近接演算による数値情報検索の精度改善

2P－2

山田洋志　福島俊一

NEC　ヒューマンメディア研究所

1　はじめに

数値の記述は多くの文書で頻繁に使われ,内容にも強く関わっている．検索の際に,製品の値段や性能で結果を絞り込みたいなど,数値の指定で,より正確に検索できる．このことから,従来,数値を検索に利用するシステムが開発されている[3, 4]．これらのシステムでは,構文解析を利用して数値が表す対象を判断し,高度な検索を実現している．反面,解析の誤りや曖昧性による再現率低下が発生する．

筆者らは,構文解析や意味解析を使わず,単位付きの数値と単語のAND検索で数値検索を実現する方式を開発した[1, 2]．本方式を情報検索システム評価用テストコレクションBMIR-J2[5]で評価したところ,単語だけによる検索と比較して,数値条件の指定により再現率をほとんど下げずに適合率を大幅に上げられることが分かった．

しかし,検索要求によってはまだ十分な適合率が得られていない．その原因の一つが,数値と単語の関係を考慮していないため,検索要求とは無関係の数値を含む文書が検索されることである．

本稿では,この問題を解決するために単語と数値との近接演算の導入を提案し,検索精度の評価結果についても報告する．

2　数値情報検索方式と課題

文献[1, 2]で提案した方式では,テキストを形態素解析し,そこから数値表現を抽出してインデックスを作成する．数値に付随する範囲表現や概数表現(「約」「以上」など)を考慮して数値に換算することで,概数や範囲表現も検索対象にできる．

検索時にユーザは単語と数値(+単位)の条件を指定する．検索システムは,指定された単語と数値条件に合う数値表現の両方を含む文書を検索結果とする．

検索精度評価には,(社)情報処理学会・データベースシステム研究会が,新情報処理開発機構との共同作業により,毎日新聞CD-ROM'94データ版を基に構築した情報検索システム評価用テストコレクションBMIR-J2[5]を利用した．

表 1: 検索要求と検索条件

検索要求115

検索要求	1ドル＝100円を超える円高	
検索語	円高	
数値条件	100円以下	(50–100)円の範囲

検索要求117

検索要求	千人以上の人員削減を計画している企業
検索語	人員 & 削減
数値条件	1000 人 以上

表 2: 数値条件による検索精度

単語のみ		単語+数値1		単語+数値2	
適合率	再現率	適合率	再現率	適合率	再現率
検索要求115					
72/370	72/80	72/106	72/80	72/100	72/80
19.5%	90.0%	67.9%	90.0%	72.0%	90.0%
検索要求117					
5/25	5/11	5/11	5/11		
20.0%	45.5%	45.5%	45.5%		

評価に用いた検索条件を表1に示す．“検索要求”はBMIR-J2で定義されている検索要求であり,詳細な説明がついている(表では省略)．“検索語”と“数値条件”は検索要求を見て人手で作成した．検索要求115には2通りの数値条件を設定した．

適合率・再現率を表2に示す．表2で上段は記事数,下段が百分率である．再現率を下げずに適合率を大幅に向上できていることが分かる．

本方式の課題として,数値と単語の関係を考慮しないことによる適合率向上の限界がある．例えば,円高の話題の中で値上げについて書かれていて,それが数値条件に一致すると検索されてしまう．

3　近接演算の導入

前節で挙げた課題を解決するために,数値と単語の関係を考慮した検索条件を導入する．処理の高速性や頑健性を維持するために,厳密な構文・意味解析の代わりに,近接演算を使用する．

“近接”の基準として段落を使用した．すなわち,検索語と数値情報が同じ段落に含まれているものを

Improvement of a Numerical Expression Retrieval Method with Proximity

Hiroshi Yamada and Toshikazu Fukushima

Human Media Research Labs. , NEC Corp.

表 3: 近接演算による検索精度(検索要求115)

	単語+数値1		単語+数値2	
	適合率	再現率	適合率	再現率
近接	57/83	57/80	57/77	57/80
演算	68.7%	71.3%	74.0%	71.3%
近接	66/93	66/80	66/87	66/80
+見出し	71.0%	82.5%	75.9%	82.5%

表 4: 近接演算による検索精度(検索要求117)

	単語のみ		単語+数値	
	適合率	再現率	適合率	再現率
近接	5/21	5/11	5/8	5/11
演算	23.8%	45.5%	62.5%	45.5%
近接	5/21	5/11	5/9	5/11
+見出し	23.8%	45.5%	55.6%	45.5%

検索した．見出しは一つの段落として扱った．また，記事によっては，見出しに数値や主題(「円高」など)が書かれ記事本文には出てこない場合がある．そこで，見出しを特別に扱い，見出し中の数値や単語については全段落に含まれるものとした評価も行った．すなわち，単語が見出しに出現し数値が本文に出現した場合(又その逆)は近接条件を満たすとした．

4 結果と考察

結果を表3，表4に示す．“近接+見出し”は，見出しを特別扱いした場合である．検索要求117では二つの検索語に対しても近接条件を適用したため，単語のみの検索の場合にも結果に変化が生じる．

検索要求117については，検索もれの増加がなく，不要の結果が減少し，近接演算の効果が現れている．不要な検索結果が6件から3件に減少した．減少したのは，無関係の数値の記載，「人員」と「削減」が別の段落に出現，全従業員数の記述，が各1件である．近接演算でも絞れなかった3件は，企業以外の人員削減が2件，削減実績の記事が1件で，話題は関連しているが，検索要求の詳細に合っていない．これらは検索要求と記事の理解が必要となる．

検索要求115については，適合率が1-4%向上したが，再現率が8-19%落ちている．近接演算の導入によって，過剰な検索結果が28件から20件に減った．減少した8件のうち5件は「円高」とは無関係な数値が含まれるもので近接演算の効果が現れている．減少した残り3件は円高を予想する記事で，「円高」と数値が別の段落に書かれていた．以下に改善例を示す．この記事は「100円以下」の金額を含んでいるが，電気料金の値下げ分で円相場とは関係がない．

(第2段落) 昨年十一月から実施している【円高】差益還元のための暫定料金引き下げが、今年十月以降も延長される見通しが高まっている。

(第6段落) 家族四人の標準世帯で電気料金が十社平均で月額【九十八円】程度、都市ガス三社が同百三十六円程度引き下げとなった。

近接演算で削減できなかった20件は，「100円に接近する」などの表現を含む記事や，円高の予想についての記事が多く，数値抽出の改良や内容理解を必要とする．また，無関係の数値を含む記事が1件あり，これはマルクの相場が「円高」と同じ段落にあった．

本来検索対象であるのに，近接演算によって漏れた記事が15件ある．このうち9件は，前節で述べた見出しの特別扱いで対処できた．残りの6件は検索要求に合う内容であるが，数値と「円高」が別の段落に記述されており，近接演算では対処が難しい．

検索要求115と117で再現率への影響が異なる原因として，数値と単語の関係の違いが考えられる．

検索要求115「100円を超える円高」では，「90円」などの数値が円高の意味を含んでいるため，「円高」と記述しなくても情報を伝えられる．そのため，検索要求115の正解記事では，数値と「円高」が分散して出現する傾向があり，近接演算の効果が少なくなる．一方，検索要求117では，「人員削減」というテーマと人数の両方がそろって初めて一つの情報になる．そのため，数値と単語(「人員」，「削減」)がまとまって出現しやすく，近接演算の効果が大きくなる．

5 おわりに

数値条件と単語を組み合わせて使用する数値情報検索方式において，適合度を改善するために近接演算を導入した．

BMIR-J2に含まれる二つの検索要求を用いた実験で，近接演算によって適合率が1-10%向上することを確認した．また，再現率の低下の程度は，検索要求によって大きく差があることが分かった．

今後，より多くの評価によって，近接演算がどのような検索要求に適しているか分析を進めたい．また，再現率よりも適合率が重視されるWWWの検索で，数値検索および近接演算の効果を評価したい．

参考文献

[1] 山田，福島，“テキスト中の数値表現を用いた情報検索方式の評価”，情処58回大会，1U-03，1999

[2] 山田，福島，“数値情報を用いたテキスト検索方式の提案と評価”，情報学基礎研究会，FI-53-3，1999

[3] 岸本ほか，“テキストの構造化に基づく検索システム”，情処論文誌，1994

[4] 斉藤ほか，“数値情報をキーとした新聞記事からの情報抽出”，情処，NL125-6，pp.63-70，1998

[5] 木谷ほか，“日本語情報検索システム評価用テストコレクションBMIR-J2”，情処，DBS114-3，1998