of a given company, brand or person ("customer"), so that public relations activities can be launched to counter such trends. For this purpose it is important to track news stories that talk about an issue that affects the customer. In the blogosphere news stories may be republished for a number of reasons. In our data sets (see Section 4 for details), we have come across instances where bloggers want to share a news item with colleagues or students[1] or where a blogger aims to kick off a discussion around the original news article within his own online community,[2] or where someone uses excerpts from a news article as references in a post where they discuss their own opinion.[3] In addition to this "strict" interpretation of the RAF task (where most or all of a source article is being republished), ORM analysts are also interested in a somewhat looser interpretation, where a key part of a source article (e.g., its lead) is being republished in social media. Republished articles matter to ORM analysts as they may become springboards where intense, possibly negative discussions flare up.

SIGIR'11, July 24–28, 2011, Beijing, China.

[1] E.g., a very large part of NYT article "A Boy the Bullies Love to Beat Up, Repeatedly" `http://www.nytimes.com/2008/03/24/us/24land.html?_r=1` was republished verbatim in The Kentucky School blog written by the school's teachers, at `http://theprincipal.blogspot.com/2008/03/boy-bullies-love-to-beat-up-repeatedly.html`.

[2] E.g., all of "Financial Russian Roulette" by NYT journalist Paul Krugman was reposted by Mark Thoma (Professor of Economics at University of Oregon) at `http://economistsview.typepad.com/economistsview/2008/09/paul-krugman-fi.html`, with a one sentence commentary by Thoma, followed by about 110 follow-up comments.

[3] See, e.g., "What parts of the agenda would you sacrifice to try to put bushies in jail for torture?" `http://nomoremister.blogspot.com/2009/01/what-parts-of-agenda-would-you.html`.

Having motivated the task of finding republished news articles in the blogosphere, we now turn to addressing the task. At first glance the strict version of the task looks like a duplicate detection task. As we show in Section 5 below, on a strict interpretation of the RAF task, state-of-the-art duplicate detection methods show very reasonable performance in terms of MRR but in terms of MAP they leave room for improvement. Under the more liberal interpretation of the RAF task, the performance of state-of-the-art duplication detection methods drops rapidly, on all metrics.

These initial findings motivate the use of standard information retrieval methods for the RAF task, viewing the original news article as a query to be submitted against an index consisting of, say, blog posts [19, 36]. We follow the latter and focus on language modeling (LM) techniques for the RAF task. Language modeling in IR is usually based on distributions that emerge from *sampling with replacement*, e.g., 2-Poisson, bernoulli, binomial, multinomial [32]. This allows a generative model of language to serve its purpose, namely, to produce infinite amounts of word sequences from a finite word population. However, in the particular case of the RAF task, we are dealing with long (document-size) queries. Here, sampling with replacement can lead to overgeneration of unseen terms; when paired with the long query length, this can have a cumulative and negative effect on performance. It is well-known from general statistics that when the sample size grows close to the population size, i.e., when it is less than 10 times the population, models based on sampling with replacement become less and less accurate [30]. In our case, we consider documents and queries as bags of word level unigrams; unigrams from the document form the population, and unigrams from the query form the sample. In the standard ad hoc retrieval setting, queries tend to be much shorter than documents, i.e., the sample is much smaller than the population. For example, title queries in the TREC Robust 2004 test set have 3 words, while documents are on average 500 words long [37]. However, in the case of our RAF task, the assumption that documents (blog posts) are at least 10 times longer than queries (source news articles) is blatantly violated: in our data set, the former are 800 words long, the latter as many as 700 words: the two are of comparable length.

Our main contribution is an LM-based retrieval model for the RAF task that builds on statistical distributions that emerge from *sampling without replacement*. Documents and queries are considered as urns that contain terms where multiple examples of each term can coexist simultaneously. A document's relevance to an information need, translates into the probability of sampling the query (the source news article) from the document (blog posts). Then, documents are ranked by this probability [33]. A suitable statistical distribution for this model is the *hypergeometric distribution* which describes the number of successes in a sequence of n draws from a finite population without replacement, just as the binomial/multinomial distribution describes the number of successes for draws with replacement.

Our approach to the RAF task consists of deriving a document model and a retrieval model. The document model is based on one of the two multivariate hypergeometric probability distributions we present here: (a) the central hypergeometric distribution and (b) the Wallenius' hypergeometric (also called non-central) distribution. Both can take into account local term weights (such as raw term frequency (TF), while the model based on the Wallenius' distribution also allows one to also incorporate global term weights (such as inverse document frequency (IDF)).

The main research question that we seek to answer is whether distributions based on sampling without replacement provide for a more effective retrieval model for the RAF task than (the usual) distributions based on sampling with replacement. We provide a positive answer to this question and complement the answer with a thorough experimental analysis of our proposed models plus a comparison to existing retrieval models tuned for the RAF task.

The rest of the paper is organized as follows. We present two hypergeometric distributions in Section 2, two retrieval models based on those distributions in Section 3 We present our experimental setup in Section 4, report on results and analysis in Section 5, discuss alternatives in Section 6, present related work in Section 7. and conclude in Section 8.

2. HYPERGEOMETRIC DISTRIBUTIONS

We present two hypergeometric distributions which we will use later for sampling a query from a document: (a) the central hypergeometric, and (b) non-central hypergeometric (also known as Wallenius' hypergeometric distribution). The difference between the two is in how we perform sampling and whether bias in sampling is involved. Under the non-central distribution the probability of drawing a term depends on the outcome of the previous draw, while under the central hypergeometric distribution, terms can be sampled independently [3, 18, 28].

Let us first describe the specific form of language model that we consider in this paper, which builds on the *query unigram model* model proposed in [42]. This model postulates that the relevance of a document to a query can be measured by the probability that the *query is generated by the document*.

Consider a query $\mathbf{q}$ and a document collection C of N documents, $C := \{\mathbf{d}_l\}_{l=1,\ldots,N}$, with both queries and document being represented as vectors of indexed term counts:

$$\begin{aligned} \mathbf{q} &:= (q_1, \ldots, q_i, \ldots, q_V) \in N^V \\ \mathbf{d}_l &:= (d_{l,1}, \ldots, d_{l,i}, \ldots, d_{l,V}) \in N^V \end{aligned}$$

where q_i is the number of times the term i appears in the query and V is the size of the vocabulary. Let us also define the *length* of a query ($n_\mathbf{q}$) and of a document (n_l) as the sum of their components: $n_\mathbf{q} := \sum_i q_i$ and $n_l := \sum_i d_{l,i}$.

2.1 The central hypergeometric distribution

The *multivariate central hypergeometric distribution* is derived from the observation that since the sampling is done without replacement, the unordered sample is uniformly distributed over the combinations of size $n_\mathbf{q}$ chosen from $\mathbf{d}_l$:

$$P_{ch}(\mathbf{q}; n_\mathbf{q}, n_l, \mathbf{d}_l) = \frac{\prod_{i=1}^{V} \binom{d_{l,i}}{q_i}}{\binom{n_l}{n_\mathbf{q}}}, \tag{1}$$

Terms are sampled independently or simultaneously, reflecting the term independence assumption.

2.2 The non-central hypergeometric distribution

What if terms had an additional property that affected their probability of being sampled, for example, in how many documents they occur? In the urn model, we can think about objects that, except of their different color, can be heavier or bigger than others. This additional property can bias sampling and can be modeled as a weight for each object type. We call this weight ω_i for the ith term in the vocabulary.

Under the *multivariate non-central hypergeometric distribution* the probability of sampling a term depends on the terms sampled so far and also on the remaining terms in the urn. Further, it supports biased sampling allowing for the incorporation of global term

weights directly in the probability calculation. The following formula describes the distribution:

$$P_{wh}(\mathbf{q}; n_{\mathbf{q}}, n_l, \mathbf{d}_l) = \left[\prod_i^V \binom{d_{l,i}}{q_i}\right] \int_0^1 \prod_{i=1}^V \left(1 - t^{\omega_i/\Xi}\right)^{q_i} dt \quad (2)$$

where $\Xi = \boldsymbol{\omega} \cdot (\mathbf{n_l} - \mathbf{n_q}) = \sum_{i=1}^V \omega_i (d_{l,i} - q_i)$ regulates the bias of q_i after every draw, and the integral stands for the recursive sampling from time $t = 0$ until all terms are sampled at $t = 1$.

The mathematical derivation, properties and efficient computation methods of the Wallenius' distribution are beyond the scope of this paper. Wallenius [38] provides in-depth information on the characteristics of the non-central distribution and Fog [13] presents efficient methods for sampling from it. The central and non-central hypergeometric distributions are connected in that when $\omega_i = 1$ for all i, then bias is cancelled and the non-central hypergeometric distribution degenerates into the central hypergeometric distribution.

2.3 An example

Now that we have presented the hypergeometric distributions, let us look at an illustrative example on how sampling with and without replacement can lead to different results when the sample size is close to the population size. We start with a query (sample) and we need to calculate the probability of a document (population) to generate the query. In the case of *sampling with replacement*, the probability of sampling a query term t from a document D follows the binomial distribution:[4]

$$binomial(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \quad (3)$$

with parameters n, p where n is the number of trials (query size), $p = \frac{\#t_D}{|D|}$ is the probability of a success, namely, the term frequency of t in D, and k is the number of successes, i.e., the term frequency of t in Q. In the case of *sampling without replacement* the probability of sampling a query term t from a document D follows the hypergeometric distribution:

$$hypergeometric(k; m, n, N) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \quad (4)$$

with parameters m, n, N where m is the term frequency of t in D, n the number of draws (query size), N the population size (document size), and k the number of successes, namely the term frequency of t in Q.

For our example, we let query Q have 4 terms, each occurring once, and also we define two documents A and B of length 1,000, and 15, respectively which share at least one common term with Q. Let also that query term t occurs 1 time in A and B. The probability of sampling t from A or B when we sample *with* replacement is given by (3) with $k = 1$, $n = 4$, $p_A = 1/1,000$, and $p_B = 1/15$. The calculations result in values of 0.003988 for document A and 0.216810 for B. Similarly, when sampling *without* replacement, we use (4) and set $k = 1$, $m = 1$, $n = 4$, $N_A = 1,000$, and $N_B = 15$. This results in values of 0.004000 for A and in 0.266666 for B. These numbers show that the difference in probability from the two models for document A is negligible ($1.2 \cdot 10^{-5}$) but when the population is close to the sample size (document B), the difference grows three orders of magnitude reaching 0.049. The example illustrates that when queries are of comparable size to the retrieved documents, sampling with replacement can lead to poor likelihood estimates with a cumulative negative effect in the multivariate case, i.e., we calculate probabilities for all query terms.

What is the upshot? It is known that the multinomial approximates the central hypergeometric as the population size remains many times larger than the sample size, i.e., when the document is much longer than the query. In the RAF task, this assumption is violated as queries and documents are expected of roughly the same size. This motivates us to derive retrieval models based on hypergeometric modeling of documents instead of multinomial models.

[4]We use the binomial distribution instead of the multinomial for simplifying the calculations in our example.

3. RETRIEVAL MODELS

Before deriving retrieval models based on hypergeometric modeling of documents, we revisit (1) and (2). We identify three constraints emerging from these equations, which relate to *smoothing* and play a role in the design of a retrieval model:

1. only query terms that occur in the document contribute to the probability,
2. the query should be shorter than the document,
3. the frequency of a query term should be lower than or equal to the term's frequency in the document.

The first constraint is obvious. The other two stem from the fact that is impossible to draw more terms than currently exist in the urn. The second constraint is imposed from the denominator $\binom{n_l}{n_{\mathbf{q}}}$ which becomes zero when $n_{\mathbf{q}} > n_l$ and results in infinite probability. The third constraint roots in $\binom{d_{l,i}}{q_i}$ which becomes zero if $q_i > d_{l,i}$ and results in zero probability. In general, $P(\mathbf{q})$ is positive only if

$$\max(0, n_{\mathbf{q}} + d_{l,i} - n_l) \leq q_i \leq \min(d_{l,i}, n_{\mathbf{q}}).$$

To address the three constraints listed above, we consider two types of smoothing. The performance of retrieval models that build on top of hypergeometric distributions is sensitive to the employed smoothing strategy, just like other retrieval models are that build on the multinomial or other distributions. In the following two subsections we present two approaches to smoothing. The first approach is somewhat related to relevance feedback and an estimated document model is trained on text from both the query and the document; this approach works for both the central and non-central hypergeometric distribution. The second approach is more elaborate and is based on bayesian inference; this approach works only for the central hypergeometric distribution, as we explain below.

3.1 A log-odds retrieval model

Our first approach to overcome the limitations on q_i, $n_{\mathbf{q}}$ given a document, is basic in terms that no sophisticated smoothing methods are involved for estimating the parameters of the document model. In a sense, it is remotely related to pseudo-relevance feedback but instead of re-estimating the query model from pseudo-relevant documents, the documents models are complemented with information from the query. One way to visualize the process is to think of a bag with query terms from which we sample the query. Obviously, the probability of sampling the query from the bag is 1. Now, for a document in the collection we add the document terms in the bag and sample the query again. Documents with high vocabulary overlap with the query will result in high probability while documents with only few common terms will result in low probability.

In particular, instead of sampling the query directly from the document, we derive a hypothetical document $\mathbf{d}'$ which is a mixture of the query $\mathbf{q}$ and the document $\mathbf{d}$:

$$\mathbf{d}' := (d'_{l,1}, \ldots, d'_{l,i}, \ldots, d'_{l,V}) \in N^V,\ d'_{l,i} = r_q q_i + r_d d_{l,i}, \quad (5)$$

where r_q, r_d are parameters for regulating the mixture. The length of this hypothetical document is: $n'_l = \sum_i r_q q_i + r_d d_{l,i} = r_q n_{\mathbf{q}} + r_d n_l$.

Now, it holds that $P_\cdot(\mathbf{q}; n_{\mathbf{q}}, n'_l, \mathbf{d}'_l) \in (0, 1]$ because at least some of the terms are always sampled from $\mathbf{d}'_l$ (i.e., those originating from the query), but never all of them because $n_{\mathbf{q}} < n'_l$ by definition of $\mathbf{d}'_l$. The extreme case of $P_\cdot(\mathbf{q}; n_{\mathbf{q}}, n'_l, \mathbf{d}'_l) = 1$ is reached when there is no vocabulary overlap between $\mathbf{d}_l$ and $\mathbf{q}$ and $r_q = 1$, however, this case is hardly encountered in practice because documents without common terms are excluded from ranking.

Document length and the vocabulary intersection between the query and the document both play an important role in the probability outcome, as in other retrieval models. To this end, we normalize the probability given the observation that the probability should maximize when the document is an exact duplicate of the query, i.e., $\mathbf{q} = \mathbf{d}_l$:

$$P_\cdot^{max} = P_\cdot(\mathbf{q}; n_{\mathbf{q}}, (r_q + r_d) n_{\mathbf{q}}, (r_q + r_d)\mathbf{q}). \tag{6}$$

Given this observation, documents able to generate the query with probability close to the maximum should be favored. We express this in the following ranking function:

$$\begin{aligned} Score(Q, D) &= \frac{P_\cdot(\mathbf{q}; n_{\mathbf{q}}, n'_l, \mathbf{d}'_l)}{P_\cdot^{max}} \\ &\propto P_\cdot(\mathbf{q}; n_{\mathbf{q}}, n'_l, \mathbf{d}'_l)). \end{aligned} \tag{7}$$

The denominator can be ignored for ranking since it is constant for all documents. The expression of $P_\cdot^{max}$ holds when we look at finding near or exact duplicates of a query. Under this scenario, query terms are expected to occur in a candidate "duplicate" document in relatively similar frequencies. However, this is hardly true in other settings where retrieved documents can deviate considerably from the query in both vocabulary and term frequencies.

In this respect, the assumption we made for deriving (6), can be too strict. The assumption can be relaxed if we only take into account terms common to the query and to the document and compute the maximum probability based on those. Similarly as before, first we derive a hypothetical document:

$$\mathbf{d}''_l := \{d''_{l,i} : d''_{l,i} = (r_q + r_d) q_i \text{ for } i \in V, q_i > 0, d_{l,i} > 0, \},$$

with length $n''_l = \sum_i d''_{l,i}$. Further, we also reduce the original query to a hypothetical query $\mathbf{q}''$ that consists of terms common to $\mathbf{q}$ and $\mathbf{d}_l$:

$$\mathbf{q}'' := \{q''_i : q_i \text{ for } i \in V, q_i > 0, d_{l,i} > 0, \}.$$

This results in the following definition of maximum probability, previously defined in (6):

$$P_\cdot'^{max} = P_\cdot(\mathbf{q}; n_{\mathbf{q}}, n''_l, \mathbf{d}''_l), \tag{8}$$

and the ranking function in (7) becomes:

$$Score(Q, D) = \frac{P_\cdot(\mathbf{q}; n_{\mathbf{q}}, n'_l, \mathbf{d}'_l)}{P_\cdot'^{max}}. \tag{9}$$

In this representation, $P_\cdot'^{max}$ cannot be ignored because it is dependent on the vocabulary overlap of the query and the document.

3.2 A bayesian retrieval model

A second approach to overcome the limitations on q_i, $n_{\mathbf{q}}$ noted at the start of this section, is to use bayesian inference. Recall that when documents are modeled as a multinomial distribution of terms, and we apply bayes' rule, the conjugate prior distribution to the multinomial is the Dirichlet distribution [41, 42]. Setting the parameters of the Dirichlet distribution accordingly, leads to the well known Dirichlet smoothing method. Here, we follow the same line of reasoning for the multivariate central hypergeometric, and arrive at the *Dirichlet compound multinomial distribution* (DCM, also known as the multivariate Polya distribution) for estimating the parameters of a document model. To the best of our knowledge no closed form is known for the conjugate priors of the non-central hypergeometric distribution; hence, we do not offer a bayesian non-central hypergeometric model.

Now, let us consider that terms $\mathbf{t} = (t_1, \ldots, t_i, \ldots, t_V)$ arise from a multivariate central hypergeometric process where parameter n_N, the vocabulary length, is known ($n_N = \sum_{l=1}^{N} \sum_{i=1}^{V} d_{l,i}$ and $n_N > 0$) and $\theta_l = (\theta_{l,1}, \ldots, \theta_{l,V})$, the vector of term frequencies in the vocabulary that make up the population, are unknown ($0 \leq \theta_{l,i} \leq n_l$ and $\sum_i \theta_i = n_l$).

Under this model, the probability of generating a particular query q with counts $\mathbf{q}$ is given by:

$$P_{ch}(\mathbf{q}|\theta_l) = \frac{\prod_{i=1}^{V} \binom{\theta_{l,i}}{q_i}}{\binom{n_l}{n_{\mathbf{q}}}}, \tag{10}$$

In the case where documents consist of all vocabulary terms, we can obtain the point estimate $\theta_{l,i} = d_{l,i}$. However, such documents rarely exist. Rather than find a point estimate for the parameter vector θ_l, a distribution over θ_l is obtained by combining a prior distribution over the model parameters $P(\theta_l)$ with the observation likelihood $P(\mathbf{d}_l|\theta_l)$ using Bayes' rule:

$$P(\theta_l|\mathbf{d}_l) = \frac{P(\theta_l) P(\mathbf{d}_l|\theta_l)}{P(\mathbf{d}_l)}, \tag{11}$$

where the observation likelihood is given by:

$$P_{ch}(\mathbf{d}_l|\theta_l) = \frac{\prod_{i=1}^{V} \binom{\theta_{l,i}}{d_{l,i}}}{\binom{n_N}{n_l}}. \tag{12}$$

The conjugate prior of a multivariate hypergeometric process is the DCM with hyperparameters H, an integer greater than zero, and $\alpha = (\alpha_1, \ldots, \alpha_i, \ldots, \alpha_V)$ where $\alpha_i > 0$ and $\sum_{i=1}^{V} \alpha_i = 1$:

$$P(\theta) = \frac{n_l!}{\prod_{i=1}^{V} \theta_i!} \frac{\Gamma(\sum_{i=1}^{V} \alpha_i)}{\prod_{i=1}^{V} \Gamma(\alpha_i)} \frac{\prod_{i=1}^{V} \Gamma(\alpha_i + \theta_i)}{\Gamma(\sum_{i=1}^{V} (\alpha_i + \theta_i))}. \tag{13}$$

where $\theta_i > 0$ and $\sum_i^V \theta i = n_l$. The resulting posterior distribution is also DCM:

$$\begin{aligned} P(\theta|\mathbf{d}_l) &= \frac{(n_V - n_l)!}{\prod_{i=1}^{V} (\theta_i - d_{l,i})!} \\ &\cdot \frac{\Gamma(n_l + \sum_{i=1}^{V} \alpha_i)}{\prod_{i=1}^{V} \Gamma(d_{l,i} + \alpha_i)} \frac{\prod_{i=1}^{V} \Gamma(\alpha_i + \theta_i)}{\Gamma(\sum_{i=1}^{V} (\alpha_i + n_V))}, \end{aligned} \tag{14}$$

with $\theta_i > 0$ and $\sum_{i=1}^{V} \theta_i = H$.

The query likelihood then becomes:

$$P(\mathbf{q}|\mathbf{d}_l) = \int_\theta P(\mathbf{q}|\theta_l) P(\theta_l|\mathbf{d}_l) d\theta_l \tag{15}$$

A standard approximation to the Bayesian predictive distribution $P(\mathbf{q}|\mathbf{d}_l)$ is the use of the maximum posterior (MP) distribution. The approximation consists of replacing the integral in (15) with its maximum value [41, 42]:

$$P_{ch}(\mathbf{q}|\theta_l^{MP}) = \frac{\prod_{i=1}^{V} \binom{\theta_{l,i}^{MP}}{q_i}}{\binom{n_l^{MP}}{n_{\mathbf{q}}}}, \tag{16}$$

Although, there is no closed form solution for the maximum likelihood estimate θ_i of DCM [40], we can use the expected value of

θ_i [20, p.80]:

$$\theta_{l,i}^{MP} = (n_N - n_l)\frac{\alpha_i + d_{l,i}}{\sum_{i=1}^{V}(\alpha_i + d_{l,i})} = (n_N - n_l)\frac{\alpha_i + d_{l,i}}{n_l + \sum_{i=1}^{V}(\alpha_i)},$$

Following [42], we assign $\alpha_i = \mu P(i|C)$ where μ is a parameter and $P(i|C)$ is the probability of the ith term in the collection and the equation above becomes:

$$\theta_{l,i}^{MP} = (n_N - n_l)\frac{d_{l,i} + \mu P(i|C)}{n_l + \mu}. \quad (17)$$

The derivation of DCM from the central hypergeometric distribution is important, because it establishes a similar link to that between multinomial and Dirichlet smoothing. In this respect, the use of DCM is expected to result in positive performance differences over Dirichlet when the sample size is close to the population size but these differences will become smaller when the sample is a small fraction of the population. Indeed, Elkan [12] compared the performance of DCM and the multinomial for document clustering (sample and population are expected to be of comparable size) with results favoring DCM. Xu and Akella [40] introduced a probabilistic retrieval model with experiments on ad hoc retrieval (when the sample is just a fraction of the population size) using Dirichlet and DCM smoothing with results that favour DCM, but small, although statistically significant, differences.

4. EXPERIMENTAL SETUP

We present our research questions, experiments, dataset and evaluation method. For the purpose of finding instances of articles that have been published in one source and republished more or less verbatim in another, we choose to focus on a single target source in our experimental evaluation, namely the blogosphere. This choice is based on the fact that blog posts unlike status updates or microblog posts, can be of arbitrary length and therefore they can be verbatim copies of a news article.

4.1 Experiments

In addressing the RAF problem in both its *strict* and *loose* interpretation, we concentrate on the retrieval effectiveness of the hypergeometric retrieval models for finding how news content propagates in the blogosphere. In this respect our goals are comparable to those of [19, 22, 23, 34]. In particular, we want to know the effectiveness of our log odds hypergeometric retrieval models and of bayesian hypergeometric retrieval model, both for finding republished articles.

To answer these research questions, we compare our methods to seven state-of-the-art retrieval methods listed in Table 1. Among them, simhash is one of the best-performing near-duplicate detection methods [17, 26]; kl has proven successful in plagiarism detection [5]; cosine, probabilistic, and language modeling based methods have performed well in the related topic detection and tracking [2] task.

In our experiments we use the Indri framework for indexing. Each experimental condition returns maximum 1,000 results. For parametric retrieval models we find parameter values that optimize their performance for our dataset. We set $\mu = 1120$ for kl, lm, indri, hgm-central-bayes, $r_q = 1, r_d = 1$ for hgm-central, and hgm-noncentral, and $k_1 = 2.0, b = 0.75$ for bm25f. For hgm-noncentral we set ω_i, to the term's inverse document frequency (IDF).

Table 1: Retrieval models we consider.

Model	Gloss
State-of-the-art models	
simhash	Hamming distance between two simhashes
cosine	Cosine similarity using IDF term weighting
kl	Kullback-Leibler divergence
lm	Unigram language model with Dirichlet smoothing
indri	Language modeling with inference networks and Dirichlet smoothing
bm25f	Okapi BM25F
tf·idf	TFIDF retrieval model
Hypergeometric models	
hgm-central	Log odds retrieval model with multivariate central hypergeometric distribution (9)
hgm-central-bayes	Multivariate central hypergeometric distribution with Dirichlet compound Multinomial smoothing (16)
hgm-noncentral	Log odds retrieval model with multivariate noncentral hypergeometric distribution (16)

4.2 Dataset

The data set that we use as our target social media collection is the Blogs08 collection provided by TREC; the collection consists of a crawl of feeds, permalinks, and homepages of 1.3M blogs during early 2008–early 2009. This crawl results in a total of 28.4M blogs posts (or permalinks). We only used feed data, the textual content of blog posts distributed by feeds and ignored the permalinks. Only using feed data is common practice and requires almost no preprocessing of the data. Extracting posts from the feed data gave us a coverage of 97.7% (27.8M posts extracted). As a second preprocessing step we perform language identification and remove all non-English blog posts from the corpus, leaving us with 16.9M blogs posts. Our index is constructed based on the full content of blog posts.

Our news article dataset is based on the headline collection from the top stories task in TREC 2009. This is a collection of 102,812 news headlines from the New York Times that includes the article title, byline, publication date, and URL. For the purposes of our experiments we extended the dataset by crawling the full body of each of the articles.

4.3 Ground truth and metrics

As there is no standard test collection for the republished article finding task, we created our own.[5] The ideal ground truth for our task would consist of tuples (n, s) consisting of a news article and a social media utterance, where s is a republication of n.

As a proxy, we follow [15, 27, 29] and use blog posts that are explicitly linked to a given news source. Our ground truth is assembled in two phases. First, for each news article we find blog posts that include the article's URL. Second, for each discovered blog post we look for other blog posts that include its URL. The process continues recursively until no more blog posts are discovered. For our experiments we sample headlines with more than ten explicit links and where social media possibly plays a role. For each news article, we take only explicitly linked blog posts within ± 1 day from the article's publication date to reduce the search space.

In the second phase, we removed the explicit links and for each (backlinked) blog post we manually examined whether it is a republication of the news article. In the strict interpretation of the RAF task, the blog post needs to be a copy all of the material from the source news article, possibly interleaved with comments etc.

[5]The ground truth may be retrieved from `http://ilps.science.uva.nl/resource/hypergeometric-lm`

Table 2: Relevance assessments for *strict* and *loose* interpretations of the RAF task.

GT	Topics #	Topics Max length	Topics Min length	Topics Avg. length	Relevant documents #	Relevant documents Max length	Relevant documents Min length	Relevant documents Avg. length	Per topic average
Loose	404	1,723	28	912	5,269	5,362	3	339	13
Strict	160	1,667	324	883	257	2,205	258	774	2

Table 3: System performance for the *strict* interpretation of the RAF on 160 news articles using three hypergeometric models, and seven other retrieval methods. Significance tested against simhash.

runID	P@5	MRR	Rprec	MAP
Baseline				
simhash	0.2838	0.8139	0.6806	0.7794
Hypergeometric retrieval models				
hgm-central	$0.3088^{\blacktriangle}$	$0.8948^{\blacktriangle}$	$0.8160^{\blacktriangle}$	$0.8874^{\blacktriangle}$
hgm-central-bayes	$0.3100^{\blacktriangle}$	0.8521	$0.7390^{\triangle}$	$0.8429^{\blacktriangle}$
hgm-noncentral	$0.3088^{\blacktriangle}$	$0.8969^{\blacktriangle}$	$0.8098^{\blacktriangle}$	$0.8858^{\blacktriangle}$
Other retrieval models				
cosine	$0.3088^{\blacktriangle}$	$0.8833^{\blacktriangle}$	$0.7702^{\blacktriangle}$	$0.8691^{\blacktriangle}$
bm25f	$0.3075^{\blacktriangle}$	$0.8896^{\blacktriangle}$	$0.7692^{\blacktriangle}$	$0.8713^{\blacktriangle}$
kl	$0.3100^{\blacktriangle}$	0.8542	$0.7442^{\triangle}$	$0.8457^{\blacktriangle}$
lm	$0.3100^{\blacktriangle}$	0.8500	0.7358	$0.8406^{\blacktriangle}$
indri	$0.3100^{\blacktriangle}$	0.8479	0.7358	$0.8409^{\blacktriangle}$
tf·idf	$0.1762^{\blacktriangledown}$	$0.4524^{\blacktriangledown}$	$0.2775^{\blacktriangledown}$	$0.4389^{\blacktriangledown}$

In the loose interpretation our assessors made sure that a key part of the source news article was republished in the blog post (e.g., a highly informative title, the news articles's lead or a central paragraph). Two assessors created this ground truth and discussed any differences they encountered until agreement was reached. See Table 2 for details of the resulting test collection; recall that in this paper, news articles are the queries that are submitted against an index of blog posts.

We report on standard IR measures: precision at 5 (P@5), mean reciprocal rank (MRR), mean average precision (MAP), and r-precision (Rprec). Statistical significance is tested using a two-tailed paired t-test and is marked as $^{\blacktriangle}$ (or $^{\blacktriangledown}$) for significant differences for $\alpha = .01$, or $^{\triangle}$ (and $^{\triangledown}$) for $\alpha = .05$.

5. RESULTS AND ANALYSIS

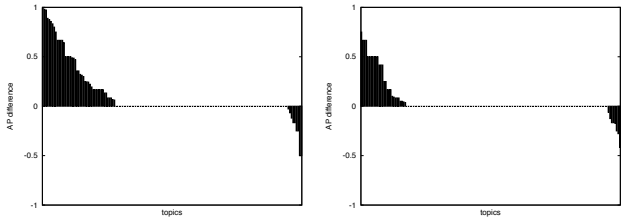
In this section, we report on the results of our experiments and conduct an analysis of their outcomes.

Strict interpretation. In our first experiment we study the retrieval effectiveness of our methods with regards to the strict interpretation of the RAF task. To this end, we choose simhash, the state-of-the-art for near-duplicate detection, as our baseline. The performance of three hypergeometric models, and seven retrieval models is listed in Table 3. We see that hgm-central and hgm-noncentral outperform the baseline with statistically significant differences in all metrics. Second and third best (in terms of MAP) come bm25f and cosine similarity with small differences between them; kl, hgm-central-bayes, lm, and indri follow with performance that hovers at the same levels. In general, all methods show strong performance in all metrics, with an exception for tf·idf.

Turning to individual metrics, we find of particular interest Rprec and MAP. For hgm-central Rprec peaks at 0.8160, 20% more than for simhash. In terms of MAP, hgm-central achieves the best score at 0.8874, a 14% improvement over the baseline. With regards to other language modeling based methods, hgm-central outperforms kl, lm, indri (statistically significantly so, in MRR, Rprec, and MAP). In terms of early precision (P@5), all methods show similar performance, which is mainly due to the small number of relevant documents per news article.

To better understand the differences between hgm-central and simhash, we look at per topic differences in average precision. Fig. 1 shows that out of 160 articles, 45 favor the use of hgm-central, and 9 simhash. Manual inspection of the results revealed that hgm-central is able to account for small changes in language: For example, if the title of the republished article had been changed in the blog post, then, according to hgm-central, this blog post will rank lower than a blog post where the title was kept the same as the original. simhash seems unable to capture these differences. This is partially due to its nature which although allows document compression which improves efficiency, it looses in precision. Another finding was the robust ranking capabilities of hgm-central even in lower ranks: blog posts there used only a couple of sentences from the original article. In contrast, ranked lists from simhash were polluted quite early (rank 10), with long documents that are irrelevant to the article, but that do share language with the article; this is in line with findings in [34].

Turning to hgm-central and lm, we find no striking differences in the resulted ranked lists. Differences in MAP are mainly due to how the ground truth is constructed. More specifically, there exist topics for which either method is penalized because the first ranking document is not assessed, however, found relevant after manual inspection. In general, lm was found to rank blog posts higher that contain either short excerpts of the article without commentary, or blog posts that are verbatim copies of the article with lots of commentary. This behavior can be explained by the accumulation of term probabilities using Dirichlet smoothing: probability mass is assigned to terms occurring in the original article. We see that hgm-central counters this problem with the use of P'^{max} which ensures that documents are ranked by how much the blog post "deviates" from the original article.



(a) hgm-central vs. simhash (b) hgm-central vs. lm

Figure 1: Per topic difference in average precision (AP) for the strict RAF task.

Loose interpretation. In our second experiment we test retrieval methods with regards to the *loose* interpretation of the RAF task. We set our baseline to hgm-central as it proved the best performing method in the previous experiment. Results in Table 4 show that when we move away from near-duplicates, retrieval effectiveness drops for all methods. hgm-central achieves the best scores overall, followed by bm25f in MRR, and lm, indri, kl in MAP. In this interpretation of the RAF task, simhash, our previous baseline, is one of the least effective along with tf·idf.

Looking at the results in more detail, hgm-central shows ro-

Table 4: System performance for the *loose* interpretation of the RAF task of 404 news articles using three hypergeometric models, and seven other retrieval methods. Significance tested against hgm-central.

runID	P@5	MRR	Rprec	MAP
Hypergeometric retrieval models				
hgm-central	0.5446	0.7612	0.4642	0.4413
hgm-central-bayes	0.5411	0.7197$^{\blacktriangledown}$	0.4708	0.4322$^{\triangledown}$
hgm-noncentral	0.5550	0.7627	0.4702	0.4093$^{\blacktriangledown}$
Other retrieval models				
cosine	0.5198$^{\blacktriangledown}$	0.7379$^{\triangledown}$	0.4292$^{\blacktriangledown}$	0.4138$^{\blacktriangledown}$
bm25f	0.5505	0.7561	0.4662	0.4253$^{\blacktriangledown}$
kl	0.5426	0.7252$^{\blacktriangledown}$	0.4603	0.4351
lm	0.5351	0.7165$^{\blacktriangledown}$	0.4587	0.4366
indri	0.5361	0.7145$^{\blacktriangledown}$	0.4593	0.4360
simhash	0.2683$^{\blacktriangledown}$	0.5423$^{\blacktriangledown}$	0.1692$^{\blacktriangledown}$	0.1337$^{\blacktriangledown}$
tf·idf	0.1485$^{\blacktriangledown}$	0.3084$^{\blacktriangledown}$	0.1242$^{\blacktriangledown}$	0.1044$^{\blacktriangledown}$

bust performance in MRR which is statistically significant over the rest of retrieval methods. hgm-noncentral shows marginally better results in terms of P@5, MRR, and Rprec over hgm-central at the cost of MAP. Finally, we find interesting that bm25f used to outperform language modeling based methods in our first experiment, however, in the current scenario we observe the opposite. This change can be ascribed to the parameter estimation of the models, which is related to the nature of the relevant documents. hgm-central, and hgm-noncentral as parameter free models are not as sensitive to changes in the notion of "relevance."

Document length. Finally, we examine our hypothesis on the effect of document length (population size) and query length (sample size) in retrieval effectiveness between modeling documents as hypergeometric and multinomial distributions of terms. Fig. 2 illustrates the correlation of MAP, MRR, and the length of relevant documents over query length, for hgm-central and lm. Hypergeometric document modeling shows to have strong positive effects in both metrics when document length is up to 0.1 times the query length. As the query and the document length become equal, the differences between the hypergeometric and the multinomial diminish.

Our experimental results demonstrate the utility of hypergeometric retrieval models for the republished article finding task in both its *strict* and *loose* interpretation.

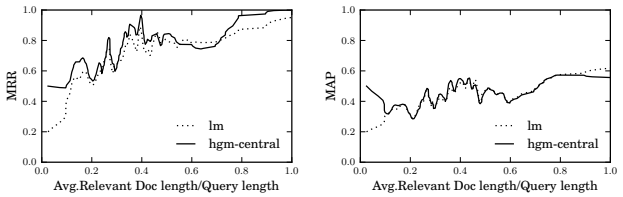


Figure 2: Moving average (window of 30) of MRR (left), and of MAP (right) over the ratio of average relevant document length and query length.

6. DISCUSSION

So far we have examined how different retrieval models perform on the two interpretations of the RAF task. In this section, we take a closer look at the distributions used for document modeling, namely, the multinomial and the hypergeometric and conduct a direct comparison of them by keeping the retrieval model the same and changing the underlying distribution. Further, we study the log odds retrieval model by experimenting with document/query representations, such as TF and TF·IDF, and with different mixture ratios r_q, r_d (see Section 3). Finally, we explore the use of hgm-central, hgm-noncentral and hgm-central-bayes in ad hoc retrieval.

Table 5: System performance on the *loose* interpretation of the RAF task using the log odds retrieval model and changing the underlying distribution to: multinomial, multivariate central hypergeometric, and multivariate non-central hypergeometric distribution. The parameters r_q, r_d are set to 1. Significance tested against the multinomial.

runID	P@5	MRR	Rprec	MAP
Log odds retrieval model				
multinomial	0.4297	0.6778	0.3177	0.2723
hgm-central	0.5446$^{\blacktriangle}$	0.7612$^{\blacktriangle}$	0.4642$^{\blacktriangle}$	0.4413$^{\blacktriangle}$
hgm-noncentral	0.5550$^{\blacktriangle}$	0.7627$^{\blacktriangle}$	0.4702$^{\blacktriangle}$	0.4093$^{\blacktriangle}$
Bayesian retrieval model				
lm	0.5351	0.7165	0.4587	0.4366
hgm-central-bayes	0.5411	0.7197	0.4708$^{\triangle}$	0.4322

6.1 Hypergeometric vs. multinomial

We are interested in exploring the validity of our hypothesis that hypergeometric document models are superior to multinomial ones when the query size is comparable to document length. We proceed as follows. For each of the two retrieval models we presented, i.e., log odds and bayesian, we create two runs, one using the multivariate hypergeometric distribution and one using the multinomial distribution. Keeping the same retrieval model and smoothing method and varying the underlying distribution, ensures that any observed differences in performance are solely due to the change in the underlying distribution. For our experiments we use the dataset from *loose* interpretation of the RAF task.

Log odds. We use the log odds retrieval model with the parameters r_q, r_d set to 1, and different underlying distributions: multinomial (multinomial), multivariate central hypergeometric (hgm-central), and multivariate non-central hypergeometric (hgm-noncentral). Results in the top half of Table 5 validate our hypothesis. Log-odds document models built on hypergeometric distributions outperform models built on the multinomial distribution. In particular, both hgm-central, and hgm-noncentral outperform multinomial in all metrics with statistically significant differences.

Dirichlet vs DCM. We compare the performance of Dirichlet smoothing on the multinomial distribution (unigram language model) and of DCM on the multivariate central hypergeometric. The smoothing parameter μ was found to peak at 1120 for both models when optimized for MAP. Table 5 (bottom) lists the results. Performance hovers at the same levels for both models, with DCM showing better r-precision with statistically significant difference. This can be attributed to the ability of DCM to capture word burstiness better than the Dirichlet [40] which leads to high early precision.

6.2 Mixture ratios and term weighting

We look at different mixture ratios r_q, r_d for hgm-central and hgm-noncentral; see (5). Table 6 shows that, on average, performance degrades as we deviate from $r_q = 1, r_d = 1$. When $r_d = 2$,

Table 6: System performance using log odds retrieval model and $tf \cdot idf$ for document and query representation, and several mixture ratios r_q, r_d. Significance testing against hgm-central with TF, and r_q, r_d set to 1.

runID	Weight	r_q	r_d	P@5	MRR	Rprec	MAP
Mixture ratios r_q, r_d							
hgm-central	TF	1	1	0.5446	0.7612	0.4642	0.4413
hgm-central	TF	1	2	0.5525	0.7576	0.4721$^{\blacktriangle}$	0.4382$^{\triangledown}$
hgm-central	TF	2	1	0.5198$^{\blacktriangledown}$	0.7251$^{\blacktriangledown}$	0.4189$^{\blacktriangledown}$	0.3611$^{\blacktriangledown}$
hgm-central	TF	3	5	0.5356	0.7338$^{\blacktriangledown}$	0.4436$^{\blacktriangledown}$	0.3908$^{\blacktriangledown}$
hgm-noncentral	TF	1	2	0.5515	0.7536	0.4670	0.4238$^{\blacktriangledown}$
hgm-noncentral	TF	2	1	0.5173$^{\blacktriangledown}$	0.7261$^{\blacktriangledown}$	0.4172$^{\blacktriangledown}$	0.3620$^{\blacktriangledown}$
hgm-noncentral	TF	3	5	0.5351	0.7307$^{\blacktriangledown}$	0.4428$^{\blacktriangledown}$	0.3886$^{\blacktriangledown}$
$tf \cdot idf$ representation							
hgm-central	TF·IDF	1	1	0.4238$^{\blacktriangledown}$	0.7097$^{\blacktriangledown}$	0.2912$^{\blacktriangledown}$	0.2435$^{\blacktriangledown}$
hgm-noncentral	TF·IDF	1	1	0.4861$^{\blacktriangledown}$	0.7297$^{\blacktriangledown}$	0.3581$^{\blacktriangledown}$	0.2901$^{\blacktriangledown}$

we observe a slight increase for some metrics at the cost of a lower MAP. In particular, hgm-central shows a statistically significant increase in Rprec.

Next, we explore the effect on performance of using global term weights, such as TF·IDF, instead of TF, for the representation of the hypothetical document $\mathbf{d}'$; see (5). The results in Table 6 (bottom) show that the use of TF·IDF leads to a significant decrease in performance for all metrics. Manual inspection reveals that the returned documents are very short, nearly one sentence long. The document size remains small, and comparable to two or three sentences until the end of the rank list. For the topics we examined at, the top ranked document is usually relevant, however, in most cases it is not assessed.

6.3 Ad hoc retrieval

Finally, we look at the performance of our log odds and bayesian retrieval models in ad hoc retrieval. For our experiments, we use TREC-Robust 2004. We formulate our queries using content from the title of each topic. The Dirichlet smoothing parameter μ is set to 1000. Table 7 shows results for indri (baseline), hgm-central-bayes, hgm-central, and hgm-noncentral. We see that hgm-central-bayes shows the same performance as the baseline. Runs based on the log odds retrieval model prove least effective. The reason lies in the value of $P_{\cdot}^{\prime max}$, which becomes 1 when the query and the document share only one common term–which is common for short queries. Without the normalization factor, and enough information from the query, the performance of the log odds model depends on $\mathbf{d}'$ which is mainly estimated from the document (given the negligible effect from the query due to its short length). To this end, the more elaborate smoothing methods, used in indri, and hgm-central-bayes prove most effective.

The three analyses that we performed in this section establish the following. The hypergeometric distributions are a better choice over the multinomial for modeling documents, when the system has to respond to document long queries. The Dirichlet and DCM smoothing show similar performance, with the later producing better early ranking. Further, retrieval effectiveness benefits the most from document representations that use raw term frequencies (TF), and equal mixture ratios r_q, r_d. Finally, with regards to ad hoc retrieval, retrieval models based on bayesian inference deliver the best performance.

Table 7: System performance on the TREC-ROBUST 2004 collection. Significance tested against indri.

runID	P@5	MRR	Rprec	MAP
Title				
indri	0.4570	0.6603	0.2638	0.2221
hgm-central	0.3590$^{\blacktriangledown}$	0.5406$^{\blacktriangledown}$	0.2096$^{\blacktriangledown}$	0.1650$^{\blacktriangledown}$
hgm-central-bayes	0.4578	0.6603	0.2638	0.2221
hgm-noncentral	0.3597$^{\blacktriangledown}$	0.5310$^{\blacktriangledown}$	0.2033$^{\blacktriangledown}$	0.1571$^{\blacktriangledown}$

7. RELATED WORK

Our task, republished article finding, is parallel to the tasks of text reuse which, in turn, relates to near-duplicate detection.

Near-duplicate detection. Garcia-Molina et al. [14] introduces the problem of finding document copies across multiple databases. Manku et al. [26] adopt simhash, a document fingerprinting method and hamming distance for efficient near-duplicate detection in web crawling; we used simhash as a baseline in our comparisons. Chang et al. [9] focus on finding event-relevant content using a sliding window over lengths of sentences. Muthmann et al. [31] discover near-duplicates within web forums for grouping similar discussion threads together. They construct a document's fingerprint from a four dimensional vector which consists of domain (in-)dependent text-based features, external links, and semantic features. Kolak and Schilit [23] find popular quoted passages in multiple sources, and use them to link these sources. Abdel-Hamid et al. [1] detect the origin of text segments using shingle selection algorithms. Zhang et al. [43] use two stage approach for finding partial duplicates with applications to opinion mining and enhanced web browsing: sentence level near-duplicate detection (Jaccard distance) and sequence matching; the tasks considered in this paper are similar to ours, however, the authors focus on pruning techniques, while we aim at discovering effective and robust methods, the output of which needs little, if any, further processing.

Text re-use. Broder [8] introduces the mathematical notions of "resemblance" and "containment" to capture the informal notions of "roughly the same" and "roughly contained" and propose efficient methods using document fingerprinting techniques. These notions correspond to our "strict" and "loose" interpretations of the republished article finding task. Seo and Croft [34] compare a set of fingerprinting techniques for text reuse on newswire and blog collections. One of their findings, which we also share, is how text in blogs layout affects the performance of fingerprinting methods. Kim et al. [22] propose an efficient overlap and content reuse detection in blogs and news articles. They find that blog posts contain large amount of exact quotations from the news articles. However, for the particular task, they find that blog posts raise significant challenges against retrieval [21]. Bendersky and Croft [6] consider the issue of text reuse on the web. They address the task using three methods: word overlap, query likelihood, and mixtures models. This work is of particular interest to us, as we focus on better understanding the effectiveness of query likelihood using hypergeometric document models.

Hypergeometric distributions. The univariate central hypergeometric distribution has been firstly used in the past to provide a theoretical framework for understanding performance and evaluation measures in IR [11, 35], and for proving the document-query duality [10].

Wilbur [39] was the first to use the central hypergeometric distribution in a retrieval setting. The vocabulary overlap of two docu-

ments is modeled as a hypergeometric distribution for determining the relevance to each other. Wilbur's model initially ignored local and global term weights, such as term frequencies within documents or term document frequency. Term weights are integrated into the final score only later through multiple iterations of the main model. Our retrieval models are able to support local and global term weights in a straightforward manner.

More recently, Bravo-Marquez et al. [7] derived a query reduction method for document long queries using the central hypergeometric distribution. Amati [3] used the central hypergeometric distribution within the Divergence from Randomness (DFR) framework for deriving the binomial distribution, a readily accepted distribution for the generative model. Amati's model has applications in query expansion [16], pseudo-relevance feedback [4], and enterprise search [24].

8. CONCLUSIONS AND OUTLOOK

We looked at the task of republished article finding (RAF), to discover springboards of discussion in social media related to a news article. Our approach is to find verbatim or near-verbatim copies of the news article building on the language modeling paradigm. Our task is related to near-duplicate detection with the additional challenge that in our scenario, users can inject comments in between excerpts from the original article. To this extent the documents to be retrieved can deviate considerably from the original article.

In the process of tackling the problem, we revisited the assumptions made in unigram language model, namely, using the multinomial distribution for modeling documents. We presented two retrieval models using the hypergeometric distributions, one task-driven (log odds), and one more elaborate using Bayesian inference. In the later, we found that the Dirichlet compound multinomial distribution (DCM) arises naturally for estimating the parameters of a document model. This is an important finding because it links central hypergeometric to DCM as multinomial is linked to Dirichlet. DCM has been derived in the past from hierarchical bayesian modeling techniques as a better model to Dirichlet [12, 25, 40].

Our experiments on finding republished news articles in the blogosphere demonstrate the utility and effectiveness of modeling documents using hypergeometric distributions. We found that our log odds retrieval model is most useful for documents whose size is similar to the query size.

In future work, we envisage to study more in depth different smoothing methods suitable for the hypergeometric distributions and compare them to the multinomial case. Such methods can be challenging to find as they need to meet the requirements set by the hypergeometric distribution, namely, the smoothed estimates need to be larger than those sampled. With regards to the noncentral hypergeometric distribution, we aim at exploring more elaborate ways of incorporating term bias, such as term co-occurence between the document and query. In the long term, we believe that our methods based on hypergeometric distributions hold promise to support grouping of individual news stories into topics, providing support for impact analysis.

Finally, our republished article finding task was formulated in the setting of online reputation management (ORM). ORM is related to search engine optimization, but the two do not coincide and their goals differ widely. ORM gives for a number of recall-oriented retrieval tasks: republished article finding is one, dealing with "creative" name variants and implicit references to a given target in social media is another important example.

Acknowledgments. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

9. REFERENCES

[1] O. Abdel-Hamid, B. Behzadi, S. Christoph, and M. R. Henzinger. Detecting the Origin of Text Segments Efficiently. In *Proceedings of the 18th World Wide Web Conference*, pages 61–70, New York, 2009. ACM. Henzinger M.

[2] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[3] G. Amati. Frequentist and bayesian approach to information retrieval. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 13–24. Springer Berlin / Heidelberg, 2006.

[4] G. Amati. Information theoretic approach to information extraction. In H. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreasen, and H. Christiansen, editors, *Flexible Query Answering Systems*, volume 4027 of *Lecture Notes in Computer Science*, pages 519–529. Springer Berlin / Heidelberg, 2006.

[5] A. Barrón-Cedeño, P. Rosso, and J.-M. Benedí. Reducing the plagiarism detection search space on the basis of the kullback-leibler distance. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09, pages 523–534, Berlin, Heidelberg, 2009. Springer-Verlag.

[6] M. Bendersky and W. B. Croft. Finding text reuse on the web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 262–271, New York, NY, USA, 2009. ACM.

[7] F. Bravo-Marquez, G. L'Huillier, S. Ríos, and J. Velásquez. Hypergeometric language model and Zipf-like scoring function for web document similarity retrieval. In E. Chavez and S. Lonardi, editors, *String Processing and Information Retrieval*, volume 6393 of *Lecture Notes in Computer Science*, pages 303–308. Springer Berlin / Heidelberg, 2010.

[8] A. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, Washington, DC, USA, 1997. IEEE Computer Society.

[9] H.-C. Chang, J.-H. Wang, and C.-Y. Chiu. Finding event-relevant content from the web using a near-duplicate detection approach. In *IEEE / WIC / ACM International Conference on Web Intelligence*, pages 291–294, Los Alamitos, CA, USA, 2007. IEEE Computer Society.

[10] L. Egghe and R. Rousseau. Duality in information retrieval and the hypergeometric distribution. *Journal of Documentation*, 53(5):488–496, December 1997.

[11] L. Egghe and R. Rousseau. A theoretical study of recall

and precision using a topological approach to information retrieval. *Inf. Process. Manage.*, 34:191–218, January 1998.

[12] C. Elkan. Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 289–296, New York, NY, USA, 2006. ACM.

[13] A. Fog. Calculation methods for wallenius' noncentral hypergeometric distribution. *Communications in Statistics - Simulation and Computation*, 37(2):258–273, 2008.

[14] H. Garcia-Molina, L. Gravano, and N. Shivakumar. dscam: Finding document copies across multiple databases. *Parallel and Distributed Information Systems, International Conference*, 1996.

[15] S. Geva and A. Trotman. Inex 2010 Link-The-Wiki Track, 2010. `http://www.inex.otago.ac.nz/`.

[16] B. He and I. Ounis. Combining fields for query expansion and adaptive query expansion. *Information Processing & Management*, 43(5):1294 – 1307, 2007. Patent Processing.

[17] M. Henzinger. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 284–291, New York, NY, USA, 2006. ACM.

[18] D. Hiemstra and W. Kraaij. Twenty-one at TREC-7: ad-hoc and cross-language track. In *Proceedings of the seventh Text Retrieval Conference (TREC)*, volume 500 of *NIST Special Publications*, pages 227–238. US National Institute of Standards and Technology, 1999.

[19] D. Ikeda, T. Fujiki, and M. Okumura. Automatically linking news articles to blog entries. In *AAAI Spring Symposium*, 2006.

[20] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. John Wiley & Sons, New York, 1997.

[21] J. Kim, K. Candan, and J. Tatemura. Organization and tagging of blog and news entries based on content reuse. *Journal of Signal Processing Systems*, 58:407–421, 2010.

[22] J. W. Kim, K. S. Candan, and J. Tatemura. Efficient overlap and content reuse detection in blogs and online news articles. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 81–90, New York, NY, USA, 2009. ACM.

[23] O. Kolak and B. N. Schilit. Generating links by mining quotations. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, HT '08, pages 117–126, New York, NY, USA, 2008. ACM.

[24] C. Macdonald and I. Ounis. Using relevance feedback in expert search. In G. Amati, C. Carpineto, and G. Romano, editors, *Advances in Information Retrieval*, volume 4425 of *Lecture Notes in Computer Science*, pages 431–443. Springer Berlin / Heidelberg, 2007.

[25] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 545–552, New York, NY, USA, 2005. ACM.

[26] G. S. Manku, A. Jain, and A. Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 141–150, New York, NY, USA, 2007. ACM.

[27] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242, 2007.

[28] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 214–221, New York, NY, USA, 1999. ACM.

[29] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *CIKM '08*, pages 509–518, 2008.

[30] D. S. Moore. *The Basic Practice of Statistics with Cdrom*. W. H. Freeman & Co., New York, NY, USA, 2nd edition, 1999.

[31] K. Muthmann, W. M. Barczyński, F. Brauer, and A. Löser. Near-duplicate detection for web-forums. In *Proceedings of the 2009 International Database Engineering & Applications Symposium*, IDEAS '09, pages 142–151, New York, NY, USA, 2009. ACM.

[32] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[33] S. E. Robertson. *The probability ranking principle in IR*, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[34] J. Seo and W. B. Croft. Local text reuse detection. SIGIR '08, pages 571–578, New York, NY, USA, 2008. ACM.

[35] W. M. Shaw, R. Burgin, and P. Howell. Performance standards and evaluations in ir test collections: Vector-space and other retrieval models. *Information Processing & Management*, 33(1):15 – 36, 1997.

[36] E. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. In *Fourth ACM Web Search and Data Mining (WSDM)*, Hong Kong, February 2011. ACM.

[37] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, November 16-19, 2004*, volume Special Publication 500-261, 2004. National Institute of Standards and Technology (NIST).

[38] K. T. Wallenius. Biased sampling; the noncentral hypergeometric probability distribution. Technical report, Stanford University, November 1963.

[39] W. J. Wilbur. Retrieval testing with hypergeometric document models. *J. Am. Soc. Inf. Sci.*, 44:340–351, July 1993.

[40] Z. Xu and R. Akella. A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 427–434, New York, NY, USA, 2008. ACM.

[41] H. Zaragoza, D. Hiemstra, and M. Tipping. Bayesian extension to the language model for ad hoc information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 4–9, New York, NY, USA, 2003. ACM.

[42] C. Zhai and J. 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*, SIGIR '01, pages 334–342, New York, NY, USA, 2001. ACM.

[43] Q. Zhang, Y. Zhang, H. Yu, and X. Huang. Efficient partial-duplicate detection based on sequence matching. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 675–682, New York, NY, USA, 2010. ACM.