4U-07

検索結果の適合度の定量的評価

藤崎 博也[1] 阿部 賢司[1] 飯島 岐勇[1] 武田和也[1] 大野 澄雄[2]
[1] 東京理科大学 [2] 東京工科大学

1. はじめに

通常の情報検索において、一般のユーザは異表記同義や同表記異義[1]の存在を意識していないため、ユーザの作成した検索式に基づいて検索する場合、検索もれや不要な検索が生じる。この問題に対し、我々は検索もれを低減することを最優先とし、“キー概念”[2]に基づき拡張した全てのキーワードの論理和を用いて検索する方法を採用した。しかし、この方法では不要な情報まで抽出することが多いため、ユーザに提示する際には、必要な情報のみを選び出す必要がある。

本稿では、必要な情報を自動的に推定するために、人間による適合度の評価値と高い相関のある評価値を推定する方法について検討する。

2. 検索式の拡張と検索結果の人間による評価

まず、検索式として、ユーザの生成したもの (A)、その中のキーワードをキー概念のレベルで拡張したもの (B)、拡張した全てのキーワードの論理和 (C)、の3種を設定した。なお、以下では拡張した後の全てのキーワードを単にキーワードと呼ぶこととする。

つぎに、これらの検索式を用いて検索要求30件に対する検索実験を行なった。検索対象は情報検索システム評価用テストコレクション構築プロジェクト[3]により提供されている大規模なデータベースから無作為に抽出した論文データ5425件とした。各論文データには、題目、著者名、所属、概要、重要語句などが含まれる。ここで重要語句とは論文のキーワードリストを指す。

検索もれを軽減することを最優先とした場合には、Cを用いることが最も効果的であることは明らかであるが、その反面、不要な情報も多数抽出される。これを具体的に調べるため、人間がCによる検索結果の適合度を「非常に関係が深い (5)」「やや関係している (4)」「どちらとも言えない (3)」「おそらく関係ない (2)」「全く関係ない (1)」の5段階で評価した結果、内訳は (5) が 7.0%、(4) が 9.6%、(3) が 16.2%、(2) が 32.3%、(1) が 34.9%であった。ここで適合度 (5), (4) を与えられた論文が必要な情報である。したがって、人間による評価値と高い相関のある推定値を与える推定法があれば、自動的に必要な情報が得られる。

Quantitative estimation of relevance of retrieved results
Hiroya Fujisaki[1], Kenji Abe[1], Michio Iijima[1], Kazuya Taketa[1] and Sumio Ohno[2]
[1]Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510
[2]Tokyo University of Technology, 1404-1 Katakura, Hachioji, 192-0982

3. 出現回数に着目した適合度の推定

なるべく多くの種類のキーワードがなるべく多く出現している文書は適合度が高いと期待される。

一つのキーワード W_iが文書 D の適合度の推定値に対して及ぼす寄与を $r_D(W_i)$ とし、すべてのキーワードの寄与の和を、文書 D の適合度の推定値 R_D とする。

$$R_D = \sum_{i=1}^{n} r_D(W_i) \qquad (n \text{ はキーワードの総数})$$

ここで、$r_D(W_i)$ はキーワード W_iの文書 D における出現回数 N_{D,W_i}の関数であり、式 (1) から式 (4) の4種を設定した。また、a, b, c, d はそれぞれの式のパラメータである。なお、簡単のため、以下では W_iの添字 i を省略して単に W と表記する。

$$r_D(W) = \begin{cases} a(N_{D,W}-1)+1, & N_{D,W}>0, \\ 0, & N_{D,W}=0. \end{cases} \qquad (1)$$

$$r_D(W) = \begin{cases} b\,log_e\left(1+\frac{N_{D,W}-1}{b}\right)+1, & N_{D,W}>0, \\ 0, & N_{D,W}=0. \end{cases} \qquad (2)$$

$$r_D(W) = \begin{cases} c\,log_e N_{D,W}+1, & N_{D,W}>0, \\ 0, & N_{D,W}=0. \end{cases} \qquad (3)$$

$$r_D(W) = \begin{cases} \sum_{n=1}^{N_{D,W}} \left(\frac{1}{n^d}\right), & N_{D,W}>0, \\ 0, & N_{D,W}=0. \end{cases} \qquad (4)$$

推定値の妥当性を検討するため、回帰分析により人間による評価値と各方式による推定値との相関係数を求めた結果を図1に示す。4種の関数のいずれにおいても、パラメータの値を適切にとることによって

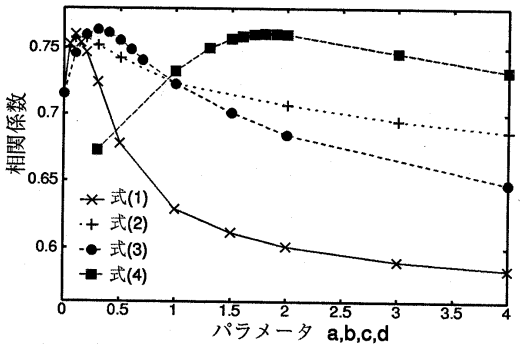

図 1. 出現回数に着目した適合度の推定の比較

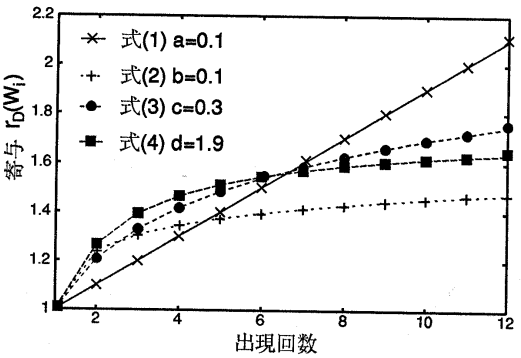

図 2. 相関の高い式の出現回数に対する寄与

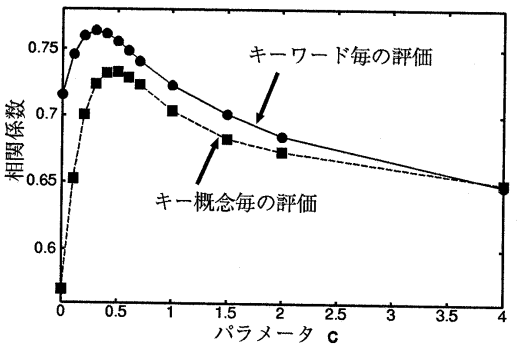

図 3. キーワード毎とキー概念毎の相関係数の比較

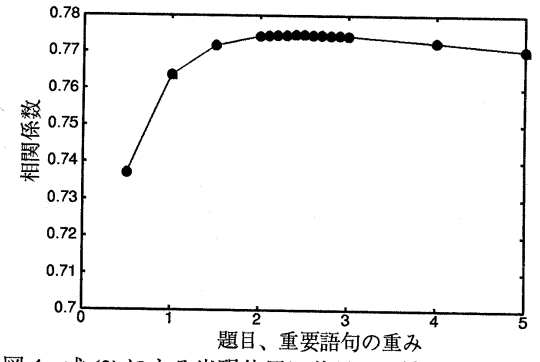

図 4. 式 (3) による出現位置に着目した適合度の推定

ほぼ等しい相関係数の最大値が得られるが、式 (3) の $c = 0.3$ が最大 (0.764) の相関係数を与えており、式 (4) の $d = 1.9$ (0.761)、式 (1) の $a = 0.1$ (0.760)、式 (2) の $b = 0.1$ (0.758) がそれに次ぐ値を示している。各式において最大の相関係数を与える推定法について、それぞれの出現回数に応じた寄与を図 2に示す。

4. キー概念毎の出現回数に着目した適合度の推定

前節では、それぞれのキーワードを別個のものとみなして出現回数を求めたが、ここでは、キー概念毎に出現回数を求め、人間による評価値とその推定値との相関係数を求めた結果を図 3に示した。出現回数をキー概念毎とした場合には、相関係数が僅かに小さくなる。なお、推定法は図 1で最大の相関係数を示した式 (3) を用いた。

5. 出現位置に着目した適合度の推定

題目や重要語句にキーワードが出現している文書は適合度が高いと期待される。ここでは、論文を (題目, 重要語句) と (それ以外) の 2 つに分け、(それ以外) の重みを 1 とし、(題目, 重要語句) の重みを任意に定め、実験を行なった。なお、推定法はこれまでの実験の結果に基づき、キーワード別の出現回数に着目した式 (3) ($c = 0.3$) を用いた。図 4はその実験結果であり、(題目, 重要語句) と (それ以外) の比率が 2.5 : 1 の場合に、推定値と人間による評価値とが最も高い相関をもつことを示している。

6. おわりに

本稿では、検索結果の適合度の推定法について、その推定値と人間による評価値とを相関係数を用いて比較した場合に、高い相関を与える推定法について検討した。今後は、より適切な適合度の推定法を検討する。

参考文献

[1] 劉 軼, 戸井田 和重, 八杉 大輔, 阿部 賢司, 大野 澄雄, 藤崎 博也, 久保村 千明, 亀田 弘之: "学術情報検索における異表記同義・同表記異義の分類・分析および処理," 言語処理学会第 4 回年次大会発表論文集, pp. 108–111 (1998).

[2] 藤崎博也, 亀田弘之, 河井恒: "新聞記事情報の階層構造に基づく記事分類・検索システム," 情報処理学会「自然言語処理」研究会資料 44 - 4 (1984).

[3] http://www.rd.nacsis.ac.jp/ ntcadm/