Latent Topic Feedback for Information Retrieval

David Andrzejewski and David Buttler

Lawrence Livermore National Laboratory
Livermore, CA (USA)

Summary

We propose a method for improving ad-hoc information retrieval by allowing explicit user feedback over topics automatically learned from the corpus using the Latent Dirichlet Allocation (LDA) [1] model. This capability may be especially useful within organizations with specialized domains or limited resources. Experiments on TREC data with simulated user feedback show improved retrieval performance, in addition to the informational benefits of the displayed topics.

Information retrieval in challenging environments

Conditions	Impaired techniques
Non-expert user	keyword queries
Lack of metadata	faceted search
Specialized domain	WordNet
Small user base	query log mining
	relevance feedback

1) Topic representation

Typical "Top N" representations of learned topics can be difficult for users to interpret. We combine several methods to construct a more easily understood topic summary.

Topic 11	oil, gas, production, exploration sea, north, company, field, energy petroleum, companies

Topic labels [2]: Petroleum; Recover capitalization

Topic n-grams [3]: state oil company; North Sea, natural gas; production, exploration, field, energy

2) Topic selection

We may have hundreds of learned topics. We use the following procedure to select a handful of topics to present to the user as feedback candidates.

Topics enriched in the top documents returned by the original query

$$\text{Enriched } E \quad \theta_d(t) = P(t|d)$$

+

Topics correlated with enriched topics

$$\text{Related } R \quad \Sigma(t,t') = \text{cov}(\theta)$$

—

Topics with low semantic coherence [4]

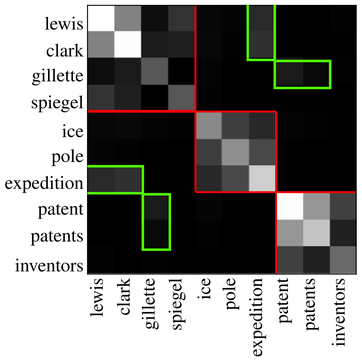
$$\text{Dropped } D \quad \text{PMI}(t) < \text{PMI}_{25}$$

3) Incorporating feedback

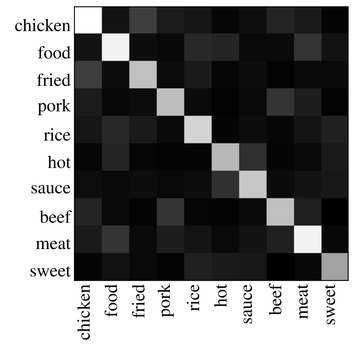
When a user marks a topic as relevant, we expand the original query using the most probable words from that topic. In order to preserve query intent, the original query is given more weight.

"euro opposition" + Emu topic → 0.375 euro 0.375 opposition ... 0.031 European 0.015 Emu

"Junk" topic (low PMI)



Good topic (high PMI)

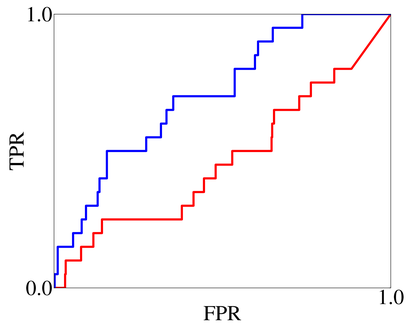


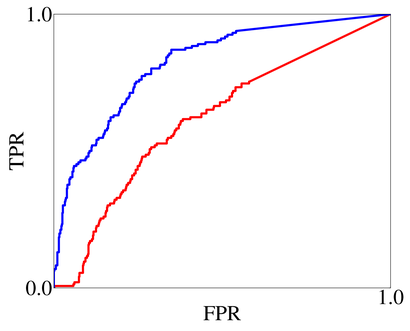
Experimental results

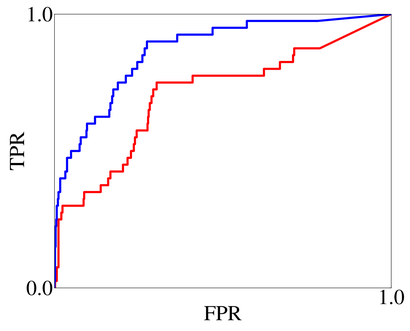
Queries and topics

ROC curves
BASELINE
vs
FEEDBACK

"euro opposition"	"King Hussein, peace"	"law enforcement dogs"
(Emu) economic monetary union Maastricht treaty, member states European, Europe, Community, Emu	**(Amman)** Majesty King Husayn al Aqabah, peace process Jordan, Jordanian, Amman, Arab	**(heroin)** seized kg cocaine drug traffickers, kg heroin police, arrested, drugs, marijuana







References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. (JMLR 2003)

[2] J.H. Lau, D. Newman, S. Karimi, and T. Baldwin. Best topic word selection for topic labelling. (COLING 2010)

[3] D. Blei and J. Lafferty. Visualizing topics with multi-word expressions. Technical report, 2009. arXiv:0907.1013v1 [stat.ML]

[4] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. Evaluating topic models for digital libraries. (JCDL 2010)

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-471258). LLNL-POST-495651