数量情報を用いた知的検索に関する研究

桑田 大輔　　森田 和宏　　泓田 正雄　　青江 順一

徳島大学 工学部 知能情報工学科 〒 770-8506 南常三島町 2-1

Email: {daisu_ke,kam,fuketa,aoe}@is.tokushima-u.ac.jp

本稿では，文書中から数量情報を抽出し，単語と組み合わせて検索するための，数量情報の数値化手法について述べる．文書中の数量情報を抽出し登録することで，数量による文書内容の検索や絞り込みができる．本方式では，数量の係り先を特定しなくても，単語条件との単純な組み合わせで数量条件による絞込み効果を実現する事が目的である．一方，一般の文章中の数量表現には，概数表現や範囲表現が含まれ，表現もさまざまである．本方式では，数量を修飾する表現を５種類に分類して，抽出した数量と共に利用することで検索もれを防ぐようにする．

Research on the intellectual search using quantity information

Daisuke KUWATA, Masao FUKETA, Kazuhiro MORITA, Junichi AOE
Master's Course in Graduate School of Engineering,
Department of information science and Intelligent Systems,
The University of Tokushima

Abstract –This paper describes technique for converting numeric information based on numerical values extraction from text. Numerical values in text are useful to decrease unnecessary retrieval results. To select proper text, our method doesn't use sentence structure analysis but uses simple combination of keywords and numeric conditions. On the other hand, numerical expressions have various modifiers which show numerical range or approximate values. Authors classify such modifiers into five types, and translate them into suitable numeric ranges.

1. はじめに

電子化された文書の量は増大する一方であり，それに伴って，ユーザが求める情報を適切に選び出す技術がますます重要になっている．現在の検索システムは，単語の検索を中心として発達してきた[１]．また，類似文書検索などは，単語による検索の精度を高める，あるいは，容易にする効果を挙げている．

また，現在の検索システムにおいて，数量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量が表わす対象を判断し，数量を含む文脈での検索を実現している．

しかし，文書から数量部分を取り出すだけでは，もとの意味が失われて検索漏れを起こす場合があるので，数量部分だけでなく，範囲内の数量や概数の前後の数量も検索できることが望ましい．

本稿では，数量の検索を有効的におこなうために，従来の検索システムのように単純に単語だけを指定するのではなく，単語と数量条件を組み合わせて検索をおこなう方式を提案する．

2. 数表現の形式

本章では，数表現の特徴と使用方法について述べる．数表現には以下の４つの使用方法が

１） 一般的形式

数表現には，構成要素として以下の５つがある[４]．

① 前置助数詞

前置助数詞は数値の前に置かれるもので，数値の意味をより規定し，助数詞の意味をより限定するものであり，また，順序を示すものもある．

（例）

平均１００人，最高２，０００人，第五回

“平均”，“最高”，“最低”，“約”，“およそ”等は数量の持つ意味をより規定するものである．前置助数詞と助数詞の間には共起関係がある．また，“第”は序数詞を示すものとしてしばしば用いられる．

② 数量

数量は整数，実数（小数点表示，べき乗表示），分数表示がある．文の中の数量は正の数がほとんどであるが，負の数を表わす場合もある．

数量の幅を示すために“から”，“，”，“～”，“－”，“より”等を使用して二つの数量が表われる．

（例）

０．６５mm，７．７%，約５，０００円，
４００～５００件

③　助数詞又は単位

助数詞，単位は多くの数表現の後に出現する．

（例）

回，円，個，枚

助数詞は数の性質を規定し，数表現で重要な役割を果たす．また，助数詞により数の取りうる範囲が限定されることがある．

そして，対象物によって助数詞の表現が変化する．例えば，大きい動物であれば“頭”，小さい動物であれば“匹”，“羽”となる．

④　単位，概数表示

計測単位，概数表示の漢字が使われる．

（例）

３千万円，４千冊

この場合の“千万”，“千”は数量を表わすものである．

（例）

３０数年，２００余社

この場合の“数”，“余”は概数を示すものである．

⑤　程度，順序を示す名詞

程度を表示するものとしては，“以上”，“以下”，“以内”，“以前”等が挙げられる．順序を表わすものとしては，“目”，“等”，“位”．“番“等が挙げられる．これらは，助数詞との結びつきが強いため，同列にして取り扱うことを考える．

２）年月日の表示

年月日の表示には西暦と年号の二通りの方法がある．これは年，月，日と表示項目の順序が一定している．しかし，月，日や年月というように省略する場合や，西暦の表示で“’８０年”というように省略して表わす場合もある．

年月日の表示は文章の中でもしばしばみられ，数表現の中で重要な部分を占めている．

３）複数の項目表示を一定順序で表示するもの

助数詞が複数個使われ，しかも表示順序が一定の順序に従うものである．これらは出現頻度があまり多くないが，特徴のある表示方法である．

（例）

５丁目４５番地，１割２分６厘

４）その他の記述形式

上記以外のもので特殊な数字の使われ方がいくつかある．

①　連体詞と数詞との結びつき

（例）

その１　，その２

②　タイトル，サブタイトルに使われる

抄録文の中には箇条書きのため，その項目番号を表示することがある．

（例）

２．１　言語と文法，４）効果について，（５）予算と決算

③　商品名，会社の一部として使われる

（例）

UNIVAC　１１０８電子計算機，セントロン８２０５

④　さまざまな表示

文中の数式や，省略語等の特別な分類として扱えない場合がある．

（例）

２x＋　２　＞　０

これらは，出現頻度は少ないが種々の場合がある．

3．システム概要

3.1．処理の流れ

本稿で提案するシステムの構成を図１に示す．数量情報抽出部では，文書中の数量情報を，検索時の数量条件と同様に，数量，単位，修飾語の分類の組み合わせで抽出する．数量抽出の際には，概数や範囲表現などできるだけ多くの種類の数量表現を正確に把握することで検索の漏れを減らすようにした．抽出手順を以下に示す．

Step1．形態素解析

与えられた文章を解析し，形態素・語のならびに分解する．そして，それぞれの形態素・語の品詞などを決定する

step2．数量情報の抽出

数量情報抽出ルールを文書に照合する．そして，数量，単位，数量修飾語を抽出し，数量を表わす文字列を数値に変換する

Step3．数量情報の換算

概数や範囲表現を分類に応じて数量の範囲に換算し，数値が表わす範囲を決定する

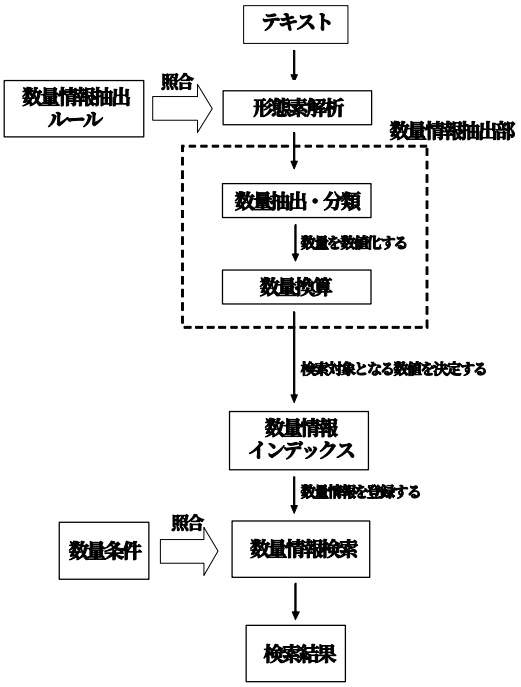

図１： 数量検索方式の構成

Step4．数量情報の登録

抽出した数量情報，数値が表わす範囲を数量情報インデックスに登録する

数量情報検索部では，ユーザが指定した数量条件と，文書中から抽出されインデックスに登録された数量情報を比較して，範囲が一部でも重なるものを検索結果として選択する．

3.2．数量情報抽出ルール

検索文から正確な数量情報を抽出するには，数量だけではなく，前後の単位・数量修飾語などの組み合わせも同時に抽出しておく必要がある．そこで，本手法では，数量・前後の単位・数量修飾語の組み合わせを抽出する数量情報抽出ルールを作成し，数量情報を抽出する．例えば，｛“上限指定”＋“数量”＋“接尾助数詞”｝というルールを作成すれば，「最大５００円のプレゼント」という数量表現から「最大５００円」を抽出することができる．「最大」，「５００円」の組み合わせを抽出することで，数量（この場合では「５００円」）のみを抽出するよりも正確な数量情報を得ることができる．

4． 数量情報抽出・変換処理

本章では，文書から，数量情報を取り出し，数値に変換する処理の詳細を述べる．前述のように，品詞をもとに数量を抽出し，次に，前後の単位，修飾表現を抽出する．以下，個々の抽出処理を説明する．

4.1．数量

形態素解析結果から各形態素の品詞を調べ，数量の文字列を抽出して，大小や範囲の演算ができる数量型に変換する．

数量の表記には算用数字・漢数字の組み合わせ方や，分数や指数，「最低５００円」のように句や文全体で数量を表現する場合など多くの形式がある．

本方式では，算用数字・漢数字のバリエーションを問わず，単独あるいは範囲表現の数量を抽出対象とする．

4.2．単位

数詞の直前・直後にある単位を，品詞をもとに抽出する．抽出する単位の制限はないが，数量検索の対象としては，金額や具体的なものの大きさの需要が大きいと考え，通貨の単位と長さ・重さ・量とする．

4.3．範囲表現

範囲を表わす記号の直前・直後が数量表現のときは，範囲表現とする．「から」，「，」，「～」，「―」を抽出対象としている．

数量は単位付きで抽出するのを原則としているが，範囲の開始を表わす数量に単位が付いていないときは，終了の数量に付いている単位を用いる．

（例）

検索文 ： １万～２万円以内でプレゼントを用意する．

この例では，「１万」には単位が付いていないが，「２万」に「円」が用いられているので，「１万」にも「円」を適用し「１万円～２万円」という表現として扱われる．

4.4．修飾表現

数量を一つ指定して範囲を表現する場合，（ｉ）数量の前後に拡げる，（ii）数量より大きな方へ拡げる，（iii）数量より小さな方へ拡げる，の３通りがある．本方式では，（ii）と（iii）については，拡げる度合いの多いものと少ないものに分割し，５通りに分類する．「下限（概数）」は，範囲が指定値より大きな方へ拡がるが，上限があまり指定値から離れていないものを表わし，「上限（概数）」は，範囲が指定値より小さな方へ拡がるが，下限があまり指定値から離れていないものを表わす．

そして，日本語の教科書，文書，新聞などから数量に付く修飾語を収集して分類をおこなった．WWW ページと新聞記事による評価によれば，数量の修飾表現の９５％を網羅している[５，６]．

表 1： 数量表現の分類

分類	修飾語
概数の表現	前後，程度，くらい，程，約， およそ，ほぼ，大体，約
下限指定	以上，最低，最小，少なくとも
下限（概数）	強，余り
上限指定	以下，以内，未満 最大，最高，最長，多くとも
上限（概数）	弱，近く

表 2： 実験結果

	文書数	正解数	正解率（%）
概数の表現	40	38	95
下限指定	32	27	84.4
下限（概数）	8	8	100
上限指定	55	49	89.1
上限（概数）	7	7	100
修飾表現を含まない	18	12	66.7
総合	150	131	87.3

今回，収集した表現のうち使用頻度の多いものを抽出対象として登録した（表１参照）．

5． 実験・考察

5.1． 実験

検索文中に含まれる数量をどれだけ正しく変換できたかを確認するために評価実験をおこなった．実験データとして Web 上から収集した数量情報を含む文１５０文を用いた．実験結果は表２の通りである．

5.2． 考察

不正解となった原因として，組み合わせの考慮不足による誤解析が挙げられる．

（例）
最高で３０００円のボーナスをくれる．

この例では，数量が「３０００円」，修飾語が「最高」であり．本来ならば，｛“上限指定”＋“数量”＋“接尾助数詞”｝の数量情報抽出ルールで抽出することができる．しかし，「最高」と「３０００円」の間に「で（格助詞）」が含まれているので抽出することができなかった．これは助詞・助動詞・副助詞などを考慮して数量情報抽出ルールを詳細にすることで解決できると考えられる．

また，「１万円の５０%」のように未登録の修飾語が現れた場合，正しく変換できていなかった．これは数量情報抽出ルールを修正し，その修飾語に適した数量変換方法を追加することで正しく数値変換できると考えられる．

6． おわりに

本稿では，数量の条件指定と検索語とを組み合わせた検索手法を提案した．

本方式では，数量を修飾する表現を５種類に分類し，抽出した数量と共に利用することで検索漏れを防ぐようにした．

今後は，検索部分の構築や WWW ページなど各種の文書で評価すると共に，単位表記の拡充や数量の修飾語の扱いの精密化などの改良をおこなう予定である．また，表形式のように単位が数字と離れて記述されている場合への対応や，近接演算との組み合わせなど，多くの選択肢を用意することで，さらに有効な絞り込みを実現できるようにする．

【参考文献】

[1] 福本，加藤，”Question and answering タスクの提案”，言語処理学会研究報告 ２００１-ＦＬ-６３-4，言語処理学会，２００１

[2] 岸本，須之内，塚田，千葉，石川，“テキストの構造化に基づく検索システム”，情処論文誌，Vol.35，No，５，１９９４

[3] 斉藤，迫田，中江，岩井，田村，中川，“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] 山田，福島，“インターネット多角的システム OTROS－数値情報の抽出と検索–”情処５７回大会，3L－０２

[5] 山田，松田，竹元，赤峯，福島，“インターネット多角的システム OTROS－全体の概要と構成”，情処５７回大会，3L－０１

[6] 田中，“日本語の数表現の解析”，情報処理学会計算言語学研究会，ｐｐ.２４－３，１９８０