5K－1

関連性の重ね合わせモデルによる問い合わせ表現の自動拡張手法

金沢 輝一† 高須 淳宏‡ 安達 淳‡

†東京大学大学院工学系研究科 ‡学術情報センター研究開発部

1 はじめに

筆者らは著者キーワードなどの関連性に基づく非排他型の文書クラスタを用いて文書ベクトルを拡張することで検索精度を高める「関連性の重ね合わせモデル（RSモデル）」を提案してきた[1]．RSモデルはtf·idfによる索引語の重み付けを補正して検索対象文書の持つ意味曖昧性に対処するものである．一方，不慣れな検索者が入力した問い合わせ表現(query)に関連語などを補うことで検索性能を向上するquery expansionでは，元々情報量の小さい問い合わせ表現を高い精度で拡張することが難しく，既存の自動化された手法では元の表現による検索結果がある程度の精度を有していない場合には効果が得られないという問題点も指摘できる．本発表ではRSモデルの文書クラスタを利用して検索者の入力した問い合わせ表現を自動的に拡張する手法を提案し，従来型の拡張手法との比較を含めた評価を行う．

2 query expansion

2.1 既存手法の問題点

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，従来は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った．

関連語の抽出はシソーラスを用いた方式と予備検索から関連語を抽出するrelevance feedbackの方式とに大別されるが，前者は辞書構築のコストや問い合わせとの関連度を動的に決定することの難しさなどの課題を抱えており，後者は予備検索の精度が関連語の質を左右するというジレンマを持っている[2]．

Mitraらはrelevance feedbackの際に補う語の共起確率を考慮することで完全自動QEの性能を向上させる手法を提案しており[3]，我々はそれを元に完全自動のQEを検索システムR^2D^2[1]に導入した．以下にその内容を示す．

2.2 自動 query expansion

tf·idfに基づく索引テーブルに対して元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自立語を抽出し，新たな検索語$\mathrm{t}_{\mathrm{new}}$として問い合わせとの関連度$r$を次式で求める．この式は$R^2D^2$における検索語の重み付け式の応用である．

$$r(\mathrm{t}_{\mathrm{new}}) \equiv \frac{1}{|D|}\sum_{\mathrm{d}\in\mathcal{D}} f(\mathrm{d})^2 \qquad (1)$$

$f(\mathrm{d}) \equiv$ (文書dに含まれる本来の検索語の数)

$\mathcal{D} \equiv$ (新しく加える語$\mathrm{t}_{\mathrm{new}}$を含む文書の集合)

次に，関連度の高いほうからT語を検索語に補う．ただし新たな問い合わせ表現による各文書の得点はMitraらの手法に基づき，

$$d \equiv \sum_{\mathrm{t}\in(\text{元の表現})} r(\mathrm{t})\times w(\mathrm{t}) + \sum_{i=1}^{T} r(\mathrm{t}_{\mathrm{new}\,i})\times w(\mathrm{t}_{\mathrm{new}}) \times \min_{j=1}^{i-1}(1 - P(\mathrm{t}_{\mathrm{new}\,i}|(\text{元の検索語} \cup \mathrm{t}_{\mathrm{new}\,j}))) \qquad (2)$$

とする．$P(\mathrm{t}_{\mathrm{new}\,i}|(\text{元の検索語} \cup \mathrm{t}_{\mathrm{new}\,j}))$は，加えた検索語$\mathrm{t}_{\mathrm{new}\,i}$と，それよりも関連度の高い検索語の共起確率であり，以下の式で推定される．

$$\frac{\left(\begin{array}{c}\text{全文献の中で検索語 } \mathrm{t}_{\mathrm{new}\,i}\text{を含み，かつ}\\ \mathrm{t}_{\mathrm{new}\,j}\text{あるいは元の検索語のいずれかを含む文書数}\end{array}\right)}{(\text{全文献の中で検索語 } \mathrm{t}_{\mathrm{new}\,i}\text{を含む文書数})}$$

また，$w(\mathrm{t})$はtf·idfに基づく語tの重みである．

3 RSモデルとquery expansionの融合

自動QEは元の問い合わせ表現に対する検索結果の適合率が低い場合に十分な効果を上げることができず，みなし正解を多くとると不正解文書が含まれる率が増し，結果としてノイズとなる語が補われてしまうという問題を持っている．そこで我々は前章

Query Expansion Method using the Relevance-based Superimposition Model.

Teruhito KANAZAWA†, Atsuhiro TAKASU‡, Jun ADACHI‡

† Graduate School of Engineering, Univ. of Tokyo

‡ R & D Department, NACSIS

で述べた従来型の手法を改良した，RSモデルにおける著者キーワードによる文書クラスタを用いて補う語を選択する手法を提案する．

3.1 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集合を作り，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)]の国内学会発表抄録データベース332,921件を対象に検索を行うものである．本研究ではNTCIRの方法に則り「テストコレクション1」として用意された自然文1フレーズずつ計83件の問い合わせについて，それぞれ最大上位1000件におけるランクA正解（完全にレレバント）の再現率と適合率を求めた．

3.2 代表ベクトルに基づいた拡張語の選択

自動QEの問題を克服するものとして，我々はRSモデルにおける著者キーワードによる文書クラスタを用いて補う語を選択する手法を提案する．すなわち，文書クラスタに付与された代表ベクトルは文書ベクトルと同次元であり，queryベクトルと代表ベクトルの間の関連度を求めることが可能であるという性質を利用して，問い合わせとの関連度の大きいC個のクラスタに含まれる索引語から式(1)による関連度の大きい順にT語を選び，問い合わせに補う．

3.3 実験

評価実験では問い合わせ#1–30をパラメータ訓練用として用い，従来手法のT, D，提案手法のT, Dの最適値として$T = 10$(共通)$, C = 20, D = 5$を得て，問い合わせ#31–83に適用して評価とした．表1では評価用に対するQEの寄与がほとんど現れないが，表2によると各々の問い合わせに平均7%以上の検索精度の変動をもたらしていることが分かる．すなわちQEの効果は問い合わせに強く依存しており，特定の条件で得られた最適パラメータを一般化して適用することは困難であるといえる．

また評価用問い合わせにおいて既存のQEと提案手法の適合率変動の相関係数を求めたところ0.6578であったことから，提案手法は既存の手法とは大きく異なる特性を持っていることが分かった．

(*1)RetRieval system for Digital Documents
(*2)NACSIS Test Collection for IR systems

表1　QEを組み合わせた場合の平均適合率
※比率はそれぞれの平均適合率（表1）に対するもの

問合せ	RS	+QE(docs)	+QE(clsts)
#1–30	.3919	.4065 (+4%)	.4064 (+4%)
#31–83	.3289	.3297 (+0%)	.3302 (+0%)

表2　問合せ毎の適合率変動（絶対値）の平均

問合せ	+QE(docs)	+QE(clsts)
#1–30	.0351 (9%)	.0405 (10%)
#31–83	.0229 (7%)	.0358 (11%)

4 考察

訓練用問い合わせの検索結果では，既存の自動QEの弱点であった，適合率が低い状況において提案手法が検索精度を向上する特性が示されたが[1]，評価用問い合わせではこれを裏付ける十分な結果は得られなかった．一方，実験で使用したデータセットの特性を分析すると訓練用は問い合わせ毎の正解文書の平均が約106文書（情報量約13.4ビット）であるのに対し，評価用は平均約36文書（約14.1ビット）とレレバント条件の厳しさに大きな違いがある．このような条件の違いが結果に影響を与えていることも考えられる．

新田らの研究[4]では文書クラスタを利用したQEの手法を提案し，TRECにおいて一定の効果を示していることなどから，今後は実験条件を変えて我々の提案手法の特性を詳しく分析していきたい．

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] “関連性の重ね合わせモデルに基づく問い合わせ表現の拡張,” 金沢 輝一, 高須 淳宏, 安達 淳, 情処研報, 99-DBS-119-51, pp.303–308, Jul., 1999.

[2] “The impact of query structure and query expansion on retrieval performance,” Keäläinen,J., Järvelin,K., *SIGIR'98*, pp.130–137, 1998.

[3] “Improving Automatic Query Expansion,” Mitra,M., Singhal,A., Buckley,C., *SIGIR'98*, pp.206–214, 1998.

[4] “文書クラスタリングを利用した検索質問展開手法の開発と評価,” 新田 清, 蓬莱 尚幸, 園部 正幸, 情処研報, 99-DBS-118-2, pp.9–16, May., 1999.

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