Topic Based Language Models for ad hoc Information Retrieval

L. Azzopardi
School of ICT
University of Paisley
Paisley UK
E-mail: leif.azzopardi@paisley.ac.uk

M. Girolami
Computing Science
University of Glasgow
Glasgow UK
E-mail: girolami@dcs.gla.ac.uk

C. J. van Rijsbergen
Computing Science
University of Glasgow
Glasgow UK
E-mail: keith@dcs.gla.ac.uk

Abstract— **We propose a topic based approach to language modelling for ad-hoc Information Retrieval (IR). Many smoothed estimators used for the multinomial query model in IR rely upon the estimated background collection probabilities. In this paper, we propose a topic based language modelling approach, that uses a more informative prior based on the topical content of a document. In our experiments, the proposed model provides comparable IR performance to the standard models, but when combined in a two stage language model, it outperforms all other estimated models.**

I. Introduction

Language modeling for Information Retrieval has been a promising area of research over the last five years. The approach represents a fundamental shift in paradigm for probabilistic IR. Language models compute the relevance R of a document d with respect to a query q, by computing the likelihood of generating q from d, i.e. $p(q|d)$ [9]. The approach avoids attempting to explicitly estimate relevance, and instead exploits the outcome of the assumption that the relevance of the document is highly correlated with $p(q|d)$. This has been the source of much conjecture concerning the theoretical underpinnings of the model [11]. Nonetheless, language models for IR have attracted much attention as they provide an elegant mathematical model for ad-hoc text retrieval with excellent empirical results reported in the literature.

An important problem in Language Modelling is the estimation of the multinomial term distribution for each document (document model θ_d). According to Ponte and Croft [9], this is essential in achieving optimal IR performance. To date most researchers have created document models using the estimated collection probabilities as a form of *a priori* knowledge [9], [8], [10], [12] . We posit, that if there is an underlying topical structure within the corpus, then utilizing this prior knowledge enables the construction of a more accurate representation of the document models. In this paper, we propose a topic based language model, that employs a document dependent term prior.

The structure of this paper is as follows; In the next section, we introduce the Language Modelling approach for ad hoc text retrieval, and briefly review previous estimators for document modelling. In Section II-A, we formulate the general approach to topic based language modelling. Section III then introduces Latent Dirichlet Allocation for the estimation of the document dependent term priors. Finally, we provide an extensive empirical evaluation and discussion of the proposed model and its implementation.

II. Language Models

The mechanics of the Language Modelling approach for ad hoc text retrieval can be described as follows [9], [13]: Given a query q, that is represented as an empirical term uni-gram distribution over the vocabulary T, a document model θ_d is instantiated for each document d, and is represented by a normalized term uni-gram distribution over the vocabulary T i.e. the probability of a term given a document model, $p(t|\theta_d)$. We assume that the query terms are independently generated from the document via the document model. Documents are then ranked according to the probability of the query being generated from the document model. This is known as the standard unigram language model, or query likelihood approach, denoting $n(t,q)$ as the number of times the term t occurs in query q then:

$$p(q|\theta_d) = \prod_{t \in q} p(t|\theta_d)^{n(t,q)} \tag{1}$$

The quintessential problem is the estimation of the document language model. Not only must the estimated document model overcome the Zero Probability Problem[9], but it must also generate an accurate representation of the underlying data. This is an important problem and many different approaches have been proposed: Ponte and Croft [9] used a shrinkage style estimator that used the average probability of a term in documents containing it. Miller et. al. [8] used a two state hidden Markov model where one state is for the document and the other state is for the background collection, where terms are assumed to be drawn from. This model can be viewed as Jelinek-Mercer Smoothing (a two part mixture model) that backs off to the collection probabilities. Song and Croft [10] evaluated generating document models using higher order n-grams, to capture term dependencies. The two stage smoothing model was proposed by Zhai and Laffery [14] to account for differences in document and query generation. A recent attempt by Zaragoza *et. al.*[12] proposed to generate a more accurate representation by developing a full Bayesian approach

0-7803-8359-1/04/$20.00 ©2004 IEEE

to query scoring. All these methods relied upon obtaining a smoothed estimate of the the document models based on the background collection term frequencies. Whilst not an entirely naive source of prior knowledge, a more informative prior could be derived for documents. Specifically, by capitalizing on any inherent underlying topical structure within the corpus. This work is, therefore, closely related to the pioneering work of Hofmann [5], [6], who proposed Probabilistic Latent Semantic Indexing (PLSI). PLSI relied heavily upon the intuition that topical structure could benefit retrieval performance. The indexing of the model linearly combined the empirical probability estimates with the PLSI estimates weighted by Inverse Document Frequency and ranking according to the cosine function. At this time Language Modelling for ad-hoc retrieval had only been recently proposed. The remainder of this paper details a topic based language model which combines the intuition developed in [5], [6], within the Language Modelling framework and provides an extensive evaluation of its utility.

A. Standard Language Model

Before describing the Topic Based Language Model, we briefly detail the standard models, which we have extended to accommodate a document dependent term prior.

The two most widely used estimators for generating document models are: Jelinek-Mercer and Bayes Smoothing. The Jelinek-Mercer estimator, as previously mentioned, linearly interpolates the maximum likelihood estimate of a term in a document $p_{ml}(t|d)$ with the probability of a term given the collection $p_c(t)$, using the λ parameter to combine the probabilities, where $0 \leq \lambda \leq 1$. In Equation 2, $p_{ml}(t|d) = \frac{n(t,d)}{\sum_t n(t,d)}$, where $n(t,d)$ is the number of times term t occurs in document d, and $p_c(t) = \frac{\sum_d n(t,d)}{\sum_{t,d} n(t,d)}$.

$$p(t|\theta_d) = \lambda p_{ml}(t|d) + (1-\lambda)p_c(t) \qquad (2)$$

The Bayes Smoothing method generates the document model by smoothing proportionally to the length of the document, by varying the β parameter in Equation 3.

$$p(t|\theta_d) = \frac{n(t,d) + \beta p_c(t)}{\sum_t n(t,d) + \beta} \qquad (3)$$

Both models perform well empirically, however, it is unclear as to how a better representation of the document is actually generated. For instance, adding the $p_c(t)$ to all documents, does not make any one particular document any more probable than any other, which does not originally contain the query term. If the document is relevant, but does not contain the query term, it is still no more probable, even though it maybe topically related.

B. Topic Based Language Models

The premise for our proposal of using topic based prior knowledge stems from the Cluster Hypothesis [7]. The Cluster Hypothesis states that: *similar documents tend to be relevant to the same request.* Thus, before a request (query) is submitted to the IR system, we can group similar documents into topics, where a document can be about one or even many topics. This topical information, in the form of a term distribution over topics, associated with the distribution of topics over a document can be used to estimate a document dependent term prior $p_d(t)$, with which to smooth the document model. It is document dependent because the document may be drawn from a topic or a number of them, and this distribution determines the term prior. This should provide a more informative prior as it relies on the term distribution over the topics as oppose to the entire collection's. Obviously, smoothing the document model according to its topical structure requires the Cluster Hypothesis to hold. The Topic Based Language Modelling approach, extends the aforementioned language models by substituting $p_c(t)$ with $p_d(t)$ in Equations 2 and 3.

If we assume that a document can only be drawn from one topic, then to implement the topic based model, we could employ a naive Bayes Mixture model to estimate the $p_d(t)$. Every document assigned to a particular class would be generated from the same term distribution and this would be used to smooth the document model. When the number of classes is one, then we return to the original collection based term prior, $p_c(t)$. Further, if the Jelinek Mercer smoothing is used with the document dependent term prior generated from such a model and $\lambda = 1$, then the ranking of documents would be equivalent to the probability of query being generated from the class that the document has been assigned. And this would result in cluster based retrieval.

However, a document maybe composed of a number of topics and so the Aspect Model (known as PLSA)[5], [6] was proposed to address this. The aspect model can be described as follows: for document d, we select a topic z with probability $p(z|d)$. Then we generate term t from topic z with probability $p(t|z)$. The document dependent term prior can therefore be estimated as the combination of the term topic distribution and topic document distribution:

$$p_d(t) \equiv \sum_z p(t|z)p(z|d) \qquad (4)$$

Further, we propose to imbed the document dependant term prior within the Two Stage Language Model[14]. This model was motivated from the empirically shown need to represent both the document and query. Therefore, we posit that we can use the document dependant term prior to model the documents and benefit from the collection probabilities to model the query. The method is the combination of the Bayes Smoothing method (which yields a *maximum a posteriori* MAP estimator) and the Jelinek-Mercer smoothing method (see Equation 5).

$$\begin{aligned} p(t|\theta_d) &= \lambda\Big\{p_{MAP}(t|\theta_d)\Big\} + (1-\lambda)p_c(t) & (5) \\ &= \lambda\Big\{\frac{n(t,d) + \beta p_d(t)}{\sum_t n(t,d) + \beta}\Big\} + (1-\lambda)p_c(t) & (6) \end{aligned}$$

Under this approach, indexing can be viewed as a principled alternative to the indexing performed in [5], [6], which is now consistent with the Langauge Modelling Paradigm.

III. Latent Dirichlet Allocation

Aspect style models, namely, Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA), have been applied to various IR related tasks [2], [3], [6], [1]. They have been noted for their ability to generate a contextually smooth document model [6], [1]. However, these models have up till now not been applied to ad-hoc retrieval within a wholly consistent language modelling framework. For the purposes of estimating the document dependent term prior, we use the latest estimator LDA. It has been shown to provide state of the art performance in terms of predictive likelihood over PLSA[1] and the naive Bayes Mixture model[1]. The estimation of a term t given document d and k latent variables under LDA $p^k_{lda}(t|d)$, is approximated, as exact inference is not possible[1], [4]:

$$p^k_{lda}(t|d) = \sum_k p(t|k) \frac{\gamma^k_d}{\sum_{k'} \gamma^{k'}_d} \tag{7}$$

LDA represents documents as a mixture over latent variables, where each latent variable is characterized by a distribution of terms $p(t|k)$ and each document is described as a variational distribution over topics $\frac{\gamma^k_d}{\sum_{k'} \gamma^{k'}_d}$.

The detailed derivation of the inference and variational parameter estimation algorithm can be found in [1].We refer to the latent variable as a topic, but this is potentially misleading, as these "topics" must be inferred as noted in [1]. The term topic distributions generated from the model are generally quite intuitive, though there are instances when a decomposition is far from intuitive. To demonstrate the utility of the LDA model we now provide an example.

A. Example

For the purposes of illustration, we have constructed a small collection consisting of 100 documents taken from three different collections which cover distinctive topics; Medicine (MED), Aeronautics (CRAN) and Information Science (CISI). Table I, reports the most probable terms for each topic, and the most probable terms given all topics. Terms are shown as the word stems after Porter stemming.

We applied the LDA model with $k = 3$ and ran twenty-five different randomly initialized parameter estimation routines. For each model generated under LDA, we examined the term topic distributions and manually assigned the distribution to the one that matched our expectations. Our expectations were based on the empirical probabilities as this is our natural understanding of the collection, and we refer to this as the intuitiveness of the inferred topic. We provide two examples; one intuitive, the other not so. Firstly, in Table II, the term distributions for each topic is very intuitive based on the empirical estimates. The corresponding aggregated topic-document distributions are provided in Figure 1. As we

MED	CRAN	CISI	ALL
patient	flow	librari	librari
cell	pressur	book	flow
studi	number	servic	number
blood	base	system	pressur
chang	effect	inform	effect
case	boundari	medic	result
rat	bodi	us	bodi
acid	veloc	catalog	case
liver	layer	need	base
group	solut	user	studi
diseas	heat	librarian	measur
present	equat	academ	present
tissu	theori	analysi	method

TABLE I

Terms occurring with the highest probability given the collection

1	2	3
blood	flow	library
patient	pressur	book
studi	number	system
chang	base	servic
cell	boundari	us
acid	effect	inform
rat	bodi	medic
increas	veloc	develop
diseas	theori	studid
effect	layer	catalog
group	solut	need
2	heat	analysi
1	equat	user
5	jet	present

TABLE II

Terms occurring with the highest probability given the most intuitive or inferrable decomposition .

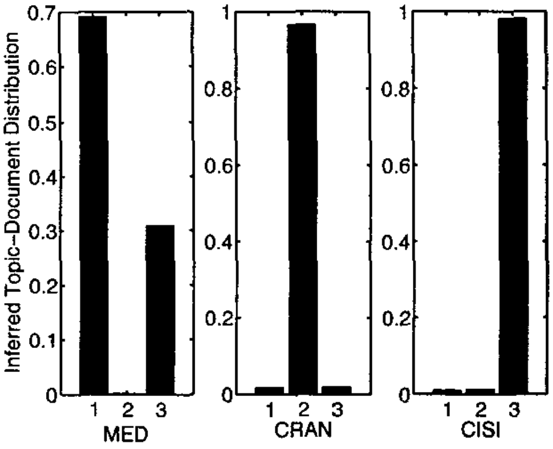

Fig. 1. Mean of the topic-document distributions for each collection given the most intuitive decomposition.

[1]Actually, PLSA is in fact a *maximum a posteriori* estimate of LDA[4].

would expect the documents from CRAN consist mainly of the second topic, CISI consists mainly from the third topic. Interestingly, MED consists of the first topic and the third topic, showing that terms in the third topic are in both MED and CISI documents. This provides an excellent example of the representational ability that can be obtained under the LDA model. On the other hand, our second example is not so intuitive, see Table III for the term distribution over topics and the corresponding topic-document distributions in Figure 2. The term distributions appear to have associated terms that are used in different contexts all together. And this intermingling of terms, is reflected in the mixture of topics required to generate a document. The decompositions provided may not necessarily reflect our intuitions, because LDA attempts to obtain the highest predictive likelihood. To determine whether there was a relationship between the predictive likelihood under the LDA model and our intuitive assessment of the topic decompositions. We measured the difference between the empirical probabilities and the inferred topic distribution with the L1 Norm measure. Figure 3, shows that there is a statistically strong correlation between the two measures (Pearson correlation test; $r = -0.857$, $p < 0.01$).

This examples shows that whilst there is variance in the intuitiveness in decompositions, that it is possible to obtain decompositions that are reflective of the inherent topical structure within the corpus. We hope to capitalize on the ability of LDA to induce such topical structure and translate this into improved IR performance. The following section provides a comprehension evaluation of the topic based language model.

1	2	3
librari	flow	librari
book	pressure	servic
studi	number	method
cell	effect	medic
catalog	base	inform
us	bodi	research
case	boundari	develop
patient	layer	system
system	result	heat
select	jet	region
circul	theori	veloc
subject	shock	program
chang	1	us

TABLE III

TERMS OCCURRING WITH THE HIGHEST PROBABILITY GIVEN THE LEAST INTUITIVE DECOMPOSITION.

IV. EVALUATION

The empirical evaluation compared the two standard language models against the proposed extended topic based language models. For convenience, we use the following abbreviations: JM as Jelinek-Mercer Smoothing with $p_c(t)$ as a prior, BS as Bayes Smoothing with $p_c(t)$ as a prior, LDA-JM as Topic based Jelinek-Mercer Smoothing and LDA-BS as Topic based Bayes Smoothing. Further, we compared the two stage smoothing model with

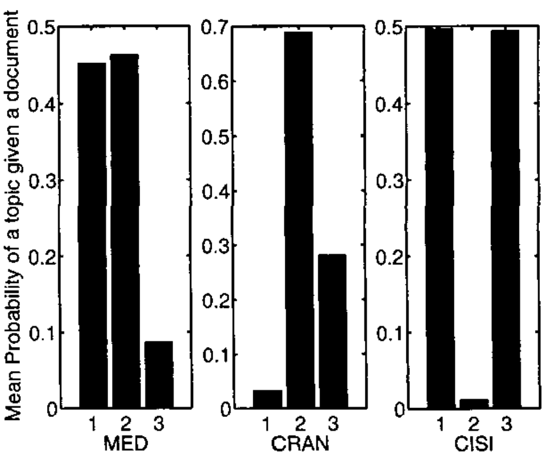

Fig. 2. Mean of the topic-document distributions for each collection given the least intuitive decomposition.

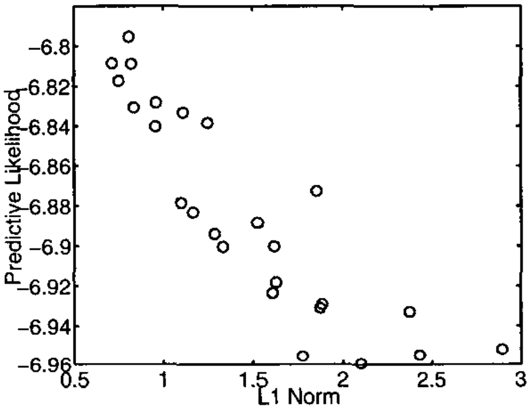

Fig. 3. Relationship of predictive likelihood to the L1 Norm as a measure of the intuitiveness of the decomposition.

the document dependent term prior(LDA-BS-JM), and the original variant[14] (BS-JM), where in Equation 5 $p_d(t)$ is substituted for $p_c(t)$. The parameter space examined for each variable was; $\lambda \rightarrow [0.05, 0.1, 0.2, 0.3, 0.5, 0.6]$, $\beta \rightarrow [1, 10, 100, 250, 500, 750, 1000]$ and $k \rightarrow [2, 4, 16, 32, 64, 128, 256]$. Due to space constraints results from only a subset of these parameters have been reported herein.

These models were then applied to four standard IR test collections; MED (1033 medical abstracts with 30 queries), CRAN (1400 Cranfield Aeronautical abstracts and 155 queries), CACM (3204 articles from the journal Communications of the ACM and 50 queries), and CISI (35 queries and 1460 documents extracted from the information science literature). These collections, whilst small in comparison to the TREC collections, were chosen because of the computational expense associated with the estimation of the model parameters for LDA. The data preparation was standard; terms were stemmed using the Porter Stemming Algorithm, standard stop words were removed and so too were infrequent terms. The collections were then partitioned by random sub-sampling to produce 10 test and train sets where 10 percent of the data was assigned to the test set. The test set was used to calculate the predictive likelihood, and the train set was used to build the language model and perform the indexing. Latent Dirichlet Allocation required a non-linear optimization, so for each test/train set we used five different initialization. This totaled 50 runs per k value. The model parameters were estimated as in [1], [4]. Note that LDA, has not been evaluated for ad-hoc retrieval within the language modelling framework, and this is the first extensive investigation of its utility.

V. Results

In Tables IV, V, VIII and IX, we report the mean Average Precision as a percentage (where the mean is taken over all sets and initializations, if applicable), for each of the methods. For the topic based models, we report all $k = 2$, the worse performing models given k, and best performing models given k. Also, the best result given a particular estimator is marked as bold, unless otherwise stated.

Model	k	λ:	0.1	0.2	0.3	0.5	0.7
JM	-		50.3	**50.7**	50.5	50.1	49.1
LDA-JM	2		44.0	46.2	47.1	47.8	**47.9**
LDA-JM	16		30.6	34.3	36.7	40.1	**42.7**
LDA-JM	256		50.0	**50.5**	50.3	49.9	49.1
Model	k	β:	1	250	500	750	1000
BS	-		43.7	50.0	**50.4**	49.9	49.7
LDA-BS	2		44.2	**45.8**	44.9	23.8	42.9
LDA-BS	4		**44.4**	40.4	38.0	36.2	34.8
LDA-BS	256		43.6	49.9	**50.1**	49.8	49.4

TABLE IV

MED - mean Average Precision results for the different smoothed estimators.

Model	k	λ:	0.1	0.2	0.3	0.5	0.7
JM	-		28.6	**29.0**	28.9	28.6	28.0
LDA-JM	2		27.0	28.0	**28.4**	28.4	28.0
LDA-JM	64		21.1	22.6	23.5	24.7	**25.3**
LDA-JM	128		28.6	**29.0**	28.9	28.6	28.0
Model	k	β:	1	250	500	750	1000
BS	-		25.9	28.9	**29.0**	28.6	28.2
LDA-BS	2		26.3	**28.3**	28.2	27.6	26.9
LDA-BS	32		**25.9**	22.8	21.3	20.3	19.6
LDA-BS	256		25.9	28.9	**29.0**	28.6	28.2

TABLE V

CRAN - mean Average Precision results for the different smoothed estimators.

Collection	Model	k	β	λ	mAP
MED	BS-JM	-	10	0.2	50.8
	LDA-BS-JM	64	100	0.1	**55.1**
CRAN	BS-JM	-	250	0.7	29.2
	LDA-BS-JM	64	100	0.3	**30.2**
CISI	BS-JM	-	500	0.7	23.8
	LDA-BS-JM	32	100	0.1	**24.4**
CACM	BS-JM	-	500	0.7	37.0
	LDA-BS-JM	2	250	0.3	**37.1**

TABLE VI

Two Stage Smoothing (standard and topic based): The best performing two stage smoothing models. Bold indicates the better model in terms of mean Average Precision.

Model	MED	CRAN	CACM	CISI
JM	50.7	29.0	34.4	23.3
LDA-JM	50.6	29.0	34.5	23.5
BS	50.4	29.0	36.9	23.1
LDA-BS	50.2	31.0*	36.7	23.0
BS-JM	50.8	29.2	37.0	23.8
LDA-BS-JM	**56.7**$\star$	**31.0**$\star$	**37.7**$\ast$	**25.0**$\bullet$

TABLE VII

Comparison of the best performers: The mean Average Precision for each model and collection - The best overall results are in bold. The $\star$,$\ast$ and $\bullet$ denote the model is significantly better statistically, where $\star$ is all other models, $\ast$ is JM and $\bullet$ is BS.

In Table VII, the mean Average Precision is reported for the best performing models. For the LDA models, from the models that gave the best mean Average Precision, we selected the best initialization from each set given the parameter set. Similarly, the best standard models were selected, given the parameter set. Comparisons between models was performed with a query-wise comparison across all sets, using the Wilcoxon Rank Sum test ($\alpha = 0.05$).

VI. Discussion

The general trend for both variants of topic based smoothing, LDA-JM and LDA-BS, that emerged from this set of results showed that when $k = 2$, the IR performance dropped substantially below the standard models, respectively, and continued to fall as the number of k increased. It would seem that a poor representation was being produced that degraded performance, at these lower values of k. Subsequent increases in the number of k brought about improved IR performance, and presumably a better document representation. This improvement, continued as the k increased until marginal improvements were gained past 128 topics. Interestingly, this suggested that a similar approximation is being generated to the standard models. And the difference at this point was not statistically significant, between the best standard model and topic based model, with respect to the type of variant. This may be a result of the document dependent term prior effectively equating to the $p_c(t)$. Or that distinctively different distributions with which to smooth documents result in similar mean IR performance.

Model	k	λ:	0.1	0.2	0.3	0.5	0.7
JM	-		33.1	34.3	**34.4**	33.6	32.3
LDA-JM	2		29.2	31.5	32.4	**32.8**	31.9
LDA-JM	32		14.9	18.0	20.1	23.4	**25.8**
LDA-JM	128		33.3	34.3	**34.4**	33.6	32.4

Model	k	β:	1	250	500	750	1000
BS	-		12.4	35.2	**36.9**	**36.9**	36.3
LDA-BS	2		12.6	**32.0**	31.6	30.7	29.9
LDA-BS	32		12.3	**15.7**	14.3	13.4	12.7
LDA-BS	128		12.5	35.1	36.3	**36.6**	36.1

TABLE VIII

CACM - MEAN AVERAGE PRECISION RESULTS FOR THE DIFFERENT SMOOTHED ESTIMATORS.

Model	k	λ:	0.1	0.2	0.3	0.5	0.7
JM	-		**23.5**	23.3	22.9	22.2	21.3
LDA-JM	2		20.6	21.7	**22.1**	22.0	21.6
LDA-JM	32		14.0	15.2	16.1	17.2	**18.0**
LDA-JM	128		**23.5**	23.4	23.0	22.2	21.4

Model	k	β:	1	250	500	750	1000
BS	-		16.7	22.4	23.1	**23.7**	23.7
LDA-BS	2		17.4	**20.6**	20.4	19.9	19.5
LDA-BS	16		**17.6**	14.9	13.9	13.3	12.8
LDA-BS	128		16.6	22.0	23.0	**23.7**	23.6

TABLE IX

CISI - MEAN AVERAGE PRECISION RESULTS FOR THE DIFFERENT SMOOTHED ESTIMATORS.

Zhai and Laffery[13] showed that the contribution of a term to the log likelihood of a query term being generated from a document is the equivalent to the $\frac{p(t|\theta_d)}{\alpha_d p_c(t)}$, where α_d is a document specific prior. They comment that the weighting is like the popular term frequency inverse document frequency weighting. Thus the importance of a term in a document is proportional to $p_c(t)$. As we have substituted the $p_c(t)$ for $p_d(t)$, this affects the importance assigned to terms on a document specific basis. For instance, given two documents which are drawn entirely from two distinct topic distributions. If the probability of the query term is under represented in one and over represented in the other, yet the term is common to the entire collection, with characteristics of a stop word, then the importance of the term for the first distribution will be higher than the latter. This means that a significant bias may be introduced through the topic based approach. This is somewhat remedied in the two stage smoothing approach, where the $p_d(t)$ is used to estimate a better representation of the document model, and the $p_c(t)$ to better characterize the query role. This is evident when we compared the estimation of parameters with the original two stage model, where the topic based model, reports improved IR performance.

Finally, if we consider that the topic based approach is in fact a novel implementation of the Cluster Hypothesis within the Language Modelling framework, then we have provided empirical evidence that shows using the inherent topical structure can achieve improved IR performance.

VII. Conclusions

This paper explored the possibility of using a document specific term prior based on inferred topics induced from the corpus. The results show that on average the method was comparable to the standard language modelling techniques. However, when linearly combined with the background probabilities, in the two stage topic based Language Model, the IR performance was consistently superior to the standard models and standard two stage smoothing model across all collections.

Due to the computational expense of the LDA method we restricted this study to relatively small test collections. Further work is required to ascertain whether these results are consistent across larger and more varied test collections. Also, it is worth considering whether there is a significant difference between the variational approximation (LDA) or the *maximum a posteriori* estimate (PLSI/PLSA) for the aspect model.

Acknowledgment

The first author is supported by Memex Technology Ltd (www.memex.com).

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proceedings of the 17th International Conference on Machine Learning*, pages 167–174. Morgan Kaufmann, 2000.

[3] D. Cohn and T. Hofmann. The missing link: A probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems (NIPS*13)*. MIT Press, 2001.

[4] M. Girolami and A. Kaban. On an equivalence between plsi and lda. In *26th Annual ACM Conference on Research and Development in Information Retrieval, SIGIR*, pages 433–434, Toronto, Canada, 2003.

[5] T. Hofmann. Probabilistc latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*. ACM Press, 1999.

[6] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[7] N. Jardine and C. J. Van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[8] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval. In *22nd Annual International ACM SIGIR conference on Research and development in information retrieval*, pages 214–221, California, US, 1999. ACM Press.

[9] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the Twenty First ACM-SIGIR*, pages 275–281, Melbourne, Australia, 1998. ACM Press.

[10] F. Song and W. B. Croft. A general language model for information retrieval. In *SIGIR ACM Research and Development in Information Retrieval*, pages 279–280, Berkeley, CA., 1999.

[11] K. Sparck-Jones, S. E. Robertson, D. Hiemstra, and H. Zaragoza. Language modeling and relevance. In W. B. Croft and J. Lafferty, editors, *Language Modeling for Information Retrieval*, pages 57–71. Kluwer Academic Publishers, 2003.

[12] H. Zaragoza, D. Hiemstra, M. Tipping, and S. Robertson. Bayesian extension to the language model for ad hoc information retrieval. In *Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–9, Toronto, Canada, July 2003.

[13] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *2001 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 49–56, New Orleans, LO, 2001. ACM Press.

[14] C. Zhai and J. Lafferty. Two-stage language models for infromation retrieval. In *2002 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 49–56, Tampere, Finland, 2002.