論　文　の　内　容　の　要　旨

論文題目　文書クラスタを用いた情報検索のモデルとその応用に関する研究

氏　名　金澤　輝一

コンピュータの能力向上とインターネットの普及によって情報の生産，流通，蓄積を低コストで大量に行うことが可能となった．一方，人間が情報を処理する能力は限られており，過多となった情報から選択的に入手し，あるいは蓄積した情報の中から素早く任意の内容を取り出すといった処理の重要性はますます増大している．しかし，検索処理における代表的な入力形態である自然言語には必ず意味的な曖昧性が存在し，検索の精度を低下させている．言葉の意味の定義は使用者ごと，あるいは状況によっても微妙に異なり，概念と表現を一対一に対応づけることはできない．言語間では文法も単語も全く異なるということも同種の問題である．問い合わせと検索対象の間で比較を行いたいのは述べられている概念だが，実際に比較できるのは表現である．表現の一致と概念の一致は必ずしも等価でないため検索精度は低下する．自然言語を問い合わせに用いる限り，これは避けられない問題であり，何らかの手法で対策を講じることが重要である．本研究では自然言語の意味曖昧性がもたらす検索精度低下の問題に対処すべく関連性の重ね合わせ (Relevance-based Superimposition) モデルを提案し，その特性の評価を行った．関連性の重ね合わせモデルの特徴は，検索対象の文書が持つ情報，特に文書関連性に着目した点である．従来，意味曖昧性への対処は問い合わせ表現に注目する手法が多かったが，問い合わせの限られた情報から的確に意図を汲み取るのは難しい．そこで本研究では検索対象の文書の持つ情報，特

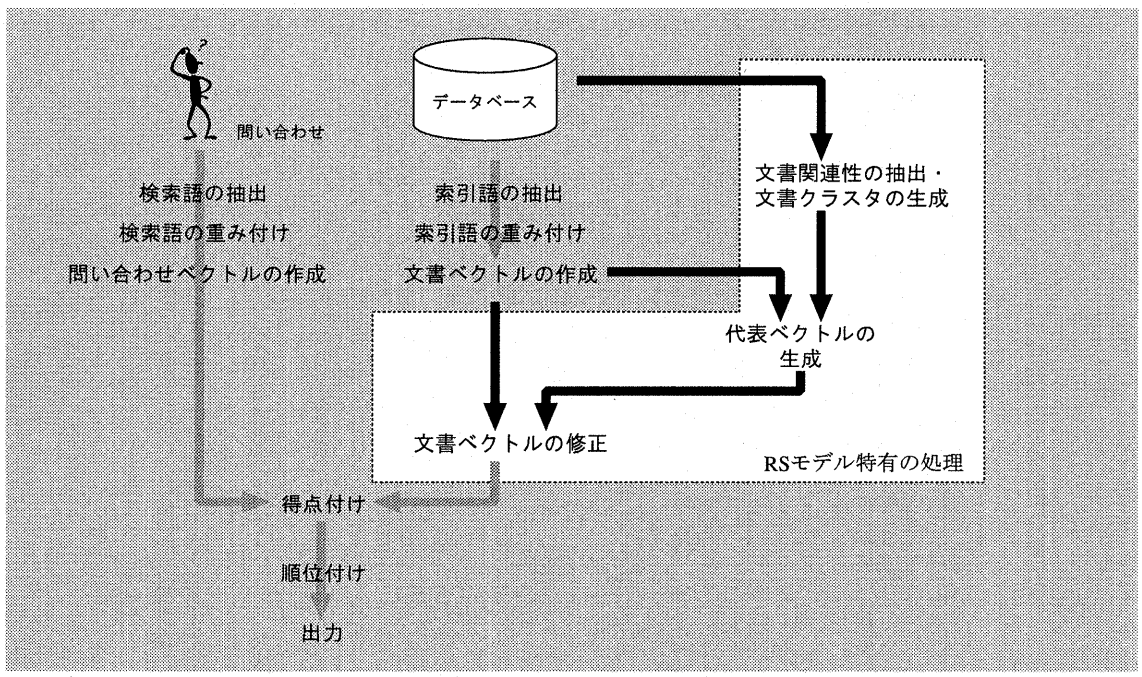

図 1　提案手法 (RS モデル) の処理の流れ

に文書関連性に着目して意味曖昧性への対処を図った．大規模テストセットを用いた評価実験では，従来手法である問い合わせ表現の自動拡張 (automatic query expansion) と比較して，検索対象の文書の話題や言語といった条件に対してロバストな効果を示し，特に学術文献の検索に対しては最大 9%の検索精度向上を達成した．また，問い合わせ表現の拡張と組み合わせた場合に相補的に効果を高めることも分かった．この場合，最大 12%の精度向上効果を得られた．

表 1　評価実験における検索精度の向上

手法の記号は baseline が tf·idf のみによるもの，QE が従来手法である問い合わせの拡張，RS が提案手法である関連性の重ね合わせモデル，QE+RS が両者の併用を意味する．
精度は平均適合率と baseline からの向上率．

	NTCIR 1 J-J	NTCIR 2 J-J	NTCIR 2 E-E	NTCIR 2 J-E	TREC 3 SJM
baseline	.3059	.2841	.2984	.2369	.2318
QE	.3270 (+7%)	.2886 (+2%)	.3044 (+2%)	.2476 (+5%)	.2578 (+11%)
RS	.3344 (+9%)	.3020 (+6%)	.3160 (+6%)	.2522 (+6%)	.2388 (+3%)
QE+RS	.3381 (+11%)	.3103 (+9%)	.3226 (+8%)	.2653 (+12%)	.2605 (+12%)