概念情報検索方式の提案

1C－1

小澤　英昭　　中川　透
ＮＴＴ　ヒューマンインタフェース研究所

1　はじめに

情報検索は、計算機利用の一分野として発展してきた。最も広く用いられている手法は、キーワード検索である。キーワード検索では、情報に対するキーワードの付与者のキーワードに対する概念と、検索者の持つキーワードに対する概念が一致しないと、効率の良い検索は行なえない。

本稿では人間の情報に対する概念的なばらつきを測定することで、個人により感じ方に差のある情報の分類を行ない、キーワードで扱いにくい情報に対する検索方式と評価法を提案する。

2　検索手法のモデル化

検索を行なう際に、利用者が求める情報は、検索の目的や、利用者の情報に対する感じ方によって異なる。利用者の情報に対する感じ方は、個々の情報が持つ内容を解釈することによって決定される。一方で情報検索では、例えばキーワードは情報の持つ特徴を記号として表現することにより、情報の内容を表現しようとしている。つまり情報検索とは、個々の情報の持つ内容に対して、利用者側の解釈と、システム側からの情報の特徴付けにはさまれた３段階の階層として、上位の層から順に次のようにモデルで表現できる。

１．感性的な関連　（人間の判断による分類）
２．概念的な関連　(内容による分類)
３．記号的な関連　（キーワードなどによる分類）

このモデルにおいて検索を効率良く行なうためには、次の２つの条件を備えていることが望ましい。

１．　検索された情報間での関係に対して、全ての人が同じ感じ方をする様に、情報を分類できる事
２.情報検索のアルゴリズムで、上記の分類を表現できること

しかし一般的に、この条件を満たすことは難しい。例えば新聞記事のデータベースの場合に、キーワードによって何らかの分類をすることはできるが、その分類は、必ずしも利用者の感じ方と一致しない。個々の新聞記事から得られる感じ方は、人それぞれによって異なり、全ての人が同じように感じる情報の分類というのは困難である。

3　人間の関連性に対する感じ方と検索法

人間の情報に対する感じ方のばら付きの有無を測定するために、以下のような実験を行なった。

実験内容：　新聞記事から記事を一つ選び出し、あらかじめ用意した別の新聞記事の集合中の各記事に対して類似する程度を記述させた。

類似の程度：４段階（関連がある　：４、関連がややある：３、関連が余りない：２、関連がない：１）

被験者の数：　　一つの設問あたり９人

設問の数：　　１２問/人

新聞記事の集合：　設問記事と同時期に記載された記事からランダムに選んだ記事群(記事数５０個)

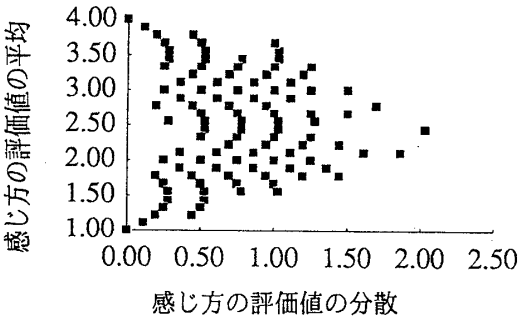

図１．　情報に対する人間の感じかた

結果は図１に示すように、人間が記事間の関係を判定する場合でも、情報により関連の有無の判断に

A Conceptualistic Information Retrieval Method
Hideaki Ozawa, Toru nakagawa
NTT Human Interface Laboratories
1-2356 Take, Yokosuka, Kanagawa 238-03, Japan

実験には、朝日新聞の記事データを用いた。

は、かなりのばらつきが生じている。つまり新聞記事のような情報を検索する場合には、多くの利用者が期待する情報と、一部の利用者が期待する情報が混在することが予想される。従来の情報検索の検索法や、検索評価法では、この人間の情報に対する感じ方の差は、ほとんど考慮されていない。

そこで感じ方の差が大きい情報に対しては、感じ方の違いを生じさせている、情報の概念的なレベルでの検索の手法が必要である。この概念レベルでの検索の手法と、次の条件を満たす手法として定義し、これを概念検索法と名付ける。

1．人間の感じ方の差が少なくなるように、検索を行なう際に、概念的なレベルでの特徴を捉える検索法

2. 人間の判断に差が少ない情報は、人間の判断と一致する場合と、人間の判断にばらつきがある場合には、検索結果に対する評価の曖昧さを許容する検索評価法

特に評価法は、人によって判断の異なる関連性を持つ情報の価値は低く、判断のばら付きの小さい関連性を持つ情報は価値が高くなる。この考え方に基づく検索効率の評価法として、本稿では、追随率という指標を提案し図２のように定義する。

$$f=\begin{cases} 0 & (x - V <= Xs <= x+V) \\ 1-x-V & (Xs > x + V) \\ x-V & (Xs < x -V) \end{cases}$$

X = {0, 1 | 0:関連性無、1:関連性有}

$x = \Sigma Xp/n$ 、 $V = \sqrt{\Sigma (Xp - x)^2/n(n-1)}$

n: 被験者の数、

Xp 被験者の与えた関連性の有無の値

Xs 検索アルゴリズムによる値

$$f_{ave} = \Sigma\Sigma f_{i,j} /nm \quad (i = 1,n \ : j = 1,m)$$

(m: データベース中のデータ数)

図２． 追随率の定義

この指標では、人間が行なった関連性の判断の有無のばらつきが小さいデータに対して、システムが判断を誤ると、fに大きな値が与えられ、ばらつきが大きい場合には、システムが判断を誤ってもfに小さな値が与えられる。理想的な検索手法の場合には、検索結果がばらつきの範囲内に納まるので、平均追随率 f_{ave} が０になる。この評価法は、再現率や適合率が関連性のある情報が検索されたことが評価の対象であるのに対し、追随率では関連性の無い情報が検索されないことも評価の対象になっている。

4 検索手法と追随率

新聞記事の多くを占める時事情報は、政治、経済社会情報のいずれを問わず、人間が地球上で何かの動作を行なうことによって生じる。誰が何をしているかは、記事中で利用されている、固有名詞やサ変名詞によって表現されている。そこで本稿では、概念的なレベルでの特徴として、使用されている固有名詞や、サ変名詞の一致量から、関連性を決定する検索アルゴリズムを作成した。作成したアルゴリズムと、新聞記事に付与されているキーワードとにより、検索効率を比較して概念レベルでの検索法の有効性を検討した。

実験としては、新聞記事に付与されているキーワードの全てを用いてのＡＮＤ検索、ＯＲ検索と、上記で作成したアルゴリズムを、提案する追随率で比較した。

	平均追随率
ＡＮＤ検索	0.16
ＯＲ検索	0.35
提案する手法	0.09

図３. 検索効率の評価

結果から見られるように、個々の記事に付与されているキーワードのAND、ＯＲ検索のいずれの結果に比べて、事件を起こす人や、その動作による概念レベルでの特徴に基づく検索法の方が、人間の情報に対する感じ方の結果に、より合致している事がわかる。

5 おわりに

本稿では、情報間の関係に曖昧性が大きいデータについて、人間による関連性の強さの判定を基盤とした検索の手法と、その評価法を提案した。