情報検索における同義語展開の有効性の評価

3G−3

下山栄子，松井くにお，富士秀

(株)富士通研究所

1. はじめに

情報検索システムにおいて、キーワードのヒット率を高め検索率を上げる手段の一つとして、各種変換テーブルを利用して行う「キーワード展開」がある。キーワード展開は、大きくとらえると、適合情報を洩れなく検索するための「ｏｒ展開」、不適合情報の検索を避けるための「ａｎｄ展開」に二分できる。それぞれいくつかの方法が考えられ効果が期待されるが、一般に膨大なデータベースを検索対象としているため再現率の算出が容易でなく、有効性が確認しにくかった。本稿では、ｏｒ展開の一つである「同義語展開」のシミュレーションを行い、その有効性を定量的に評価した実験結果を報告する。

2. 検索効率の評価方法

2.1 再現率と適合率

情報検索の検索効率の評価には、一般的に図1のような範囲で示される再現率と適合率が用いられる。

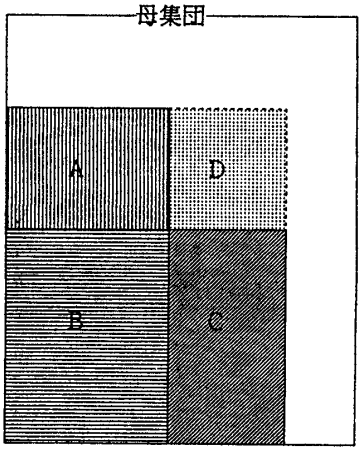

図1　母集団と適合情報

この図で、

A＋B：検索された情報の全体
B＋C：要求に適合する情報の全体
B　　：検索された適合情報

とすると、

再現率　$R_1 = B/(B+C)$
適合率　$P_1 = B/(A+B)$

で表すことができる。

ここで、AとBは検索された結果であるため測定可能であるが、Cは膨大なデータベース中に埋もれてしまうため測定不能であり、一般的には再現率は算出不可能となる。

2.2 検索効率の算出方法

われわれは、同義語展開の有効性の評価に、以下の仮定を導入することにした。

【仮定】　理想的な同義語展開を行ったキーワード検索結果は、適合情報のすべてを含むものとする。

この仮定の下では、同義語展開後の再現率・適合率は以下のように計算できる。

再現率　$R_2 = (B+C)/(B+C) = 1$
適合率　$P_2 = (B+C)/(A+B+C+D)$

ここで、CとDは、前回の差分をとって測定が可能となるため、展開前の再現率R_1の算出も可能となる。そこで、検索率を再現率と適合率の積と考えると、同義語展開の有効率は以下のように算出できる。

$$\text{有効率}\quad E = \frac{\text{展開後の検索率}}{\text{展開前の検索率}} = \frac{\overline{R_2 P_2}}{\overline{R_1 P_1}}$$

$$= \frac{\text{ave.}\{(B+C)^2\ (A+B)\}}{\text{ave.}\{B^2\ (A+B+C+D)\}}$$

3. 同義語展開のシミュレーション

3.1 実験方法

シミュレーションには特許文の検索システムを利用した。2.2 に示した方法で検索の有効性の評価を行うには、検索結果を内容の適否までチェックする必要があるため、始めに適当な量の母集団を定めた。次にサンプルキーワードを決め、それぞれ(1)(2)の方法で検索を行い、(1)の結果を基準として検索結果の変化を調べた。

Effectiveness of Using Synonyms of Keywords
Eiko SHIMOYAMA, Kunio MATSUI, Masaru FUJI
FUJITSU LABORATORIES LTD.

(1) 同義語展開しないで検索
(2) 同義語展開して検索

(2)については、利用した検索システムが実験に適当な同義語ファイルを持たないので、機械翻訳用の大語彙辞書から作成した別の辞書システムで同義語を引き、検索式に展開する方法をとった。この辞書システムから得られる同義語は意味ごとに分かれているので、意味の適合するグループに含まれる全ての同義語のorをとって1つの検索式を作るようにした。

3.2 実験結果

表1は、「キーワード AND 辞書」で得られた特許文116件を母集団として3.1の方法でキーワード検索した結果である。
(2)の空欄は(1)との差分のなかったもので、そのうち「設定」「付与」の2つのキーワードでは、同義語展開しても検索で得られる情報が増えなかった。また、「連想」「内容」の2つのキーワードについては、辞書システムから適当な同義語が得られず、展開できなかった。

4. 同義語展開の有効性

サンプルキーワード12語についてシミュレーションした結果、多くの同義語展開が行われ、全体として検索件数が62件から115件へとほぼ倍増した。検索された全件について内容の適否を調べると、展開後の増分53件中11件が適当と判断され、展開なしの場合の取りこぼしを同義語展開で救ったことになる。これについて今回の計算式を利用して数値的に評価すれば、再現率が55.9% から100%に増加したと言える。ある程度適当なキーワード展開を行えば再現率が上がるのは当然であるが、ここで問題なのは、適合率を下げないことである。実験では、展開前の適合率22.6%に対し、展開後の適合率は 21.7%で、同義語展開による適合率の低下はほとんど無かったと言ってよいだろう。

以上の評価の結果、今後さらに検索率を向上させていくためには不適当な検索を取り除く工夫が大いに必要ではあるが、実験に用いた辞書システムに基づく同義語展開は検索率の向上に有効であるという結論を得た。

5. おわりに

今回の実験では、or展開の一つである同義語展開について有効性を評価した。今後は、他のor展開（異表記、対訳、構成語への展開など）について同様のシミュレーションを行ない、今回提案した方法でそれぞれの有効性を評価していく予定である。

評価の結果、有効性の確認された展開方法は検索システムに取り込んでいきたいと考えているが、検索速度など、有効率以外の要素も考慮してシステム全体を考えた場合、全ての方法を100%取り込めないことも十分考えられる。各方法の有効性を数値化した評価結果は、そのような場合に優先順位をつける指標として利用することもできよう。

表1 同義語展開のシミュレーション結果

母集団： 「キーワード AND 辞書」 116件

展開	(1) なし		(2) あり		
キーワード	A	B	D	C	同義語（一部）
設定	9	1			確立、構える、築き上げる、…
付与	6	1			付与、授与、差し上げる、…
付加	2	1	4	1	追加、添付、附加
複合語	0	4	0	1	混合語、合成語、連語、…
展開	2	0	1	0	拡大、広がる、伸張、発展、…
関連	6	3	6	4	関係
連想	3	0			
内容	13	0			
同義語	2	1	3	2	代名詞、シノニム、同義、…
類義語	2	1	3	1	代名詞、類語、…
類語	1	0	4	2	代名詞、類義語、シノニム、…
分野	2	2	21	0	フィールド、関係、領域
合計	48	14	42	11	