On an Equivalence between PLSI and LDA

Mark Girolami
School of ICT
University of Paisley
PA1 2BE, UK
mark.girolami@paisley.ac.uk

Ata Kabán
School of Computer Science
University of Birmingham
B15 2TT, UK
a.kaban@cs.bham.ac.uk

ABSTRACT

Latent Dirichlet Allocation (LDA) is a fully generative approach to language modelling which overcomes the inconsistent generative semantics of Probabilistic Latent Semantic Indexing (PLSI). This paper shows that PLSI is a *maximum a posteriori* estimated LDA model under a uniform Dirichlet prior, therefore the perceived shortcomings of PLSI can be resolved and elucidated within the LDA framework.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language Models*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Algorithms

Keywords

Language Model

1. INTRODUCTION

Language Modelling (LM), as a statistically principled approach to information retrieval (IR), employs the conditional probability of a query ($\mathbf{q}$) given a document ($\mathbf{d}$) $P(\mathbf{q}|\mathbf{d})$, as a means of relevance ranking [4]. One particular approach to LM based IR is PLSI [2]. PLSI decomposes the joint probability of observing a term w and document $\mathbf{d}$ with the use of a latent variable k such that $w \perp \mathbf{d} \,|\, k$ and $P(w, \mathbf{d}) = \sum_k P(w|k)P(k|\mathbf{d})$. PLSI has been shown to be a low perplexity language model and outperforms latent semantic indexing in terms of precision-recall on a number of small document collections [2]. However, the generative semantics of PLSI are not fully consistent which leads to problems in assigning probability to previously unobserved documents [1]. LDA [1] is also a probabilistic LM which possesses consistent generative semantics and overcomes some of the perceived shortcomings of PLSI. However, the following section will show that PLSI emerges directly as a specific instance of LDA so the claimed shortcomings of PLSI can be understood within the LDA framework.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

2. LDA AND PLSI EQUIVALENCE

A language model $\mathcal{M}$ based on a corpus $\mathcal{D}$ with vocabulary $\mathcal{V}$ is represented by LDA as follows. For corpus $\mathcal{D}$ a k-dimensional parameter $\boldsymbol{\alpha}$ is fixed. In generating document $\mathbf{d}$ a K-dimensional variable $\boldsymbol{\theta}$ is drawn from the Dirichlet distribution $D(\boldsymbol{\theta}|\boldsymbol{\alpha})$. The parameters $P(w|\theta_k)$ denoting the probability of the term w given the k'th element of the Dirichlet variable $\boldsymbol{\theta}$ are then linearly combined to obtain the multinomial distribution $P(w|\boldsymbol{\theta})$ from which a term w is drawn. Sampling from $P(w|\boldsymbol{\theta})$ is repeated for each term in the document. Denoting the $|\mathcal{V}| \times K$ parameters $P(w|\theta_k)$ as the matrix $\mathbf{P}$ and the number of times that term w appears in the document as $c_{\mathbf{d},w}$ then the probability assigned to the document $\mathbf{d}$ under the LDA model with parameters $\boldsymbol{\alpha}$ and $\mathbf{P}$ is given as

$$P(\mathbf{d}|\boldsymbol{\alpha}, \mathbf{P}) = \int_{\triangle} d\boldsymbol{\theta} D(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{w \in \mathbf{d}} \left\{ \sum_{k=1}^{K} P(w|\theta_k)\theta_k \right\}^{c_{\mathbf{d},w}}$$

where the integral is defined over the support of the Dirichlet distribution. Exact inference for LDA is not possible, so approximate variational methods have been developed in [1] for the purposes of inference and parameter estimation.

However, another approach to approximate inference and estimation for LDA models is the *maximum a posteriori* estimator which obtains the value of the variable $\boldsymbol{\theta}$ that maximizes the posterior distribution given the document $\mathbf{d}$ and obviates the necessity to obtain the value of the posterior, so in the case of LDA, for each document we require to solve

$$\boldsymbol{\theta}_{\mathbf{d}}^{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \; \log\{P(\boldsymbol{\theta}|\mathbf{d}, \mathbf{P}, \boldsymbol{\alpha})\}$$

Once the estimate $\boldsymbol{\theta}_{\mathbf{d}}^{MAP}$ for every document in $\mathcal{D}$ has been obtained the parameters $\mathbf{P}$ and $\boldsymbol{\alpha}$ can be estimated by maximum likelihood (ML) estimation. If the Dirichlet distribution defines a uniform density across the simplex i.e. $\boldsymbol{\alpha} = \mathbf{1}$, where $\mathbf{1}$ denotes a K-dimensional vector of ones, then the MAP estimator is identical to the ML estimator and so

$$\begin{aligned} \boldsymbol{\theta}_{\mathbf{d}}^{MAP} = \boldsymbol{\theta}_{\mathbf{d}}^{ML} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \; \log\{P(\mathbf{d}|\boldsymbol{\theta}, \mathbf{P})\} \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \; \sum_{w \in \mathbf{d}} c_{\mathbf{d},w} \log \left\{ \sum_{k=1}^{K} P(w|k)\theta_k \right\} \end{aligned}$$

Once $\boldsymbol{\theta}_{\mathbf{d}}^{ML}$ is obtained the ML estimate for $P(w|k)$ requires

$$\begin{aligned}\mathbf{P}^{ML} &= \underset{\mathbf{P}}{\operatorname{argmax}} \sum_{\mathbf{d}\in\mathcal{D}} \log\{P(\mathbf{d}|\boldsymbol{\theta}_{\mathbf{d}}^{ML}, \mathbf{P})\} \\ &= \underset{\mathbf{P}}{\operatorname{argmax}} \sum_{\mathbf{d}\in\mathcal{D}} \sum_{w\in\mathbf{d}} c_{\mathbf{d},w} \log\left\{\sum_{k=1}^{K} P(w|k)\theta_{\mathbf{d},k}^{ML}\right\}\end{aligned}$$

where $\theta_{\mathbf{d},k}^{ML}$ is the k'th element of the ML estimated Dirichlet variable for document $\mathbf{d}$. As the Dirichlet variables satisfy the constraints $\theta_k \geq 0, \forall\ k$ and $\sum_k \theta_k = 1$ these can be viewed as the $P(k|\mathbf{d})$ parameters in PLSI.

As such the interleaving of the two ML estimation procedures above will recover exactly the ML estimator for PLSI [2]. Therefore PLSI is a MAP / ML estimator of an LDA document model under a uniform Dirichlet prior. Viewing PLSI as MAP LDA under a uniform prior the heuristic *folding-in* of queries or new documents can in fact be seen to be the principled MAP / ML estimation of the Dirichlet variable for the query/document. Whilst LDA has been shown experimentally to provide a lower perplexity language model than PLSI this can now be seen to be as an outcome of the approximate estimation method employed, indeed in [3] Expectation Propagation is shown to be more accurate than the variational approach developed in [1].

3. IR WITH LDA AND PLSI

The relevance of a document to a given query under such a model can be measured as the likelihood that the query is generated given a particular document and the parameterized model [4]. Formally this can be posed as the posterior probability of the query given the document and the language model adopted.

$$P(\mathbf{q}|\mathbf{d}) = \prod_{q\in\mathbf{q}} P(q|\mathbf{d})^{c_{\mathbf{q},q}}$$

What is required is $P(q|\mathbf{d})$ which follows from the LDA representation as

$$\int_{\triangle} P(q|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{d})d\boldsymbol{\theta} = \int_{\triangle} \left\{\sum_{k=1}^{K} P(q|k)\theta_k\right\} P(\boldsymbol{\theta}|\mathbf{d})d\boldsymbol{\theta}$$

which can be seen to be dependent on the expectation over the posterior distribution of the Dirichlet random variable given the document i.e.

$$\sum_{k=1}^{K} P(q|k) \int_{\triangle} \theta_k P(\boldsymbol{\theta}|\mathbf{d})d\boldsymbol{\theta} = \sum_{k=1}^{K} P(q|k) E_{P(\boldsymbol{\theta}|\mathbf{d})}\left\{\theta_{k,\mathbf{d}}\right\}$$

The required expectation is problematic due to the posterior being intractable, however if it is assumed that the posterior is approximately symmetric with one dominant mode then $E_{P(\boldsymbol{\theta}|\mathbf{d})}\left\{\theta_k\right\} \approx \theta_{k\mathbf{d}}^{MAP}$. These MAP estimates for each document have already been approximated as part of the model parameter optimization process and so

$$\sum_{k=1}^{K} P(q|k) E_{P(\boldsymbol{\theta}|\mathbf{d})}\left\{\theta_{k\mathbf{d}}\right\} \approx \sum_{k=1}^{K} P(q|k)\theta_{k,\mathbf{d}}^{MAP}$$

Therefore the probability of generating query $\mathbf{q}$ from document $\mathbf{d}$ under the LDA language model can be approximated by

$$P(\mathbf{q}|\mathbf{d}) \approx \prod_{q\in\mathbf{q}} \left\{\sum_{k=1}^{K} P(q|k)\theta_{k,\mathbf{d}}^{MAP}\right\}^{c_{\mathbf{q},q}}$$

For the case where a uniform Dirichlet prior is imposed on the LDA model then as shown above we exactly recover PLSI and $\theta_{k,\mathbf{d}}^{MAP} = \theta_{k,\mathbf{d}}^{ML} \equiv P(k|\mathbf{d})$.

$$P(\mathbf{q}|\mathbf{d}) \approx \prod_{q\in\mathbf{q}} \left\{\sum_{k=1}^{K} P(q|k)P(k|\mathbf{d})\right\}^{c_{\mathbf{q},q}}$$

The log of the above probabilistic measure can be considered as a form of cross-entropy $\sum_q c_{\mathbf{q},q} \log P(q|\mathbf{d})$ or *entropic cosine similarity* measure somewhat reminiscent of the PLSI-U similarity measure employed to good effect in terms of IR performance in [2].

An alternative LDA based similarity measure is the *a posteriori* probability of the document given the query $P(\mathbf{d}|\mathbf{q}) = \prod_{w\in\mathbf{d}} P(w|\mathbf{q})^{c_{\mathbf{d},w}}$ where now

$$P(w|\mathbf{q}) = \sum_{k=1}^{K} P(w|k) E_{P(\boldsymbol{\theta}|\mathbf{q})}\left\{\theta_{k\mathbf{q}}\right\} \approx \sum_{k=1}^{K} P(w|k)\theta_{k,\mathbf{q}}^{MAP}$$

which leads to the following expression for the required conditional probability

$$P(\mathbf{d}|\mathbf{q}) \approx \prod_{w\in\mathbf{d}} \left\{\sum_{k=1}^{K} P(w|k)\theta_{k,\mathbf{q}}^{MAP}\right\}^{c_{\mathbf{d},w}}$$

The $\theta_{k,\mathbf{q}}^{MAP}$ for the query requires to be estimated using a MAP estimator and as before for a uniform Dirichlet prior the LDA model is exactly PLSI so $\theta_{k,\mathbf{q}}^{MAP} \equiv P(k|\mathbf{q})$. As above taking the log we obtain $\sum_w c_{\mathbf{d},w} \log P(w|\mathbf{q})$. The required estimation of the posterior expected value of the Dirchlet variable given the query can now be understood as the 'heuristic' method of query 'folding-in' as originally proposed in the PLSI model [2].

4. CONCLUSIONS

This paper has clarified the relationship between PLSI and LDA. PLSI in fact is a MAP / ML estimated LDA model under a uniform Dirichlet distribution and issues surrounding 'heuristic' folding-in and likelihood computation are now resolved due to the LDA interpretation of the PLSI parameters presented.

5. ACKNOWLEDGMENTS

Mark Girolami is part of the DETECTOR project funded by the DTI Management of Information (LINK) Programme and EPSRC grant GR/R55184.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[2] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[3] T. P. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, 2002.

[4] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR 98*, pages 275–281. SIGIR, 1998.