Incorporating Context within the Language Modeling Approach for *ad hoc* Information Retrieval

Leif Azzopardi

Using context to improve retrieval performance is a current challenge within the discipline and presents a major challenge to the research community. In this thesis, the context of documents formed via semantic associations is utilized within the Language Modeling (LM) approach for *ad hoc* Information Retrieval.

Whilst, the LM approach provides a natural and intuitive means of encoding such context, it also represents a change to the way probability theory is applied to the ranking of documents in *ad hoc* Information Retrieval[5, 6, 2, 4]. This requires several assumptions to be engaged for its application. A consequence of engaging these assumptions is a key implication that better retrieval performance can be obtained through developing better representations of the documents[6]. It is posited that the context associated with a document will enable the development of such representations - *context based document models*. This premise relies upon the explicit and implicit assumptions of the Language Modeling approach being valid, which have, up until now, not been fully tested or verified.

In the thesis, we (1) formalize the assumptions of the Language Modeling approach; (2) motivated by the implications of these assumptions we present our framework for estimating context based document models; (3) perform a comprehensive analysis of the main assumptions underlying the Language Modeling approach, not only to validate the approach, but to deepen our understanding of the retrieval model itself; and (4) empirically assess the performance of the context based document models against the standard document models on various test collections and contexts. Some of the semantic associations used to define the context included using Probabilistic Latent Semantic Analysis (PLSA)[3]/Latent Dirichlet Allocation (LDA)[1], web links between documents, and interactions with the collection from topic tracking.
LMs and surmises three core assumptions for LM, which are later empirically validated.

An electronic copy of this thesis is available from: http://cis.strath.ac.uk/ leif/

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. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, 2001.

[3] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference of Uncertainty in Artificial Intelligence, UAI'99*, pages 289–296, Stockholm, Sweden, 1999.

[4] 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.

[5] J. M. Ponte. *A Language Modeling Approach to Information Retrieval*. PhD thesis, University of Massachusetts Amherst, 1998.

[6] 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.