6C-03

トピック型ブーリアンクエリモデルおよび一般的なランキングモデルを用いた学術論文検索システムの構築

福田悟志† 冨浦洋一†

九州大学†

1. はじめに

学術論文検索では，膨大な論文集合から情報要求を満たす比較的少量の論文を網羅的に発見することが求められている．多くの論文検索エンジンでは，ユーザに対して，自身の情報要求が反映されたブーリアン型検索クエリを入力することを求めている．しかし，クエリ内の検索語が論文の著者によって別の語あるいは表現で記述されている場合，その論文は検索されない．一方で，検索語と同一の意味を持つ語を人手で列挙することは困難である．

我々はこれまで，ユーザが作成したクエリに基づいたLatent Dirichlet Allocation (LDA)[1]による論文検索手法を提案した[2]．この手法では，ユーザが作成した検索クエリおよび抄録内の各単語をトピックに置き換えることで，トピック型のブーリアンクエリにヒットする抄録を出力する．また，トピック分析において，LDA に与える最適なパラメータは，分析する文書集合のサイズや単語の分布などによって異なるが，様々なパラメータの組み合わせ下における分析結果を統合することで，その集合が持つトピックの傾向を捉えることができると仮定し，論文のランク付けを行っている．このアプローチにより，ユーザが想定していない検索語の同義語や類似表現を潜在的に考慮した網羅的な論文検索モデルを構築した．

上記の論文検索モデルでは，パラメータの組み合わせ数に基づいて，クエリを満たす論文の出力回数によってランク付けする粗いランキングを行う．一方で，クエリ尤度モデルに代表されるようなクエリベースによる一般的な情報検索モデルでは，各文書における文書モデルに基づいてクエリとの関連の度合いを算出し，ランク付けを行う．このように，両モデルにおけるランク付けのアプローチは異なるため，上位にランクされる論文は大きく異なる可能性がある．そこで本研究では，学術論文検索において，2 種類の検索モデルを統合することで，更なる検索性能の向上を目指す．クエリ尤度に基づいた検索モデルとして，本研究では，LDA を用いた検索語の同義語や類似表現を潜在的に考慮したクエリベースのランキングモデルのひとつである Wei らのランキングモデル[3]を用いる．

Exhaustive Search of an Academic Paper Using a Topic-Based Boolean Query and a General Ranking Model
Satoshi Fukuda†, Yoichi Tomiura†
†Kyushu University

2. 2 種類の検索モデルを組み合わせた論文検索手法

2.1. トピック型ブーリアンクエリによる論文検索モデル (Topic search)

この論文検索モデルでは，まず，ユーザに以下の形式によるクエリの作成を要求する．

(w_1 OR w_2 OR … w_m) AND (w_1' OR w_2' OR … w_n') AND …

クエリは，同一の検索概念を表現する同義語などの別表現を OR で結合するものとする．システムにおいては，OR で結合されている検索概念を同一の特殊記号に変換する．同様に，各抄録内で出現する検索語を，上記の変換に対応した記号に変換する．これにより，これらの語が出現する位置に同一のトピックが付与されることを保証する．次に，抄録集合に対して LDA によるトピック分析を行い，ユーザが作成したクエリにヒットする抄録集合から各特殊記号に付与されたトピックを調べ，以下の形式に従ってトピック型クエリに変換する．

$$\mathrm{AND}_{i=1}^{I}\left(t_{i,1}\ \mathrm{OR}\ t_{i,2}\ \mathrm{OR}\ \cdots \mathrm{OR}\ t_{i,J_i}\right)$$

Iは特殊記号の数を表し，$\{t_{i,1}, t_{i,2}, \cdots, t_{i,J_i}\}$は，$i$番目の特殊記号に付与されていたトピックを表す．その後，抄録内の各単語(トークン)に付与されたトピックからその抄録が持つトピックを調べる．最後に，上記のトピック型クエリを満たすトピック集合を持つ論文を出力する．そして，このアプローチを，LDA に与えるいくつかのパラメータの組み合わせ下で実行し，各論文の出力回数に基づいて降順でソートする．

2.2. Wei らの情報検索モデル (LDA+LM search)

Wei らのクエリ尤度モデルは，(1)式のように定義されている．

$$P(Q|D) = \prod_{q \in Q} P(q|D) \qquad (1)$$

Dは文書(論文抄録)，qはクエリ集合Qにおけるクエリを表している．$P(Q|D)$はQを生成する文書

モデルの尤度を表し，$P(q|D)$ は，(2)式のように，トピックモデルと言語モデルを統合したモデルとして表されている．

$$P(w|D) = \lambda\left(\frac{N_D}{N_D+\mu}P'(w|D) + \left(1-\frac{N_D}{N_D+\mu}\right)P'(w|coll)\right) + (1-\lambda)\left(\sum_{t=1}^{K}\frac{n_{-i,j}^{(w_i)}+\beta_{w_i}}{\sum_{v=1}^{V}n_{-i,j}^{(v)}+\beta_v}\times\frac{n_{-i,j}^{(D_i)}+\alpha_{z_i}}{\sum_{t=1}^{T}n_{-i,t}^{(D_i)}+\alpha_t}\right) \quad (2)$$

N_DはDにおける単語のトークン数，$P'(w|coll)$は文書集合全体におけるwの最尤推定を表し，λ，μはスムージングパラメータを表す．$n_{-i,j}^{(w_i)}$は，トピックjが付与された単語w_iの数を表し，$n_{-i,j}^{(D_i)}$はjが付与された文書D_iにおける単語の数を表す．

2.3. ハイブリッドモデル (Hybrid search)

上記で述べた2種類の手法により決定されたランクに基づき，論文抄録に与えるランクを以下のように決定する．

$$r_3 = \min(r_1, r_2) \quad (3)$$

r_1，r_2はそれぞれ Topic search および LDA+LM search により決定されたランクを表す．すべての抄録に対して(3)式によりランクが決定した後，r_3の値に基づき昇順でソートする．

3. 実験

3.1. 実験設定

実験では，NTCIR-1,2 情報検索用テストコレクション[4][5]を用いた．このデータセットには，132 件の検索課題および各課題に対して「適合」「部分適合」「不適合」のラベルが付与された英語論文データ(約 1,000 から 4,000 件)が含まれている．本実験では，適合論文が 10 件～100 件程度含まれている 検索課題から 40 課題を用いて実験を行い，「適合」「部分適合」が付与された論文を関連論文として評価した．また，各課題におけるクエリは，1 人の被験者が課題内容を読むことで作成した．評価尺度には，ランキング結果の上位 1%から 100%までの各ポイント(1%刻み)における累積再現率を用いた．

パラメータ設定として，Topic search ではα，β，Kを LDA に与えるパラメータの組み合わせの対象とし，$\alpha = \{0.01,\ 0.02, 0.05, 0.1, 0.2, 0.5\}$，$\beta = \{0.01, 0.02, 0.05, 0.1,\ 0.2, 0.5\}$，$K = \{7, 8, 9, 10, 11, 12\}$とした．LDA+LM search における$\alpha$，$\beta$，$K$，$\lambda$，$\mu$は，ランキング結果における累積再現率のグラフ面積が最大であった 0.01，0.2，7，0.5，10 とした．また，LDA に対するギブスサンプリングの回数は 10,000 と設定した．

3.2. 実験結果および考察

各手法によるランキング結果における平均累積再現率の結果を図 1 に示す．図 1 から，0.77 から 0.92 までの再現率において，Hybrid search では他の 2 手法と比べて，同等の再現率をより高いランクで示していることが分かった．例えば，0.82 の再現率を，Hybrid search ではランキング結果の上位 31%，ほかの 2 手法では上位 36%で示している．また，0.85 の再現率の場合，Hybrid search ではランキング結果の上位 33%，ほかの 2 手法では上位 42%で示している．すなわち，関連論文をより多く含む論文集合を獲得する場合，Hybrid search を適用することで，他の 2 手法と比べて約 5～9%の不要な論文を除去して提示することができると考えられる．

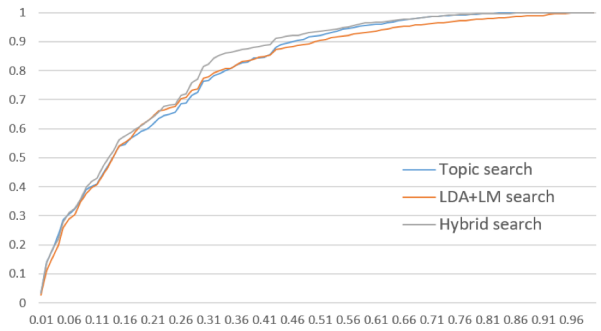

図 1 各手法によるランキング結果における平均累積再現率

4. おわりに

本研究では，学術論文検索タスクにおいて，2 種類の検索モデルを組み合わせることによる論文検索性能の向上を検討した．

謝辞

この研究は科研費 JP15H01721 の助成を受けたものである．

参考文献

[1] T.L. Griffiths and M. Steyvers.: Finding scientific topics. In: National Academy of Sciences, pp. 5228-5253 (2004).
[2] S. Fukuda and Y. Tomiura.: Exhaustive search of academic paper using topic-based boolean query. In: International Symposium on Information Technology Convergence (2018).
[3] X. Wei and W.B. Croft.: LDA-based document models for ad-hoc retrieval. In: SIGIR, pp. 178-185 (2006).
[4] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, S. Hidaka, and J. Adachi.: The NTCIR workshop: The first evaluation workshop on Japanese text retrieval and cross-lingual information retrieval. In: Information Retrieval with Asian Languages Workshop, pp. 1-7 (1999).
[5] N. Kando.: Overview of the second NTCIR workshop. In: NTCIR Workshop, pp. 35-43 (2001).