# インターネット多角的検索システムOTROS
## －数値情報の抽出と検索－

3L－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

数値に関する記述は,多くのテキストで頻繁に使われており,内容にも強く関わっている。テキストを探す際に,製品の値段や大きさなどを知りたいとか,料金の条件で結果を絞り込みたいなど,数値を指定することでより正確な検索が行える.

テキストの内容を検索する場合は,文字や単語によるインデックスが使われる.しかし,数値を以上・以下や範囲で指定するためには,文字列のままではなく数値に変換する必要がある.文章中の数値情報を抽出して,数値化して検索するシステムの従来研究[2, 3]では,文章構造の解析や登録・検索のための構造化が中心となり,数値に付随する表現の違いによる変化については,少数の表現に対応しているだけであった.しかし,文章中で使われる数値の表現は,範囲の表現方法やおおよその数の表現など非常に多彩であるため,正しく数値情報を解釈するには,前後の修飾語も含めた分析が必要である.

筆者らは,文章中で使われる数値表現の違いを,検索の際の条件に反映できる数値検索方式を開発し,多角的検索システムOTROSに実装した.本方式では,範囲の表現や曖昧な数の表現を,その表現が表す数値の範囲の広さで分類する.これにより,単純な数値だけでなく,概数や範囲表現も検索対象にできる.

本稿では,数値表現の分類と,それを利用した検索方式について述べる.また,実際にテキストから数値を抜き出す際の精度について報告する.

### 2　数値検索機能

テキストを数値で検索する場面として,ふたつが考えられる.

ひとつは,数値そのものが知りたい場合である.たとえば,ある国の人口を知りたいので,それが書いてあるテキストを探すという場合である.この場合は,求める数値がテキスト中にあるかどうかの判定が重要で,数値そのものの解釈は必ずしも必要ではない.

もうひとつは,数値による演算を用いて検索結果を絞り込む場合であり,提案方式はこちらを対象としている.たとえば,コンピュータの新製品を紹介したテキストのうち,定価が20万円以下のものについてだけ検索したいという場合である.この場合は,テキスト中で表現されている数値がいくつであるかを判別する必要がある.

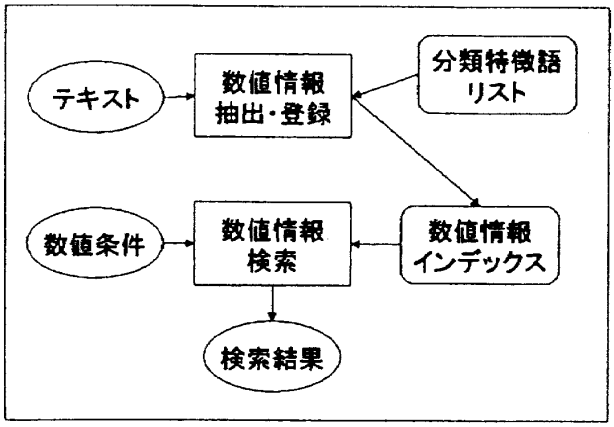


図1: 数値検索システム

提案方式による数値検索の概要を図1に示す.検索対象テキスト中の数字列を,数値表現の分類を特徴づける語(分類特徴語)のリストとマッチングして分類し,数値および単位とともに数値情報インデックスに登録する。分類特徴語リストには,各特徴語ごとに,数値,単位との位置関係が定義されている.

検索時には,ユーザが数値条件と検索語の組み合わせ(下の例)を入力する.

検索語:　パソコン AND モバイル
数値:２０万　単位:円　範囲:以下

数値条件は数値,単位,範囲の組み合わせで指定し,内部で式に変換した上で数値情報インデックスとマッチングを行う.

### 3　数値表現の分類

提案方式では,数値表現を範囲の指定方法や曖昧さによって分類し,それぞれについて検索時の扱いを変更する.各分類は,数値に付随する分類特徴語によって特徴づけられる.数値表現の分類は以下のとおりである。

---

Internet Diversity Retrieval System OTROS
– Search Functions for Number Expression –
Hiroshi Yamada and Toshikazu Fukushima
Human Media Research Labs. , NEC Corp.

表 1: 分類特徴語の数と例

| | | |
|---|---|---|
| 範囲表現 | 6 | から, ～, － |
| 上限指定 | 9 | まで, 以下, 未満 |
| 下限指定 | 5 | より, 以上 |
| 曖昧(中央) | 12 | 約, 程度, くらい |
| 曖昧(上限) | 7 | 弱, 近い, 足らず |
| 曖昧(下限) | 18 | 強, 余り, 少し |
| 合計 | 57 | |

表 2: 分類特徴語の評価結果

| | 新聞 | WWW | 合計 | |
|---|---|---|---|---|
| 数値表現の総数 | 232 | 381 | 613 | 100% |
| 単独の数値 | 177 | 325 | 502 | 81.8% |
| 範囲・曖昧表現 | 55 | 56 | 111 | 18.2% |
| 収集済パタン | 53 | 52 | 105 | 17.2% |
| 未収集パタン | 2 | 4 | 6 | 1.0% |

**単独の数値** 単一の数値で表現しているもの. 「20万円」,「100メートル」,「3年」など. そのままインデックスに登録する.

**数値の範囲** ふたつの数値表現で範囲を表現しているもの. 「10万から20万円」,「150～170人」,「500グラム以上800グラム以下」など. 数値の下限と上限を登録する.

**上限の指定** 数値の上限だけを記述しているもの. 「10万円以上」,「100人より多い」など. 数値と分類を登録する(以下同様).

**下限の指定** 数値の下限だけを記述しているもの. 「10万円以上」,「100人未満」など.

**中央を指定した曖昧表現** 数値の前後にある程度の幅があるもの. 「約200円」,「50人程度」など. 検索時には,前後に一定の幅を持った範囲として扱う(たとえば±10%).

**上限を指定した曖昧表現** 上限のみが表現されているが,ある程度の下限が想定できるもの. 「100人弱」,「1万円に近い」など. 下限は明言されていないが,数値からかけ離れた値ではないと推定される. 検索時には,少ない側に一定の幅を持った範囲として扱う(たとえば-10%).

**下限を指定した曖昧表現** 下限のみが表現されているが,ある程度の上限が想定できるもの. 「100人余り」,「1万円を越える」など. 検索時には,多い側に一定の幅を持った範囲として扱う(たとえば+10%).

特殊な例として,「代,台」を用いた表現がある. これは,上位の桁が共通な数値の範囲として扱う. 「30歳代」→「30歳～39歳」,「2万円台」→「20,000円～29,999円」.

提案方式をOTROSシステムに実装するために,新聞記事,WWWページ,文章表現の参考書などから57種類の分類特徴語を収集した(表1).

## 4 収集した数値表現の評価

提案方式の有効性を見積もるため,数値情報の抽出対象となる表現のテキスト中ででの出現数と,収集した分類特徴語の数が十分かを評価した.

収集用とは別の新聞記事,WWWページ各20テキストを用いて,数値の出現数と,そのうち収集した分類特徴語をともなう割合とを調査した(表2).

この結果から数値表現が1テキストあたり平均15か所と頻繁に使用されていることがわかる. その中では単独の数値が多いが,範囲や曖昧表現なども18%あり無視できない. 収集した分類特徴語はそのうちの95%をカバーしている. カバーできなかった表現は数値表現全体の1%とわずかである.

未収集の分類特徴語には,「○○規模」,「○○以降」など,通常のテキストを大量に調査することで集められそうなものが多く,現状から大きく増やさなくても大多数の場合に対応できると予想している.

## 5 おわりに

テキスト中の数値表現を検索するために,範囲や曖昧な数値をあらわす表現を分類し,分類に応じて検索条件を変える検索方法を開発した.

数値表現を7種類に分類し,それぞれの分類を特徴づける分類特徴語を,新聞記事などから57語収集した. 実際のテキストによる調査では,収集した分類特徴語で95%を網羅しており,収集を続けることでさらに多くの表現をカバーできるようになる.

今後の課題としては,曖昧な表現を数値の範囲に換算するパラメータと検索精度との関係の検討と,より多くのテキストによる数値表現の分析と網羅率の調査がある. また,より高度な検索に利用するために文章解析と組み合わせて,数値が示す対象と組み合わせた抽出を行う.

## 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-全体の概要と構成-", 情処57回大会,3L-01

[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