テキスト中の数値表現を用いた情報検索方式の評価

1U－3

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な検索が行える. 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[2, 3]. 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,高度な検索を行っている. 半面,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なしでも,数値条件を組み合わせることで単語だけを指定する場合より適合率を向上させることに着眼して検索方式の開発を行った. さらに,テキストから数値を抽出する際に,概数や範囲表現など多彩な数値表現に対応することで,再現率の低下を抑えた.

本稿では,数値条件の指定により,再現率をほとんど下げず適合率が大幅に上がることを示す. 精度評価には検索用ベンチマークBMIR-J2を使用した.

2 数値検索機能の概要

本方式の評価のため,筆者らの開発した多角的検索システムOTROSに数値検索機能を実装した[1]. 指定された数値条件は,単語とのAND条件として検索結果の絞り込みに利用する.

本方式では,あらかじめ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. 抽出の際に,数値に付随する範囲表現や概数表現(「約」「以上」など)を6種類に分類し,分類に応じて数値に換算する. これにより,概数や範囲表現も検索対象にできる. 修飾表現は,57種類を収集した. これらで新聞,WWWページに出現する数値の修飾表現の95%以上を網羅している[1]. 今回は,このうち使用頻度の多い30種類を実装し,評価に利用した.

3 検索精度の評価

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[4]を利用した.

表 1: 検索要求と検索条件

ID	検索要求
115	1ドル＝100円を超える円高
116	2期以上連続の減益企業
117	千人以上の人員削減を計画している企業

ID	検索語1	検索語2
115	円高	
116	減益	減益 & 連続
117	人員 & 削減	(人員 or 社員) & 削減 & (予定 or 計画)

ID	数値条件1	数値条件2
115	100円以下	(50–100)円の範囲
116	3期 以上	3期以上 or (3–10)年
117	1000 人 以上	

BMIR-J2は,5080件の新聞記事と60件の検索要求,正解記事集合からなる. 標準セットには,数値を条件に含む検索要求が3件ある(表1). 各検索要求に対し,1つまたは2つの検索語と数値条件を人手で作成し(表1),組み合わせて検索精度を測定した. なお,記事集合中のランクA,Bとも正解として扱った.

4 評価結果

適合率と再現率を表2,図1に示す.

図表から分かるとおり,構文や文脈を利用しなくても,数値条件を指定することで適合率が大幅に向上した. 一方,数値条件の指定による再現率低下は少なく,本方式の有効性を確認できた.

4.1 検索もれの原因

検索要求115と117では,数値指定による検索もれは無い. しかし,検索語だけの段階でもれが多い. 主なもれとして,115では,「ドル安」「円相場」「為替相場」,117では,「従業員」を含む記事がある(8件). これらは,同義語辞書の併用で解決できる.

検索要求116では,数値条件1を使用したときに6件の検索もれがあった. その原因を以下にあげる.

単位の同義語 数値条件1では,期数を「年」で記述

Evaluation of a Numerical Expression Retrieval Method

Hiroshi Yamada and Toshikazu Fukushima

Human Media Research Labs. , NEC Corp.

表 2: BMIR-J2による検索精度

検索要求	検索条件	検索語のみ		数値条件1		数値条件2	
		適合率	再現率	適合率	再現率	適合率	再現率
115	検索語1	19.5%	90.0%	67.9%	90.0%	72.0%	90.0%
116	検索語1	34.1%	100.0%	100.0%	60.0%	61.9%	86.7%
	検索語2	83.3%	100.0%	100.0%	60.0%	92.9%	86.7%
117	検索語1	20.0%	45.5%	45.5%	45.5%		
	検索語2	29.4%	45.5%	55.6%	45.6%		

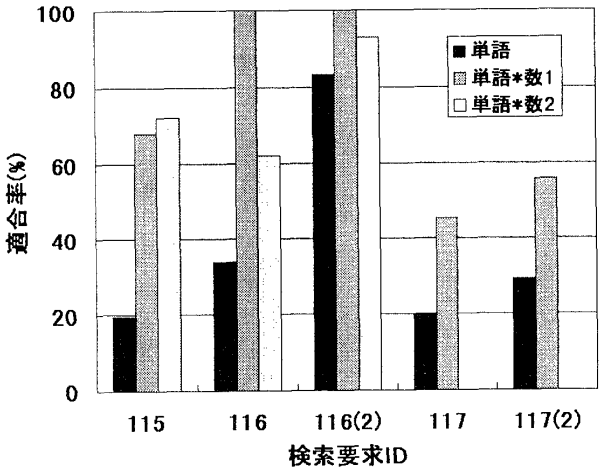

図 1: BMIR-J2による適合率

している記事が検索もれになる(4件). 重要な単位についてはシステム側で同義や表記の違いに対応する必要がある.

内容理解が必要 各期の利益が記述され,内容を理解しないと「3期連続」なのが分からない(1件).
「2期連続の二ケタ減益……95年3月期も……経常利益は1千億円(同9.4%減)と予想」

形態素解析誤り 形態素解析結果から数詞と単位の組み合わせを抽出しているため,解析を誤ると抽出もれに通じる(1件).

4.2 検索過剰の原因

数値条件の指定によって適合率が大幅に向上しているが,まだ,過剰な検索結果が含まれている. その原因を分析する. このうち,始めの2種類は本方式に起因し,残りは文章理解の問題である.

無関係な数値 検索要求とは無関係の数値が同じテキスト内にある場合に検索されてしまう(14件). 実際の記事を調べると,構文・意味解析なしでも,検索語と数値表現との近接演算を使うことでさらに精度を上げられるケースが多い.

修飾表現 本方式では,「約」「以上」「未満」などの数値を修飾する表現を,一定の割合で数値の範囲に置き換えている. そのため,置き換えた範囲が広すぎて絞り込みに失敗する場合がある(2件). 表現の意図する範囲を正確に推測することは困難なため,検索もれが起こらないことを重視して広めの範囲設定にしたことによる.

省略表現 単位や数値の省略で抽出誤りを起こす(1件). 今回,「3〜7万」を「3万〜7万」と判断できなかったために検索過剰を起こした.

内容理解が必要 単なる検討や仮定など,数値を記述していても検索要求に沿わない記事や,「100円に迫る」など文脈なしでは数値の大小関係が判定できない表現がある(23件). 他に,今回の評価結果には影響しなかったが,記事内容から計算で数値を求める必要のある場合がある(3件). 例えば,「千人以上の削減」で「二万一千六百五十人を一万七千四百人にする」を検索するには引き算が必要である.

5 おわりに

数値の条件指定と検索語とをAND条件とする手法の検索精度を,BMIR-J2を利用して評価した. 数値条件を追加することで適合率が大幅に向上し,構文解析などの高度な文書処理を行わなくても数値を適合率向上に利用できることが分かった. 数値の抽出は,構文解析などに比べると,高精度に実現でき,テキストの変化の影響をうけにくいため,本手法は種々のテキストを扱うシステムに特に有効である.

今後,WWWページなど各種のテキストで評価するとともに,単位表記の拡充や数値の修飾語の扱いの精密化など,実用化のための改良を行う.

参考文献

[1] 山田,福島ほか,"インターネット多角的検索システムOTROS", 情処57回大会,3L-01〜04

[2] 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

[3] 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

[4] 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