Inf Retrieval (2011) 14:390–412
DOI 10.1007/s10791-010-9147-3

A multi-collection latent topic model for federated search

Mark Baillie · Mark Carman · Fabio Crestani

Received: 5 February 2010 / Accepted: 20 September 2010 / Published online: 7 October 2010

Abstract Collection selection is a crucial function, central to the effectiveness and efficiency of a federated information retrieval system. A variety of solutions have been proposed for collection selection adapting proven techniques used in centralised retrieval. This paper defines a new approach to collection selection that models the topical distribution in each collection. We describe an extended version of latent Dirichlet allocation that uses a hierarchical hyperprior to enable the different topical distributions found in each collection to be modelled. Under the model, resources are ranked based on the topical relationship between query and collection. By modelling collections in a low dimensional topic space, we can implicitly smooth their term-based characterisation with appropriate terms from topically related samples, thereby dealing with the problem of missing vocabulary within the samples. An important advantage of adopting this hierarchical model over current approaches is that the model generalises well to unseen documents given small samples of each collection. The latent structure of each collection can therefore be estimated well despite imperfect information for each collection such as sampled documents obtained through query-based sampling. Experiments demonstrate that this new, fully integrated topical model is more robust than current state of the art collection selection algorithms.

Keyword Distributed information retrieval · Topic models · Retrieval · Collection selection

1 Introduction

Distributed information retrieval (DIR) encompasses a body of research investigating solutions for searching online content which cannot be discovered using the standard Web

M. Baillie
Department of Computer and Information Sciences, University of Strathclyde, Glasgow, Scotland, UK

M. Carman (✉) · F. Crestani
Faculty of Informatics, University of Lugano, Lugano, Switzerland
e-mail: mark.carman@usi.ch

crawling techniques (Callan 2000). This content is often referred to as the *deep* (Madhavan et al. 2008) or *hidden Web* (Price and Sherman 2001), since it lies buried behind Web forms and text search interfaces.[1] The hidden Web contains a wide variety of content from academic research libraries to online retail Web sites, whose content is often generated dynamically in response to user queries.

The aim of a DIR system (also known as federated search (Avrahami et al. 2006) or selective meta-search (Craswell et al. 2004)), is to retrieve documents from a set of distributed collections through a centralised broker. To enable retrieval, the broker maintains a representative description of the content held in each collection. When cooperation with a particular collection is possible, content statistics can be accessed through a shared protocol (Gravano et al. 1997; Paepcke et al. 2000). Typically cooperation cannot be guaranteed and techniques such as *query-based sampling* (Callan and Connell 2001) or *focused probing* (Gravano et al. 2003) are used to obtain a representative sample of documents from the collection. Sampling is terminated when it is believed a sufficiently good representation of the underlying collection has been acquired that facilitates effective retrieval (Avrahami et al. 2006). The index maintained by the broker is required for both collection selection and results merging. The form it takes depends on the underlying retrieval model used for collection selection, with potential index representations including the *big document model* (basic term statistics across the whole sample) (Xu and Croft 1999; Si et al. 2002), the *small document model* (term statistics for each document in the sample) (Si and Callan 2003), or a *hierarchical topical summary* (where the sample is classified into a subject category from a hand-crafted taxonomy) (Gravano et al. 2003) or the full collection index (Callan 2000).

In this paper we address the problem of collection selection using a principled hierarchical Bayesian modelling approach commonly referred to as *topic modelling* (Blei et al. 2003; Griffiths and Steyvers 2004; Wei and Croft 2006; Wallach 2008). We introduce a new model for collection selection which combines the best features of all three indexing approaches, namely the ability to calculate robust term statistics across a sample of documents, the ability to use all information (including document boundaries) within the sample, and the ability to leverage term statistics from other topically-related samples from other collections. We do this by estimating the parameters of a multiple collection latent topic model of text. This generative process enables us to model the latent topic structure (major themes) between documents both within and across collections. Modelling the hidden thematic structure within each collection and leveraging it for collection selection has a number of distinct advantages over simpler approaches:

- By recovering the topic distribution for each collection, we can estimate which collection is most likely to contain documents that are relevant to the topic of the query (rather than the collection is most likely to contain documents with the same terms as the query).
- The parameters of the generative model are estimated based on the co-occurrence of words across documents and collections. Thus the term statistics for each sample are implicitly smoothed using the statistics of topically related samples, making the prediction robust to small sample sizes. Moreover, the use of co-occurrence

[1] An online deep-web resource is an information collection that is searchable. This could be a free-text or Boolean search system, relational database, etc. We only make the assumption that a site will have a discoverable search text box. Therefore, solutions such as sampling via queries submitted to a interface are adopted for indexing deep-web content rather than crawling (Madhavan et al. 2008; Callan and Connell 2001; Ipeirotis et al. 2006; Bar-Yossef and Gurevich 2006).

information addresses to a certain extent problems of synonymy and polysemy (Blei et al. 2003; Griffiths and Steyvers 2004). We note that synonymy, polysemy and missing vocabulary are particularly important problems for federated search because of the small samples of documents that are used to represent very large collections.
- The generative modelling approach results in a collection selection algorithm that in theory requires no parameter tuning, since all parameters (including hyperparamters) of the generative process can be chosen so as to maximise model fit on the sampled documents. This is in contrast to some state-of-the-art approaches, e.g. that of Shokouhi (2007), which contain arbitrary parameter settings that need to be chosen based on training data (a query log and relevance judgements).
- The Bayesian framework allows us to include additional information such as a collection size estimate into the ranking function in a consistent and coherent manner.
- The topic-based characterisation of each collection can be used to find important terms for query-based sampling of collections or to compare different collections in terms of topic prevalence. The latter being important for visualising the content of different collections.

The paper is structured as follows. In the next section we review the current state-of-the-art centralised-index-based collection selection algorithms and investigate further motivations for our topic modelling approach. We then introduce two topic models, latent Dirichlet allocation (LDA) and a hierarchical extension designed to take document groupings (i.e. collections) into account and discuss how these models can be used for collection selection. We evaluate this new approach to collection selection, comparing performance with a number of existing methods. Finally, we discuss the implications of this study before concluding the paper and outlining future work directions.

2 Previous approaches to collection selection

Collection selection is a critical function of a DIR system in which the broker attempts to route queries only to those collections which (potentially) contain relevant information. Collection selection can be summarised into two phases: the first phase ranks collections with respect to the user query, where the ordering reflects how likely a collection is to contain relevant information (the expected density of relevant documents in the collection). Depending on this ranking, the second phase determines which collections to route the query to and how many documents to retrieve from each. After this second phase, the retrieved documents from all searched collections are merged into a single coherent ranked list to present to the user. Typical merging strategies involve the normalisation of local collection relevance scores from the retrieved documents (Callan et al. 1995; Si and Callan 2003).

A number of solutions have been proposed for collection selection, which can be broadly grouped into two categories: *big-document* and *centralised sample index* approaches. The first category is so called because collections are represented by a large virtual document which is the concatenation of the acquired set of representative sampled documents. Analogous to standard document retrieval, the collection-representative documents can be ranked with respect to a query using a retrieval algorithm. Therefore big-document approaches essentially differ by how the resource descriptions are ranked, for example, using a bayesian inference network (CORI) (Callan et al. 1995), the vector space model (vGlOSS) (Gravano et al. 1999), or language models (Xu and Croft 1999; Si et al. 2002).

The decision to remove document boundaries within the representation set of documents is thought to impact on collection selection performance (Xu and Croft 1999), with a number of recent empirical studies supporting this claim (Si and Callan 2003; Hawking and Thomas 2005; Thomas and Hawking 2009). As a consequence, a new group of techniques called centralised sample index algorithms have been proposed which retain document boundaries (Si and Callan 2003; Hawking and Thomas 2005; Shokouhi 2007). The sampled documents obtained from each collection are indexed centrally at the broker to form a partial centralised index, which is an approximation of the global virtual collection index. Thus, given a user query, documents in the sampled index are first ranked. This document ranking is then used to predict which collections have the largest number of relevant documents, informing the decision process for selecting the subset of collections to search.

We note that the problem of collection selection is a critical function not only for DIR but also arises in a number of other contexts including expert search (Balog 2008) (where each "collection" contains documents regarding a particular person), and in blog search (Elsas et al. 2008) (where blog posts are considered to be a sample of the documents that a blog author could write). A fundamental goal for any collection selection algorithm is to rank collections, experts or blogs by the expected density of relevant documents.

2.1 Language modelling framework

We will now discuss in more detail state-of-the-art approaches to resource selection, introducing the formulae that we will both compare with as a baselines and extend using our topic modelling approach. We base our explanation on the language modelling framework for IR (Manning et al. 2008), since it allows for ease of comprehension and because many collection selection algorithms can be easily reformulated in the Bayesian setting. In this framework, collections are ranked according to their likelihood given a query. The likelihood of a collection c given a query q can be calculated using Bayes rule as follows:

$$P(c|q) = \frac{P(q|c)P(c)}{P(q)} \propto P(q|c)P(c) \tag{1}$$

where $P(q|c)$ is the likelihood that collection c generates query q, $P(c)$ is the (query independent) prior probability of retrieving a document from the collection and $P(q)$ is a normalising constant (the collection-independent query prior) which can be dropped from the calculation without effecting the overall ranking of collections. Assuming all documents are equally likely, we can estimate the value collection prior using the relative size of the collection:

$$P(c) = \frac{\hat{\mathcal{D}}_c}{\sum_{c=1}^{C} \hat{\mathcal{D}}_c} \propto \hat{\mathcal{D}}_c \tag{2}$$

where $\hat{\mathcal{D}}_c$ is the estimated size (in documents) of collection c. In uncooperative environments the collection size is estimated using population estimation techniques such as sampling-resampling (Si and Callan 2003). Larger collections are assigned more weight prior to ranking resources making explicit the assumption that bigger collections are more likely to contain relevant information. Alternatively, the collection prior can be calculated based on the expected prior usefulness of the collection, which can be estimated from training data. Other sources of evidence that could be leveraged for calculating

query-independent priors for collections include hyper-link and anchor text evidence when it is available (Hawking and Thomas 2005).

Combining (1) and (2) gives the following simple ranking function, where the query likelihood for each collection still needs to be specified.

$$P(c|q) \propto \hat{\mathcal{D}}_c P(q|c) \tag{3}$$

2.2 Big-document approaches

In the big-document approach, the set of sampled documents from a collection are concatenated to form a large virtual document. The query likelihood can then be approximated using term statistics for this large document by applying the Naive Bayes conditional independence assumption:

$$\hat{P}(q|c) = \prod_{w \in q} P(w|c) \tag{4}$$

where $P(w|c)$ denotes the probability of word w in collection c. Simple approximations of $P(w|c)$ include the relative term frequency in the big document or the relative document frequency in the sample. Usually the maximum likelihood estimates are smoothed in order to deal with the zero probability problem for example using Jelinek-Mercer smoothing against a background distribution containing the union of all samples (Si et al. 2002).

$$\hat{P}(w|c) = \lambda P_{\mathtt{ML}}(w|c) + (1-\lambda) P_{\mathtt{ML}}(w|\cup_i c_i) \tag{5}$$

where $P_{\mathtt{ML}}$ denotes a maximum likelihood estimate and λ is a smoothing parameter.

Ipeirotis and Gravano (2008) noted that according to Zipf's law, a sample of a collection sample will fail to recover a large proportion of terms which occur less frequently in the collection than the sampling rate, but may nonetheless be important terms for defining the topics present in the collection. As a consequence, short queries or queries containing infrequent terms not represented in the sample, but important to a collection, will affect retrieval performance. Thus smoothing techniques have been explored which attempt to exploit the observation that similar (potentially topically related) collections share similar vocabularies.

For example, Xu and Croft (1999) smoothed the term distribution of topically grouped documents using Laplace smoothing. Topics were generated by first clustering documents. Topics were then represented by a smoothed language model (i.e. $P(w|c)$) where a small constant probability mass was assigned to terms not occurring in the cluster set of documents for a topic.

Ipeirotis and Gravano (2008) applied shrinkage to estimate $P(w|c)$ by first classifying collections into a topical hierarchy. Collections are then smoothed based on this topical hierarchy. Shrinkage over a topic hierarchy smoothes the collection estimate not with a global collection but with a set of topically related collections in a classification T:

$$\hat{P}(w|c) = \lambda_0 P_{\mathtt{ML}}(w|T_c) + \sum_{i=1}^{m} \lambda_i P_{\mathtt{ML}}(w|T_i) \tag{6}$$

such that $\sum_{i=0}^{m} \lambda_i = 1$. Instead of a fixed smoothing weight, the λ_i mixture weights are estimated by expectation-maximisation over training data.

Despite the elegance of their approach, there are three obvious disadvantages of their smoothing method, which are not shared by our topic modelling approach to collection

selection, namely: (1) that a hierarchy of content areas must be defined in advance; (2) that each collection may only belong to a single topic node in the hierarchy; and (3) that both the topic assignments and smoothing weights must be learnt from training data.

2.3 Small-document approaches

In the small document approach, exemplified by the ReDDE (Si and Callan 2003) algorithm, the sampled documents from each collection are not concatenated but indexed individually along with documents sampled from other collections to form a "centralised sample index" that approximates the unified index over all documents in the different collections. The documents in the sample index are ranked for each query, and based on this ranking density of relevant documents in each collection is estimated.

From a Bayesian perspective, estimating the likelihood of relevant documents in each collection is equivalent to marginalising over the documents in the sample. Thus the small document model can be written as (Elsas et al. 2008):

$$\hat{P}(q|c) = \sum_{d \in c} P(q|d)P(d|c) \tag{7}$$

Here $P(q|d)$ denotes the likelihood of a query given a document, which is essentially the "retrieval score" for the document and could be estimated using standard language modelling smoothing techniques. $P(d|c)$ is the probability of a document in a collection. It can either be set to the uniform distribution ($\hat{P}(d|c) = 1/D_c$) or be estimated as a measure of the representativeness of a document in a collection, for example using the geometric mean of a term in a collection (Elsas et al. 2008). (The latter has been shown to provide a robust model for ranking blogs.) For ease of comparison with other methods we will assume a uniform distribution over documents in the sample.

The likelihood of the query given the document $P(q|d)$ can be estimated using a variety of smoothing methods including Jelinek-Mercer smoothing:

$$\hat{P}(q|d) = \prod_{w \in q} (\lambda_1 P_{\mathrm{ML}}(w|d) + \lambda_2 P_{\mathrm{ML}}(w|c) + \lambda_3 P_{\mathrm{ML}}(w| \cup_i c_i)) \tag{8}$$

where $P_{\mathrm{ML}}(w|d)$ is the relative frequency of term w in document d, $P_{\mathrm{ML}}(w|c)$ is the relative frequency across all documents in collection c, and λ_i is a smoothing parameter such that $\sum_i \lambda_i = 1$.

The original ReDDE algorithm (Si and Callan 2003) involved a different estimate for $P(q|d)$, which they denoted $P(rel|d)$, meaning the probability of relevance of a particular document. We include their original estimate for comparison purposes and use it as an additional baseline for the experiments. In this case, the query likelihood is estimated by ranking all the documents contained in the centralised sample index using a document retrieval algorithm such as a vector space, InQuery, BM25 or language model.

$$\hat{P}(q|d) \propto \begin{cases} 1 & \text{if rank}(d) < nD/\hat{\mathcal{D}} \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

where n is the desired number of relevant documents that should be found, D is the total number of documents across all samples in the index and $\hat{\mathcal{D}}$ is the estimated size of all collections combined.

Recently, a central-rank based collection selection (CRCS) algorithm has been proposed by Shokouhi (2007). As with ReDDE, first documents in the sample are ranked using a

document retrieval model. CRCS then measures the proportion of documents from each collection that are highly ranked in the centralised sample index with respect to the query and uses it to estimate of the proportion of relevant documents likely to be held in each collection. Two CRCS formulas have been shown, empirically, to improve performance over ReDDE on a number of testbeds. The CRCS algorithms can be seen simply as two different estimates for the query likelihood that rely on a document's rank within the index rather than its retrieval score. The first estimate is dependent on the negated rank:

$$\hat{P}(q|d) \propto \begin{cases} \gamma - \text{rank}(d) & \text{if rank}(d) < \gamma \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

While the second is dependent on a weighted exponent of the negated rank:

$$\hat{P}(q|d) \propto \exp(-\beta \, \text{rank}(d)) \tag{11}$$

The parameters γ and β need to be tuned on labeled training data (relevance judgements) during an initial training phase. We will refer to these estimates in the experiments as CRCS(l) and CRCS(e), respectively.

2.4 Recap and motivation

After reviewing the existing approaches to collection selection we have identified a number key desirable properties a collection selection algorithm should consider as well as a number of limitations with prior research: (1) a model should be able to account for incomplete information, term disambiguation and vocabulary smoothing under a single framework; (2) model coherence, i.e. the approach should model the problem at hand directly; (3) the model should estimate the topical relatedness of collection. Previous work required an ontology and labelled collections for training the classifiers for this purpose (Ipeirotis and Gravano, 2008). However, in this paper we want to infer any possible structure from the data itself; and (4) prior information, the approach should be able to include other evidence sources. In the following section we define a new model for collection selection based on latent topic modelling.

3 Latent topic models

Topic modelling (Griffiths and Steyvers 2004) is an active area of research which combines ideas from dimensionality reduction techniques like Latent Semantic Indexing (LSI) and its probabilistic reformulation Probabilistic Latent Semantic Indexing (PLSI) (Hofmann 1999) with generative modelling techniques using Bayesian Networks approaches (Buntine 1994). In topic modelling, documents are represented by a distribution over a semantic topic space where each topic is characterized by a distribution over words. The reliance on Bayesian techniques for developing these models prevents them from overfitting the data and allows them to generalise well to unseen documents. These techniques have recently been applied to a number of problems in IR including the modelling and tracking of scientific publications (Griffiths and Steyvers 2004) and document retrieval (Wei and Croft 2006).

There are a number of different ways that the topics of a collection could be estimated including the use of document clustering techniques (Manning et al. 2008) and Probabilistic Latent Semantic Indexing (PLSI) (Hofmann 1999). We examine in this paper

$\mathcal{D}_c$	Size (in documents) of collection c
D_c	Size (in documents) of sample of collection c
N_d	Length (in words) of document d
N_c	Total length (in words) of sample of collection c
N_z	Total occurrences of topic z across all samples
$N_{w,z}$	Occurrences of word w in topic z
$N_{z,d}$	Occurrences of topic z in document d
$N_{z,c}$	Occurrences of topic z in sample of collection c
V	Vocabulary: distinct words across all samples
Z	Number of topics
C	Number of collections
$\phi_{w\|z}$	Probability that word w is chosen for topic z
$\theta_{z\|d}$	Probability that topic z is chosen in document d
$\psi_{z\|c}$	Probability that topic z is chosen in collection c
β	Hyperparameter smoothing word distributions
α, α_i	Hyperparameters smoothing topic distributions
R_k	Recall-based metric for top k collections
P_k@10	Precision@10-based metric for top k collections

Fig. 1 Main notation used in this paper

methods based on probabilistic topic models (Griffiths and Steyvers 2004). Latent Dirichlet allocation (LDA) (Blei et al. 2003) is the most frequently used topic model. It is a probabilistic generative model for documents within a collection, where each document is modelled as a mixture of topics and each topic is a distribution over terms. More specifically, LDA is a Bayesian reformulation of PLSI where Dirichlet prior probability distributions over model parameters (distributions over topics for each document and distributions over terms for each topic) are used to prevent over-fitting of the model to the data, and thereby allow for good generalisation to unseen documents. The ability of LDA to generalise well to unseen documents is critical for the application to collection selection, where the description of each resource must be induced from a small sample of the documents present in the whole collection (Fig. 1).

LDA has been applied to the problem of modelling topics in text corpora, including modelling and tracking the development of scientific topics (Griffiths and Steyvers 2004); classification, collaborative filtering (Blei et al. 2003), and retrieval (Wei and Croft 2006) amongst others. The LDA model specifies how a document may have been generated, the underlying assumption being that documents are mixture of (sub-)topics. Representing concepts as probabilistic topics enables each topic to be interpretable and thereby presentable to the user.

In the following section we will briefly describe LDA as it pertains to the current work. We will follow the formulation of LDA given by Griffiths and Steyvers (2004) and the notation used by Wallach (2008). A list of the notation used throughout the paper is given in Table 1. In Sect. 3.2 we will outline a multi-collection topic model which models the latent topical relationship of documents in and across collections.

3.1 Latent Dirichlet allocation

Figure 2 shows LDA as a graphical model[2] using plate notation (Buntine 1994). In the generative model each document is modeled as a distribution over topics θ_d that is used to

[2] Graphical models are used to represent the dependence between random variables in a statistical model (such as a Bayesian Network). Random variables are shown as labeled circles and arrows denote dependence between them. Shaded circles represent observed variables, while unshaded circles represent latent variables. A box denotes a repeated structure, where the value in the bottom right of the box is the cardinality of the repetition.

Table 1 The most likely terms in three generated topics of a MCTM model

Topic 2		Topic 7		Topic 30	
Crime	0.028	Economic	0.017	Fish	0.025
Law	0.018	Development	0.015	Fishery	0.024
Court	0.015	Government	0.011	Species	0.023
Attorney	0.014	System	0.009	Marine	0.022
Enforce	0.014	Increase	0.008	Vessel	0.014
Criminal	0.013	Policy	0.008	Action	0.012
Violence	0.012	Economy	0.007	Permit	0.011
Victim	0.012	Change	0.007	Management	0.010
Justice	0.011	Country	0.007	Service	0.010
Prison	0.011	Nation	0.007	Population	0.010

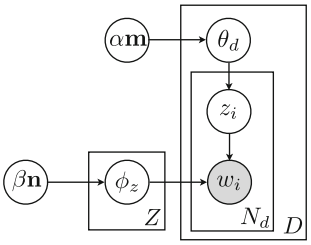

Fig. 2 Graphical model for latent Dirichlet allocation. Variable $w_i \in \{1,\ldots,V\}$ represents the ith word in the corpus (all documents concatenated together), where V is the vocabulary of the collection. N_d is the length of the dth document, $d \in \{1,\ldots,D\}$. Variable $z_i \in \{1,\ldots,Z\}$ denotes the hidden topic assignment of the ith word. For each document, we have a variable θ_d that defines the probability distribution over topics $\{1,\ldots,Z\}$ from which the values z_i are chosen. For each topic, ϕ_z gives a probability distribution over terms in the vocabulary $\{1,\ldots,V\}$, according to which w_i is chosen. The distributions θ_d and ϕ_z are selected using a Dirichlet distribution with parameters $\alpha\mathbf{m}$ and $\beta\mathbf{n}$, where $\mathbf{m}$ and $\mathbf{n}$ are uniform distributions over Z and V, respectively. The hyperparameters α and β determine to what extent the sampled distributions vary from the uniform prior

choose the words in the document according to a distribution over terms for each topic ϕ_z. The probability that a particular topic z will be emitted by a document d is denoted θ_{zld} and the probability that a vocabulary word w will chosen for a topic z is denoted ϕ_{wlz}. According to the generative model, the likelihood of a corpus of documents, denoted $\mathbf{w} = \langle w_1,\ldots,w_N\rangle$ (which consists of all documents concatenated together) and an assignment of values to the hidden topic variables $\mathbf{z} = \langle z_1,\ldots,z_N\rangle$, given the model parameters $\Phi = \{\phi_z\}_{z=1}^{Z}$ and $\Theta = \{\theta_d\}_{d=1}^{D}$, is then given by:

$$P(\mathbf{w},\mathbf{z}|\Phi,\Theta) = \prod_{i=1}^{N} \phi_{w_i|z_i}\theta_{z_i|d_i} \tag{12}$$

where N is the length of the corpus (in word occurrences) and d_i is the document associated with the ith position in the corpus. Z and D denote the number of topics and documents, respectively.

In LDA, the model parameters ϕ_z (the topic term distribution) and θ_d (the document topic distribution) are themselves chosen according to a Dirichlet distribution:

$$\phi_z \sim Dirichlet(\beta \mathbf{n}) \tag{13}$$

$$\theta_d \sim Dirichlet(\alpha \mathbf{m}) \tag{14}$$

where $\mathbf{n}$ and $\mathbf{m}$ are uniform distributions over words and topics, respectively. Thus, $n_i = \frac{1}{V}$ and $m_i = \frac{1}{Z}$, where V is the size of the vocabulary and Z is the number of topics. The posterior estimate given the data (the corpus $\mathbf{w}$ and topic assignments $\mathbf{z}$) for the model parameter $\phi_{w|z}$ (the probability of topic z producing word w) is:

$$\hat{\phi}_{w|z} = \frac{N_{w,z} + \beta \frac{1}{V}}{N_z + \beta} \tag{15}$$

where $N_{w,z}$ is the number of occurrences of vocabulary word w for topic z and N_z denotes the number of times topic z occurs in the corpus as a whole. Similarly, the posterior estimate given the data for the model parameter $\theta_{z|d}$ (the probability of document d emitting topic z) is given by:

$$\hat{\theta}_{z|d} = \frac{N_{z,d} + \alpha \frac{1}{Z}}{N_d + \alpha} \tag{16}$$

where $N_{z,d}$ is the number of occurrences of topic z in document d and N_d is the length of the document.

For an accurate estimation of the coverage of topics in a sample with respect to the collection, a good representation of the collection is required using LDA. As exact inference using LDA is intractable, we use the approximate inference approach defined by Griffiths and Steyvers (2004) which uses Gibbs sampling to approximate the posterior distribution.

The Gibbs sampling procedure involves first generating a random assignment of values for the topic vector $\mathbf{z} = \langle z_1,\ldots,z_N \rangle$. This is followed by repeated steps of re-estimating all the values in the vector. At each iteration, the values for individual topic variables z_i are updated in turn by sampling a value from the conditional probability distribution for z_i given the word w_i, using estimates for $\phi_{w|z}$ and $\theta_{z|d}$ based on current assignments to all the other topic variables $\mathbf{z}_{\neg i} = \langle z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N \rangle$. The estimate for the conditional probability of topic variable assignment $z_i = z$ is given by:

$$\hat{P}(z_i = z | w_i = w, d, \mathbf{w}, \mathbf{z}_{\neg i}) = \frac{\hat{\phi}_{w|z} \hat{\theta}_{z|d}}{\sum_z \hat{\phi}_{w|z} \hat{\theta}_{z|d}} \tag{17}$$

The procedure is usually repeated until a preset number of iterations have been reached. The stopping criterion could also be a threshold on the delta improvement in the model Likelihood (given in (12)).

We will discuss how the LDA model can be used to rank resources for collection selection in Sect. 3.3

3.2 A multi-collection latent topic model

For the problem of collection selection, we are interested in investigating the latent topical structure within and across different resources. In its current form, LDA does not take

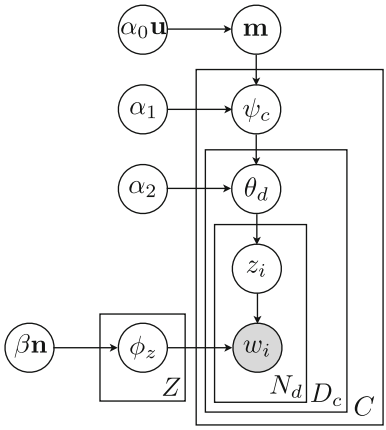

Fig. 3 Graphical model for hierarchical latent Dirichlet allocation where documents are grouped into collections. There are C collections in the corpus and D_c documents in collection c. The variable ψ_c denotes a probability distribution over topics $\{1,\ldots,Z\}$, which characterizes the collection c by defining the mean of the Dirichlet distribution that generates a document topic distribution θ_d for each document. The hyperparameter α_2 determines the amount by which the document topic distributions vary from ψ_c. The variable $\mathbf{m}$ is now the mean of the distributions over topics for the corpus as a whole and the hyperparameter α_1 determines the extent to which the different collections vary from one another. Finally, $\mathbf{u}$ is a uniform distribution over topics, and α_0 regulates how far the corpus deviates from uniform

document groupings (collections) into account and thus is not necessarily learning the best model for this task. We now investigate a more expressive model capable of dealing with document groupings in the form of multiple collections within the one corpus of documents. In particular we describe a hierarchical extension to latent Dirichlet allocation with observed document collections, that was first introduced by Wallach (2008). In this model, each document is assigned to a particular collection and the statistics of document generation (in terms of the relative frequencies of different topics) are different for the different collections. Figure 3 provides a graphical representation of the model with document groupings.

In the new model, the prior over document topic distributions is no longer constant across the whole corpus but depends on the collection that the document comes from. More specifically, the topic distribution for each document θ_d is generated by a Dirichlet distribution conditioned on a collection specific topic distribution ψ_c (the mean of the topic distributions in the collection). Moreover, the collection specific distribution is generated by a second Dirichlet distribution, conditioned on a corpus level topic distribution $\mathbf{m}$. Finally, the corpus level parameter is itself dependent on a uniform Dirichlet prior. This *multi-collection topic model* (MCTM) is defined as follows:

$$\theta_d \sim Dirichlet(\alpha_2\psi_c) \tag{18}$$

$$\psi_c \sim Dirichlet(\alpha_1\mathbf{m}) \tag{19}$$

$$\mathbf{m} \sim Dirichlet(\alpha_0\mathbf{u}) \tag{20}$$

This new model has four hyperparameters α_0, α_1, α_2 and β as opposed to LDA's two. The most interesting of the new parameters is α_1, which controls the amount to which the topic descriptions ψ_c for different collections can vary from each other. The third Dirichlet distribution (20) facilitates the modelling of collections within the corpus. The extra level

of hierarchy in the model allows for more flexibility and result in a better model fit by allowing different topics in the corpus to have quite different relative frequencies from one another (determined by the parameter **m**).

The posterior estimate for the document level topic distribution is then computed as follows:

$$\hat{\theta}_{z|d} = \frac{N_{z,d} + \alpha_2 \hat{\psi}_{z|c}}{N_d + \alpha_2} \tag{21}$$

$$\text{where} \quad \hat{\psi}_{z|c} = \frac{N_{z,c} + \alpha_1 \hat{m}_z}{N_c + \alpha_1} \tag{22}$$

$$\text{where} \quad \hat{m}_z = \frac{N_z + \alpha_0 \frac{1}{Z}}{N + \alpha_0} \tag{23}$$

Here $N_{z,c}$ denotes the total number of occurrences of topic z in the documents of collection c and N_c is the length (in words) of collection c. Using this new estimate, the Gibbs sampling algorithm for estimating model parameters is the same for MCTM as it was for LDA, (see Wallach (2008) for a derivation).

One of the advantages of MCTM over LDA when applied to the problem of collection selection is that $\hat{\psi}_{z|c}$ provides an estimate of the prevalence of a topic in the collection c. For example, Figure 4 illustrates the topical distribution of two collections estimated by a 50 topic MCTM model. The topical distribution in each collection differs significantly indicating that the topical distribution induced from each sample may provide a good representation for the differing contents of the collections. Table 1 indicates the most likely terms (and associated probability) from the prevalent topics in collection 1 (i.e. topic 2) and collection 2 (i.e. topic 30), as well as a topic with equal weight in both collections (i.e. topic 7). From these topic distributions we can gain an insight into the content held in each collection, where the two most prevalent topics appear to have very different semantic associations of law (topic 2) and aquaculture (topic 30), while both collections share a similar prevalence of a topic related to economic policy (topic 7). We can therefore utilise this information to rank collections based on a user query as well as potentially leverage the topical distributions for further query-based sampling of a collection (Callan and Connell 2001).

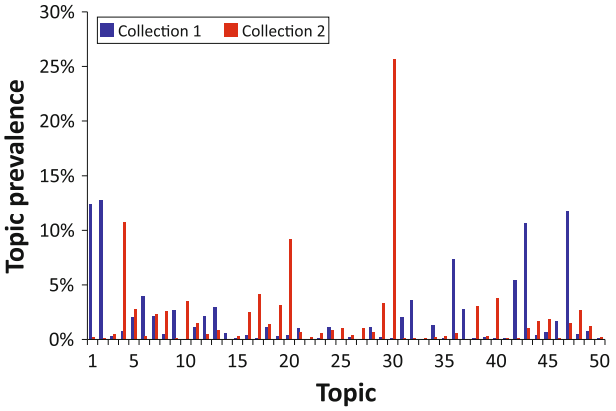

Fig. 4 An example of the distribution of topic prevalence for two distributed collections modelled by MCTM, (best viewed in colour). We see that the two collections differ greatly in terms of their topic-based characterisation with few common topics of non-neglibile probability

3.3 Topic-based collection selection

By modelling the sampled set of resource descriptions using topic models we are able to estimate the likelihood that each collection model would generate a document containing the query terms. We can then rank collections according to their likelihood given the query using (3). In other words, we will rank collections according to the number of relevant documents we expect to see in each collection, given the generative model we have learnt for each collection.

Given a trained LDA model, we derive the ranking function as follows. We first calculate the likelihood of a word w being emitted by a document d. This is calculated by summing over all topics the likelihood of that word and topic given the document:

$$\hat{P}(w|d) = \sum_{z=1}^{Z} \hat{\phi}_{w|z}\hat{\theta}_{z|d} \tag{24}$$

The likelihood of a query given a document is then simply the product of word probabilities for all terms in the query:

$$\hat{P}(q|d) = \prod_{w \in q} \hat{P}(w|d) \tag{25}$$

We can calculate the query likelihood given the collection by simply averaging the likelihood over all the documents in the collection as was done for the small document model:

$$\hat{P}(q|c) = \sum_{d \in c} \hat{P}(q|d)\hat{P}(d|c) = \frac{1}{D_c}\sum_{d \in c} \hat{P}(q|d) \tag{26}$$

Here D_c is the number of documents in (our sample of) the collection c. Note that we could have used a more complicated estimate for likelihood of a document given a collection rather than the uniform distribution $P(d|c) = \frac{1}{D_c}$, to deal with the fact that some documents are more "central" to the themes of the collection sample than others. Finally we combine (26) and (3) to rank collections for the LDA model according to:

$$\hat{P}(c|q) \propto \frac{\hat{\mathcal{D}}_c}{D_c}\sum_{d \in c}\prod_{w \in q}\sum_{z=1}^{Z} \hat{\phi}_{w|z}\hat{\theta}_{z|d} \tag{27}$$

For the hierarchical MCTM model the estimation is much simpler. We use the posterior estimate for the model parameter ψ_c (the collection level topic distribution) to calculate the query likelihood and multiply the latter by the estimated collection size to rank collections according to their likelihood:

$$\hat{P}(c|q) \propto \hat{\mathcal{D}}_c \prod_{w \in q}\sum_{z=1}^{Z} \hat{\phi}_{w|z}\hat{\psi}_{z|c} \tag{28}$$

We note that query processing for the MCTM model is faster than for the LDA model since a single probability distribution ψ_c, the collection level distribution, is used to represent the collection rather than a set of distributions for each document in the collection sample.

The astute reader may question what appears to be a "big document" approach for MCTM, since we are generating a single representation of each collection (in the topic

space) that we compare with the query. We note however, that this collection description is in fact learnt from a topic model that takes the co-occurance of words across documents within (and across) each collection sample into account. Said in another way, the dimensions of the latent topic space depend on the document boundary information within each sample. Were we to throw away that information, and start with a "big document" representation of each sample, it would indeed be impossible to learn a topic based representation of collection.[3]

4 Experiments

We now describe a series of experiments comparing the MCTM model with a number of baseline collection selection algorithms.

4.1 Baseline models

ReDDE, ReDDE-LM, CRCS(l) and CRCS(e) as well as LDA were used as a comparison to the new model. To provide consistency with previous evaluations of these baselines, the internal retrieval model used for both ReDDE and CRCS was the standard InQuery model included within the Lemur framework.[4] ReDDE-LM represents the language modelling version of ReDDE. All models used the same collection prior with complete information as a control when comparing models.

4.2 Testbeds

The collection selection algorithms were compared over a number of standard DIR testbeds[5]: Trec123-100col-bysource, Trec4-100col-bysource, Trec4-100col-global, Trec6-100col-bysource and Trec6-100col-global testbeds (Xu and Croft 1999). Each testbed holds a set of 100 collections with documents grouped either by the source and date (bysource) or by topical similarity (global). The bysource test beds have approximately equal sized, overlapping homogeneous collections, while the global test beds represent a more varied distinct and diverse set of heterogeneous collections. Each collection was represented by 300 sampled documents obtained through query-based sampling using uniform term selection, retrieving 4 documents per query submitted (Callan and Connell 2001). A centralised sampled index was generated from the resource descriptions for each testbed. The resource descriptions were stemmed using the Porter stemmer and all stopwords were removed.

4.3 Measurements

All models were compared using *short* title queries. Collection selection accuracy was measured using two metrics, the first being the recall-based R_k metric. R_k is a measure of

[3] In (28) we ignore the individual document topic distributions contained in the Θ matrix when ranking collections according to the MCTM model. There may be ways to use this information to improve ranking performance. We leave that investigation to future work.

[4] http://www.lemurproject.org/lemur/.

[5] The testbeds are accessible at http://boston.lti.cs.cmu.edu/callan/Data/.

the overall percentage of relevant documents contained in the top k collections searched (Callan and Connell 2001).

$$R_k = \frac{\sum_{i=1}^{k} E_i}{\sum_{i=1}^{k} B_i} \tag{29}$$

where E_i is the number of relevant documents in collection i ranked by a collection selection algorithm, while B_i is the number of relevant documents in collection i ranked by a perfect, oracle based ranking. In accordance with previous studies (Si and Callan 2003; Shokouhi 2007), we present the R_k value for the first 20 collections searched.

We also compared systems to an oracle baseline (Shokouhi et al. 2007; Puppin et al. 2010), which we will refer to as *relative precision*. The performance of each collection selection method was analysed using a centralised system with a complete index of all documents from all collections. The aim of this evaluation is to determine how well the collection selection models compare to the centralised approach. The ranking provided by the centralised model for the test query set is used as pseudo-relevance judgements. For the oracle baseline we used InQuery, which is also used for the document retrieval model with ReDDE and CRCS. Systems were compared using relative $P_k@10$ (Puppin et al. 2010):

$$P_k@10 = \frac{1}{10}|Top10 \cap Top10_k| \tag{30}$$

where $Top10$ denotes the 10 most relevant documents (as ranked by InQuery) for a particular query over all the collections, while $Top10_k$ denotes the 10 most relevant documents over the k highest ranking collections.

We calculate the mean and standard error for both R_k and relative $P_k@10$ over the topic set. Mean results are reported and a 95% confidence interval (CI) with Bonferroni correction[6] for multiple comparisons are displayed where possible to provide an indication of statistical precision and variance. However, for clarity of presentation we avoid displaying a CI for all models and indicate significant differences in the text when appropriate.

4.4 Model parameters

We set the model hyperparameters for LDA and MCTM using previous literature (Griffiths and Steyvers 2004). We set the α and β parameters to 0.1. Future work will investigate estimating the hyperparameters directly from the available data, for example, using a Gibbs Expectation-Maximisation approach to parameter learning (Wallach 2008).

We estimated the number of topics Z using a discriminative approach. Using the task of collection selection directly, we evaluated a number of topic sizes using a training set of data, observing the effect of varying Z on collection selection performance. For example, Figure 5 illustrates the effect of increasing Z from 100 to 500 topics on the Trec6-100col-global testbed. Results indicated that setting $Z = 500$ provided stable performance across all testbeds, although the variation in performance was minimal across this range. Future work will investigate alternative approaches to estimating Z through modelling using hierarchical Dirichlet processes (Teh et al. 2006), where a non-parametric prior is placed on Z allowing for the number of topics to be estimated directly from the data.

[6] The Bonferroni correction is used when we are testing multiple hypotheses at the same time. It involves using a significance level of α/n instead of α where n is the number of hypotheses being tested (i.e. the number of systems being compared).

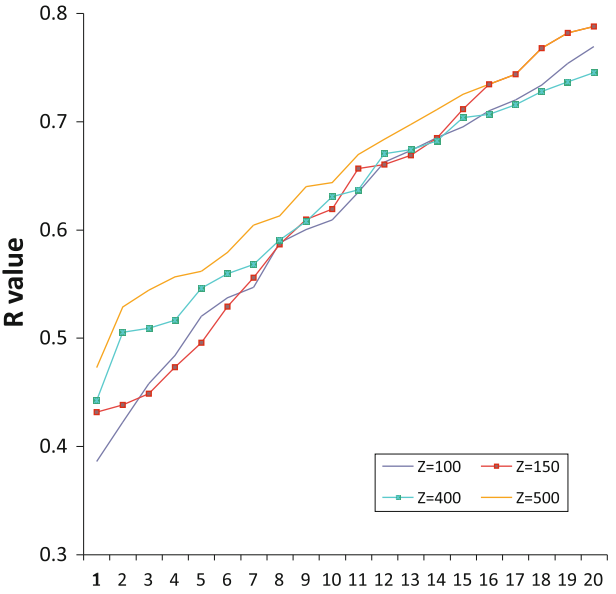

Fig. 5 Estimating the number of topics for MCTM model on the Trec6-100col-global testbed

5 Results

5.1 Homogenous testbeds

Figures 6, 7 and 8 display the results over the three homogenous bysource testbeds using both metrics. The general trend across this set of testbeds was that CRCS(e), ReDDE-LM and MCTM performed consistently better than the remaining methods. MCTM observed comparable performance with CRCS(e) and ReDDE-LM, indicating better mean performance at later cut-off k values, while CRCS(e) reported better early precision. The experiments over the bysource testbeds also indicated that MCTM was better for the task of collection selection than using non hierarchical LDA. Other trends include that ReDDE-LM was on average more consistent than ReDDE, and CRCS(e) was more stable than CRCS(l).

5.2 Heterogeneous testbeds

Figure 9 and 10 present the results of the two heterogeneous global testbeds. Again comparable performance was comparable between the MCTM, CRCS(e) and ReDDE-LM models. ReDDE-LM and CRCS(e) performed better at lower cut-offs for k while MCTM improved over both models at higher cut-offs.

5.3 Performance across test beds

We also examined the consistency of techniques across test beds. To do so, we standardised the performance scores using the technique outlined by Webber et al. (2008). This is a standard statistical technique used in meta-analysis, where the performance scores for

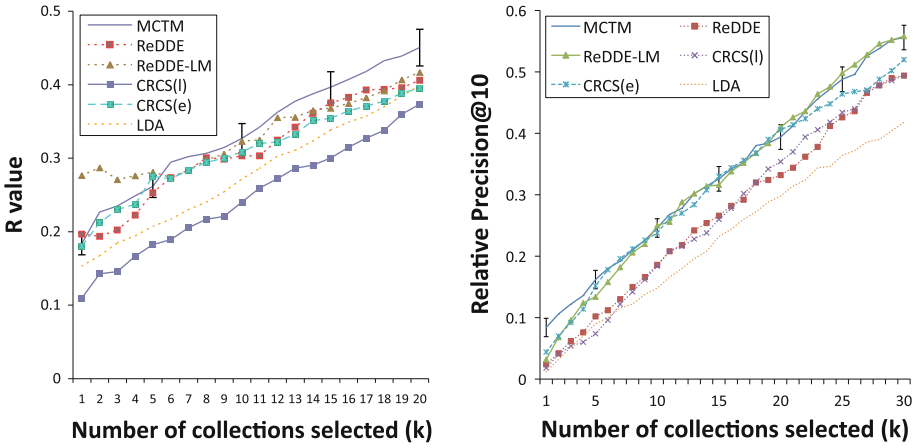

Fig. 6 Comparison over the TREC123-100col-bysource using the R_k (*left*) and the relative P_k@10 (*right*) metrics

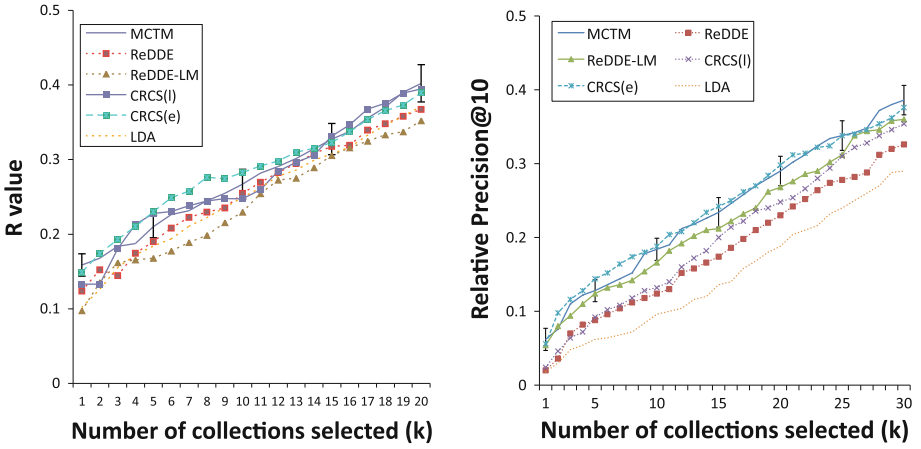

Fig. 7 Comparison over the TREC4-bysource testbed using the R_k (*left*) and the relative P_k@10 (*right*) metrics

each method is transformed into a standard Normal distribution with mean zero and unit variance, producing a dimensionless quantity irrespective of topic or testbed. The transformed scores are also known as a 'z' or standard score. The standardised scores are then converted onto a [0,1] scale as probabilities through cumulative density function for the standard Normal distribution. The final transformed performance scores are therefore comparable across testbeds. By adopting these technique we are able to analyse the consistency of each method across the different test beds.

Figures 11 and 12 present the output of such an analysis for the *R*-value and relative P@10 metrics, respectively. To simplify the analysis, we evaluated each technique at a single cut off point of ten collections selected. Figure 11 and 12 (top) presents the normalised scores from each test bed for all methods. Figures 11 and 12 (bottom) presents the combined normalised scores across collections, allowing for an approximate

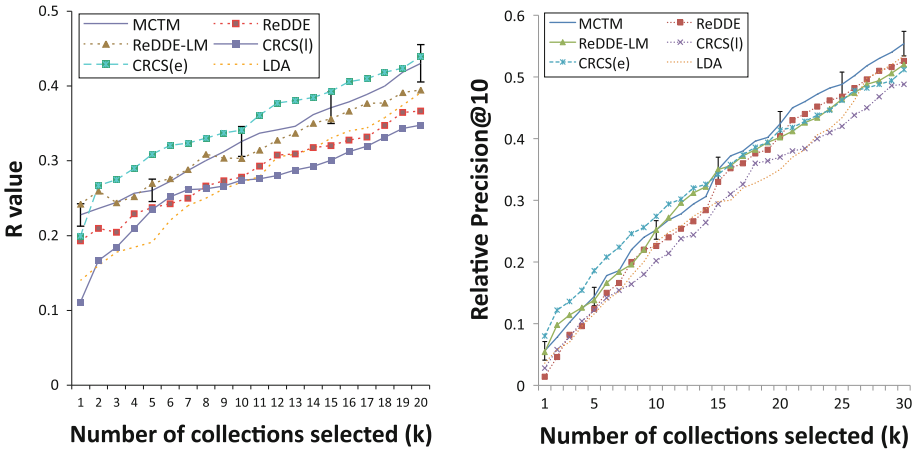

Fig. 8 Comparison over the TREC6-bysource testbeds using the R_k (*left*) and the relative P_k@10 (*right*) metrics

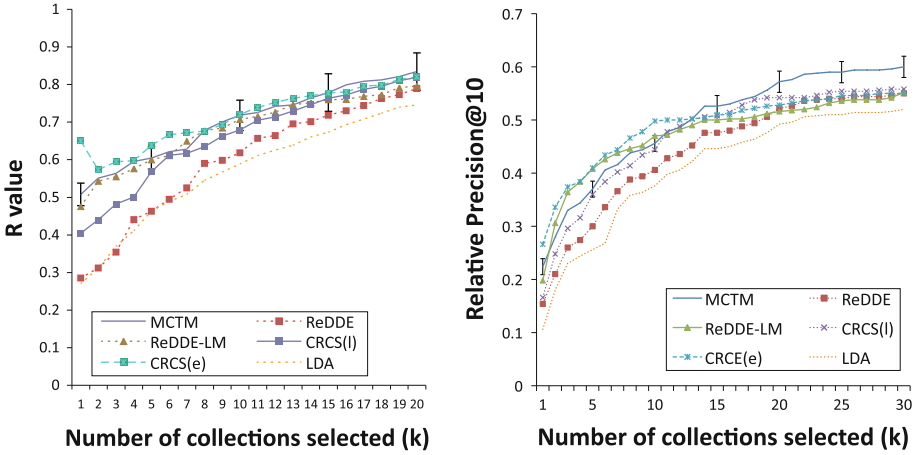

Fig. 9 Comparison over the Trec4-100col-global testbed

comparison of all methods. We also indicate variability in this average standardised score using a 95% confidence interval for the mean.

These results reiterate the findings from the analysis across the individual testbeds. Focusing on the R-value metric, the MCTM and CRCS(e) methods are consistently the better performing approaches. Although no significant different was observed between both methods, MCTM indicated less variable performance than CRCS(e) as illustrated by the smaller confidence interval around the average performance (see Fig. 11 (bottom)). Both ReDDE methods recorded more variable retrieval performance, while LDA adapted for collection selection was the worse performing method. Focusing on the relative P@10 metric, Figure 12, both MCTM and CRCS(e) again were the best performing methods along with the ReDDE-LM approach.

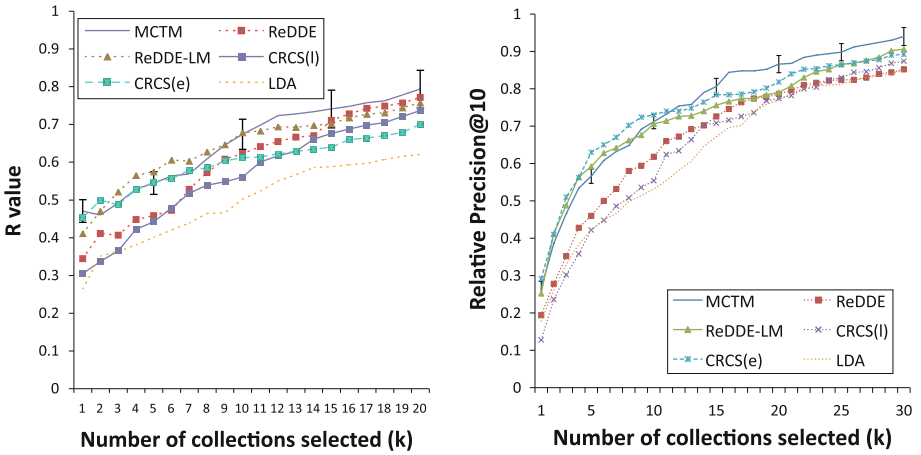

Fig. 10 Comparison over the Trec6-100col-global testbed

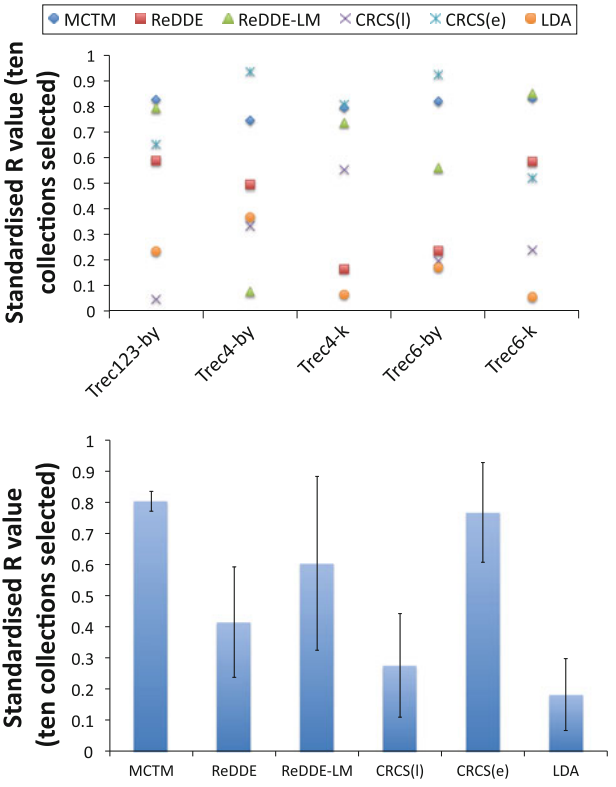

Fig. 11 Standardised *R* value at ten collections selected. The *top plot* reports the standardised scores for each technique and the *bottom plot* is the mean performance score across all collections. The 95% confidence interval for the mean is also presented

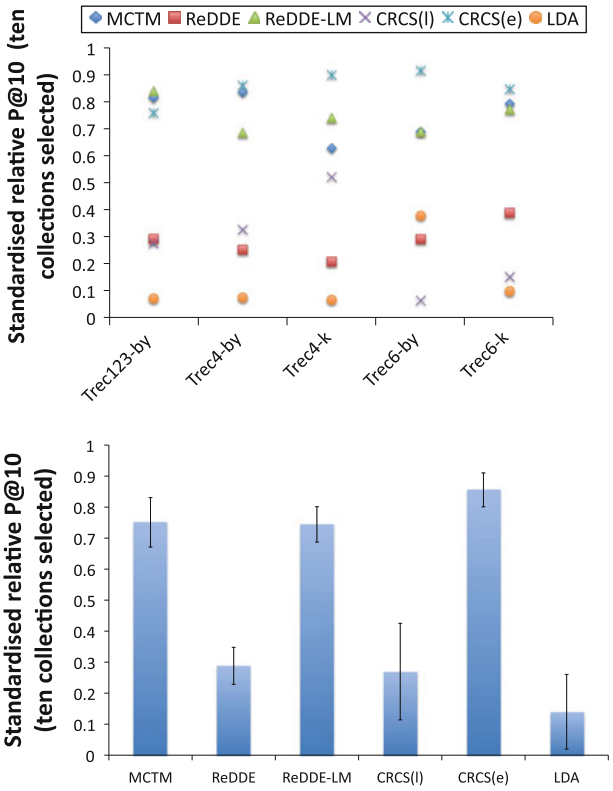

Fig. 12 Standardised relative P@10 value at ten collections selected. The *top plot* reports the standardised scores for each technique and the *bottom plot* is the mean performance score across all collections. The 95% confidence interval for the mean is also presented

6 Discussion

In general we observed that no single technique dominates performance across the different testbeds. This indicates that there is a certain amount of variance in the performance of state-of-the-art approaches depending on the test corpus.[7] Moreover, the fact that MCTM performed well (albeit not always the best) across the different collections provides strong indication of the quality of the technique. For example, although the approach did not provide statistically significant improvements in comparison to CRCS(e), the variability (or uncertainty) of performance was found to be smaller when using MCTM.

The results indicated that MCTM consistently outperforms the non-hierarchical LDA approach. Showing that the additional modelling complexity results in a better estimation for the likelihood of a query given the collection $P(q|c)$. The results also indicated a trend that MCTM improved at larger cut-off values. In other words, MCTM performed better as more servers were selected. This suggests that using information from topically related collections does improve the representation of collections. CRCS(e) and ReDDE-LM, by

[7] The variance in selection performance across different test collections has been observed previously by researchers in DIR including French et al. (1998).

design, perform well at lower cut-off values. This is because if a collection sample already contains one or more relevant documents for the query, these documents are more likely to be ranked high in the centralised sample index resulting in high precision. However, if a collection representation does not contain a relevant document then that collection is less likely to be ranked highly. The results would indicate that the MCTM model could be addressing this limitation by implicitly smoothing collection representations with information from other topically related collections.

Finally, it is possible that a combination of different techniques (in particular a mixture of topic modelling and the simple language modelling used in ReDDE-LM) might result in more robust performance across different testbeds. A fusion approach to collection selection may result in significant gains combining a mixture of different strategies.

7 Conclusions

In this paper we have introduced a topic modelling approach to collection selection based on a hierarchical latent Dirichlet allocation (MCTM) model with document groupings. We have shown with extensive experiments that the topic model performs comparably with state-of-the-art collection selection approaches and that it outperforms a non-hierarchical topic modelling approach, LDA, when applied to this problem.

We note that while the MCTM model contains a number of parameters (four smoothing parameters and a topics count), good or optimal values can be chosen for these parameters based on model fit, by maximising the likelihood of the sampled documents. Thus the parameter values are selected in a way that is independent of the collection selection problem itself. Since the model parameters can be chosen to best fit the data and not to maximise collection selection performance on training data, we can consider the MCTM based resource ranking algorithm to be an unsupervised learning technique. Thus our approach is somewhat different from that of central-rank-based collection selection (CRCS) and other collection selection algorithms, where the parameters of the algorithm need to be tuned using a set of test queries and relevance judgements for best performance on a new set of resources. This tuning requires considerable effort in generating representative sets of queries, and labelling relevant documents. Although estimating the MCTM is computationally intensive in comparison to other models, this is an off-line task and parameter estimation can be parallelized (Asuncion et al. 2008). It is important to note, however, that at retrieval time the model is comparable to existing collection selection models.

Topic modelling approaches such as LDA can deal to a certain extent with problems of synonymy and polysemy due to the fact that topics are defined and discovered by the co-occurrence of words across documents. Synonymy and polysemy are particularly important problems for federated search and collection selection in particular because of the small samples of documents that are often used to represent large collections. Thus LDA and MCTM may offer a principled way of dealing with this problem. One of the biggest problems with query-based sampling (QBS) of uncooperative collections is missing vocabulary in the sample. The MCTM model may be capable of dealing with this problem by "inferring" the presence in each sample of additional vocabulary terms from high density topics, since the term distributions for those topics are estimated across the samples from the different collections.

The usefulness of the MCTM model is not limited to collection selection. Once a model has been learnt using samples of each collection, the model can be used for a number of different purposes, including predicting the source of a document, assigning new

documents to collections, determining the similarity between collections (based on their topic representation), visualizing the contents of collections (e.g. using the topical equivalent of a "tag cloud"), and facilitating navigation through different collections.

Future work includes investigating more complicated collection models which allow for correlation between topics within individual documents, such as correlated topic models (Blei and Lafferty 2007) or Pachinko allocation (Li and McCallum 2006), to see if collection "aware" versions of these models can be developed and adapted to the collection selection problem. A second interesting direction would be to investigate hyperparameter estimation (Wallach 2008) as well as non parametric topic modelling techniques based on Dirichlet processes (Teh et al. 2006), where the optimal number of topics for the model is discovered during model estimation.

References

Asuncion, A., Smyth, P., & Welling, M. (2008). Asynchronous distributed learning of topic models. In *Neural information processing systems (NIPS'08)* (pp. 81–88). Cambridge: MIT Press.

Avrahami, T. T, Yau, L., Si, L., & Callan, J. (2006). The fedlemur project: Federated search in the real world. *Journal of the American Society for Information Science and Technology 57*(3), 347–358.

Balog, K. (2008). The SIGIR 2008 workshop on future challenges in expertise retrieval (fCHER). *SIGIR Forum 42*(2), 46–52.

Bar-Yossef, Z., & Gurevich, M. (2006). Random sampling from a search engine's index. In *WWW'06: Proceedings of the 15th international conference on world wide web* (pp. 367–376). New York: ACM.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics 1*, 17.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Buntine, W. L. (1994) Operations for learning with graphical models. *Journal of Artificial Intelligence Research 2*, 159–225.

Callan, J. P. (2000). Advances in information retrieval. In *Distributed information retrieval* (pp. 127–150). Dordrecht: Kluwer Academic Publishers.

Callan, J. P., & Connell, M. (2001). Query-based sampling of text databases. *ACM Transactions of Information Systems 19*(2), 97–130.

Callan, J. P., Lu, Z., & Croft, W. B. (1995). Searching distributed collections with inference networks. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 21–28). New York: ACM Press.

Craswell, N., Crimmins, F., Hawking, D., & Moffat, A. (2004). Performance and cost tradeoffs in web search. In *ADC'04: Proceedings of the 15th Australasian database conference* (pp. 161–169).

Elsas, J. L., Arguello, J., Callan, J., & Carbonell, J. G. (2008). Retrieval and feedback models for blog feed search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 347–354). New York: ACM.

French, J. C., Powell, A. L., Viles, C. L., Emmitt, T., & Prey, K. J. (1998). Evaluating database selection techniques: A testbed and experiment. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 121–129). New York: ACM.

Gravano, L., Chang, C. C. K., Garcia-Molina, H., & Paepcke, A. (1997). Starts: Stanford proposal for internet meta-searching. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on management of data* (pp. 207–218). New York: ACM Press.

Gravano, L., García-Molina, H., & Tomasic, A. (1999). GlOSS: Text-source discovery over the Internet. *ACM Transactions on Database Systems 24*(2), 229–264.

Gravano, L., Ipeirotis, P. G., & Sahami, M. (2003) Qprober: A system for automatic classification of hidden-web databases. *ACM Transactions of Information Systems 21*(1), 1–41.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science 101*, 5228–5235.

Hawking, D., & Thomas, P. (2005). Server selection methods in hybrid portal search. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 75–82). NY: ACM Press.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57). NY: ACM.

Ipeirotis, P. G., & Gravano, L. (2008). Classification-aware hidden-web text database selection. *ACM Transactions of Information Systems 26*(2), 1–66.

Ipeirotis, P. G., Agichtein, E., Jain, P., & Gravano, L. (2006). To search or to crawl?: Towards a query optimizer for text-centric tasks. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on management of data* (pp. 265–276). New York: ACM Press.

Li, W., & McCallum, A. (2006), Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML '06: Proceedings of the 23rd international conference on machine learning* (pp. 577–584). New York: ACM.

Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., & Halevy, A. (2008). Google's deep web crawl. *Proceedings of the VLDB Endowment 1*(2), 1241–1252.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Paepcke, A., Brandriff, R., Janee, G., Larson, R., Ludaescher, B., Melnik, S., et al. (2000). Search middleware and the simple digital library interoperability protocol. *D-Lib Magazine 6*(3).

Price, G., & Sherman, C. (2001). *The invisible web: Uncovering information sources search engines can't see*. Medford: CyberAge Books.

Puppin, D., Silvestri, F., Perego, R., & Baeza-Yates, R. (2010). Tuning the capacity of search engines: Load-driven routing and incremental caching to reduce and balance the load. *ACM Transactions on Information Systems (TOIS) 28*(2), 1–36.

Shokouhi, M. (2007). Central-rank-based collection selection in uncooperative distributed information retrieval. In *Advances in information retrieval, 29th European conference on IR research*. ECIR 2007, Rome, Italy, 2–5 April 2007, Proceedings (pp. 160–172).

Shokouhi, M., Baillie, M., & Azzopardi, L. (2007). Updating collection representations for federated search. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 511–518). New York: ACM.

Si, L., & Callan, J. (2003). Relevant document distribution estimation method for resource selection. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 298–305). New York: ACM.

Si, L., Jin, R., Callan, J., & Ogilvie, P. (2002). A language modeling framework for resource selection and results merging. In *CIKM '02: Proceedings of the eleventh international conference on information and knowledge management* (pp. 391–397). New York: ACM.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Thomas, P., & Hawking, D. (2009), Server selection methods in personal metasearch: A comparative empirical study. *Information Retrieval 12*(5), 581–604.

Wallach, H. M, (2008). Structured topic models for language. PhD thesis, Cambridge: University of Cambridge.

Webber, W., Moffat, A., & Zobel, J. (2008). Score standardization for inter-collection comparison of retrieval systems. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 51–58). New York: ACM.

Wei, X., & Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 178–185). New York: ACM.

Xu, J., & Croft, W. B. (1999). Cluster-based language models for distributed retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 254–261). New York: ACM Press.