情報検索における関連性の重ね合わせモデルの効果

3U-1

金沢 輝一† 高須 淳宏‡ 安達 淳‡
†東京大学大学院工学系研究科 ‡国立情報学研究所

1 はじめに

筆者らは情報検索における自然言語の意味曖昧性への対処として関連性の重ね合わせモデル (RS モデル) を提案している．この手法は，著者キーワードなどの情報に基づいて文書をクラスタリングして特徴ベクトルを補正するもので，索引語の重要度計算を tf·idf などの手法より高い精度で行うものである．RS モデルは文書間の同概念異表記の問題に対して大きな効果を持っており，学術文書の検索などに有効であることが確認されている．一方，報道記事のように表記が統制される傾向にある文書に対しての RS モデルの効果は学術文書の場合に比べて小さい．報道記事検索では問い合わせ表現 (query) の曖昧性の問題が相対的に大きくなると考えられるため，本発表では query expansion などの問い合わせを補正する手法との融合による検索精度の向上を検討する．

2 RS モデルと検索システム R^2D^2[1]

関連性の重ね合わせモデル (RS モデル) は，筆者らが提案している意味曖昧性への対策手法で，検索対象の文書間に存在する関連性に基づき非排他的な文書クラスタを作り，これを解析することで文書ベクトルを拡張するものである．

R^2D^2[(*1)]は RS モデルを適用した文献検索システムで，NTCIR[(*2)][2] の国内学会発表抄録データベースあるいは TREC[(*3)][3] の報道記事，行政文書データベースを対象に検索を行うことができる．現在の実装では各文書に予め付与されているキーワードを解析し，同一キーワードを含む文書で RS モデルのクラスタを形成している．

Effect of the Relevance-based Superimposition Model on Information Retrieval.
Teruhito KANAZAWA†, Atsuhiro TAKASU‡, Jun ADACHI‡
† Graduate School of Engineering, Univ. of Tokyo
‡ National Institute of Informatics
(*1)RetRieval system for Digital Documents
(*2)NII–NACSIS Test Collection for IR systems
(*3)Text REtrieval Conference

3 query expansion

問い合わせ表現の拡張 (query expansion; 以下 QE) は，検索者が入力した問い合わせ表現の関連語をシソーラスあるいは検索対象のデータベースから選択して問い合わせに加えることで，問い合わせの意味曖昧性に対処する手法である．検索者の意図と合致する語だけを自動的に選択して補うことは困難であるため，候補となる語を列挙するにとどめ，選択は検索者自身が行うという方式が一般的である．

関連語の選出はシソーラスを用いた方式と予備検索の結果から関連語を抽出する relevance feedback の方式とに大別されるが，本実験では後者の手法による自動の QE を実装した．

tf·idf に基づく索引を用いて本来の問い合わせ表現で検索を行った結果の上位 D 件の文書に含まれる索引語のうち，tf·idf の平均が大きい T 語を検索語に補う．D と T の最適値は $D = 30, T = 10$ であった．

4 実験

今回の実験では TREC4 テストセットの一部である San Jose Mercury News(90,257 文書) に対して，用意されている問い合わせ 50 件のうち正解文書を 5 つ以上含む 37 件を対象にそれぞれ最大上位 1000 件における再現率と適合率を求めた．

表 2 は各手法ごとの平均適合率と，RS モデル，QE のいずれも適用しなかった場合を baseline とした増減である．RS モデルと QE を併用した場合が最も検索精度が高く，baseline に比べて約 12%向上した．

表 1 は，RS モデルによって平均適合率が 0.05 以上増加した 2 つの問い合わせに関して，各手法ごとの平均適合率と basline からの増減である．2 件とも，RS モデルと QE を併用した場合が最も検索精度が高くなっている．

表 3 は RS モデルと QE の併用による効果が大きかった問い合わせにおいて，問い合わせのベクトルと関連性が高いとされた文書クラスタが QE によってどのように変化したかを示しており，元の問合わ

表 1　RS と QE の併用による効果

Query ID	baseline	RS	QE	RS+QE
216	.4582	.5775(+26%)	.6134(+34%)	.6414(+40%)
230	.1538	.2224(+45%)	.3672(+139%)	.4638(+202%)

せに対して最も関連性が高いと認識されたが正解を含まないクラスタ「automobile」は QE で補正された問い合わせに対する関連性が低くなっており，正解を含むクラスタ「recharge」や「electric」は QE 後の問い合わせに対して関連性が高くなっている．RS モデルでは問合わせに対して関連性の高いクラスタが多くの正解文書を含んでいる場合に効果が大きく，すなわち QE が適切に問い合わせを補正することで RS モデルの寄与を高めていることを示す結果といえる．

5　考察

RS モデルに関する過去の実験 [4] において，TREC の文書に付与されているキーワードは種類，文書あたりのキーワード数共に NTCIR に比べて少ないために RS モデルの寄与も小さいことが分かっている．TREC4 SJM では RS モデルの適用によって平均適合率が 0.05 以上向上した問い合わせは 37 件中 2 件であった．表 2 は 37 件の平均適合率であるために，QE のみ適用した場合と RS モデルと QE を併用した場合の違いは 1%程度に留まっている．しかし RS モデルの寄与が大きい 2 件に注目すると RS モデルと QE の併用によって，それぞれを単体で適用した場合を上回る検索精度が得られている（表 1）．

表 3 でキーワード recharge を含む文書のクラスタは全文書数 8 のうち 4 件が正解であり，このようなクラスタは検索精度の向上に大きな作用を持っていると考えられる．一方，キーワード electric を含む文書のクラスタは全正解の 8 割を含んでいるが，クラスタ内文書の 9 割は不正解であり，このクラスタによる文書ベクトル補正はノイズの多いものとなってしまう．RS モデルの効果を高めるには文書クラスタの粒度を下げる必要があると考えており，フレーズキーワードによるクラスタリングや，クラスタのサブトピック分割などの手法の導入を検討している．

表 2　各手法の平均適合率

	QE なし	QE あり
RS なし	.2318	.2578 (+.0260, 11%)
RS あり	.2388 (+.0070, 3%)	.2605 (+.0287, 12%)

表 3　**Query 230: ～ develop and produce an electric-powered automobile ～** に関連性の高い文書クラスタの，**QE** による変化

正解文書数は 34.

QE なし

順位	キーワード	クラスタ内文書数	正解文書数
1	automobile	70	0
2	electric	294	28
3	recharge	8	4
4	efficient	50	0
5	GE	24	0

QE あり

順位	キーワード	クラスタ内文書数	正解文書数
1(↑)	recharge	8	4
2(—)	electic	294	28
3(↑)	efficient	50	0
4(↓)	automobile	70	0
5(↑)	solar	47	3

謝辞

筆者らは，NACSIS コレクション (NTCIR) ワークショップに参加し，本研究では，NACSIS 研究開発部が「学会発表データベース」のデータの一部を使用して，データ提出学会(*3) の理解の下に構築した「テストコレクション 1」を利用した．

参考文献

[1] Kanazawa, T., "R^2D^2 at NTCIR: Using the Relevance-based Superimposition Model," *Proc. of NTCIR Workshop 1*, pp.83 – 88, Aug. 1999.

[2] NTCIR: `http://www.rd.nacsis.ac.jp/~ntcadm/`

[3] TREC: `http://trec.nist.gov/`

[4] 金沢輝一, 高須淳宏, 安達淳, "英語テキストにおける関連性の重ね合わせモデルの検索特性," 情処研報 2000-DBS-122, pp.57 — 64, July 2000.

(*3) `http://www.rd.nacsis.ac.jp/~ntcadm/acknowledge/thanks1-ja.html` 参照