Hypergeometric Language Models for Republished Article Finding (Abstract)[*]

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 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 (ORM).

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, we have come across instances where bloggers want to share a news item with colleagues or students or where a blogger aims to kick off a discussion around the original news article within his own online community, or where someone uses excerpts from a news article as references in a post where they discuss their own opinion. 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.

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 Table 1 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 (Table 2).

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 [1, 4]. 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 [3]. 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 without replacement become less and less accurate [2]. In

[*] The full version of this paper appeared in *SIGIR 2011* [5].

DIR'12, February 23–24, 2012, Ghent, The Netherlands.

Table 1: 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$^{\vartriangle}$	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$^{\vartriangle}$	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$^{\blacktriangledown}$

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 . 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. 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)). We present 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.

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. Our experiments on finding republished news articles in the blogosphere demonstrate the utility and effectiveness of modeling documents using hypergeometric distributions, and provide a positive answer to this question; see Table 1, and 2. Our log odds retrieval model is found most useful for documents whose size is similar to the query size.

Table 2: 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$^{\blacktriangledown}$

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.

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.

2. REFERENCES

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

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

[3] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, New York, NY, USA, 1998. ACM.

[4] E. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. In *WSDM 2011*, Hong Kong, February 2011. ACM.

[5] E. Tsagkias, M. de Rijke, and W. Weerkamp. Hypergeometric language models for republished article finding. In *SIGIR 2011*, pages 485–494, Beijing, China, 2011. ACM.