定量的基準に基づく検索結果の順位付けの検討

1U－2

藤崎 博也　大野 澄雄　阿部 賢司　飯島 岐勇　片見 憲次　鈴木 匡芳

東京理科大学

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と、検索もれや不要な検索が生じる。このような観点から、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] を用いて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それらを取り除く必要がある。本稿では、検索結果に対して適切な順位付けを行ない、必要な情報をユーザに優先的に提示する手法について検討した結果を述べる。

2. 検索結果の分析

まず、検索式として、ユーザが生成したもの (検索式 A)、検索式 A をキー概念に基づいて拡張したもの (検索式 B)、拡張した全てのキーワードの論理和 (検索式 C)、 の 3 種を設定した。

つぎに、これらの検索式を用いて検索要求 10 件に対する検索実験を行なった。検索対象は学術情報センター電子図書館サービス [3] によりテキストデータとして提供される論文概要 5425 件 (1998 年 1 月時点) とした。各テキストデータには、題目、著者名、所属、概要、キーワードなどが含まれる。一例として、検索要求：“日本語に関する構文解析についての論文” に対する検索式 A、検索式 B、検索式 C を以下に示す。また、それらを用いた場合の結果を表 1に示す。

検索式 A　日本語 *and* 構文解析

検索式 B　(日本語 *or* 邦語) *and* (parsing *or* syntactic parsing *or* syntax analyses *or* syntax analysis *or* パージング *or* 構文解析)

検索式 C　日本語 *or* 邦語 *or* parsing *or* syntactic parsing *or* syntax analyses *or* syntax analysis *or* パージング *or* 構文解析

(ここで “邦語” とは、EDR 電子化辞書中で用いられている用語で、外国語の表現を邦訳したものを指す)

表 1　3 種類の検索式による検索結果

検索式	適合文書数	抽出した文書数
検索式 A	2	4
検索式 B	3	5
検索式 C	7	68

各検索式における検索結果とデータベース内の適合文書との関係は、一般に、図 1のようになる。

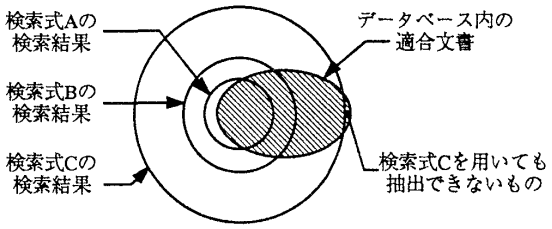

図 1. 適合文書の分布の概念図

すなわち、検索もれを軽減することを最優先とした場合には、検索式 C を用いることが最も効果的であるが、その反面、不要な情報も多数抽出される。したがって、順位付けにより適合文書をユーザに優先的に提示することが望ましい。順位付けの方法を検討するため、まず、適合文書の特徴を分析した。その結果、一般に **(1)** 検索式を構成するキーワードの集合の中の特定の一つのキーワードが頻出する、**(2)** 複数のキーワードが含まれる、**(3)** 論文の題目などにそれらのキーワードが含まれる、といった特徴があった。これらの特徴を考慮し、出現するキーワードの種類と回数とを重視して得点を与える方法 (方法 (1)) と、キーワードの出現位置と回数とを重視して得点を与える方法 (方法 (2)) について検討した。

3. 順位付けの手法

3.1 方法 (1)

まず、文書 d におけるキーワード w の出現回数 $N_{d,w}$ から文書 d の重み $v(d,w)$ を計算する。ここで、$v(d,w)$ は式 (1) で表される。

$$v(d,w) = \sum_{n=1}^{N_{d,w}} \frac{1}{n} \qquad (1)$$

つぎに、検索式中の全てのキーワードの $v(d,w)$ の和を文書 d の得点とし、順位付けに用いる。

A study on the rank-ordering of results in information retrieval based on quantitative criteria

Hiroya Fujisaki, Sumio Ohno, Kenji Abe, Michio Iijima, Kenji Katami, and Masayoshi Suzuki

Science University of Tokyo, 2641 Yamazaki, Noda, 278-8510

3.2 方法 (2)

まず、文書 d における位置 i でのキーワード w の出現回数 $n_i(d,w)$ と、位置 i による得点 p_iから重み $v(d,w)$ を計算する。ここで、位置 i は、題目、著者名、キーワード、その他 (概要等) の 4 通りとする。位置 i による得点 p_iは題目など、重要な位置ほど高くなっている。したがって、$v(d,w)$ は式 (2) で表される。

$$v(d,w)=\sum_{i=1}^{4} n_i(d,w)\cdot p_i \qquad (2)$$

つぎに、検索式中の全てのキーワードの $v(d,w)$ の和を文書 d の得点とし、順位付けに用いる。

3.3 方法 (3)

上記の方法 (1),(2) との比較のため、通常用いられる$tf\cdot idf$法を順位付けに用いる。以下では、これを方法 (3) と呼ぶ。

4. 順位付けの結果の評価

文書群の得点は、最大値を 1 として正規化し、それがある閾値を越えるものをユーザに提示するものとする。一般に、検索結果は検索もれ率 (適切なものを検索しなかった割合) と誤検索率 (検索した結果の中での不適切なものの割合) を用いて評価することができる。したがって、ここではその 2 つに基づいて検討する。いま、再現率を R とすれば検索もれ率は $(1-R)$ で、適合率を P とすれば誤検索率は $(1-P)$ で表わされ、両者の荷重和が検索の失敗に対する定量的な指標となる。

図 2は閾値の 0 から 1.0 までの値に対する検索もれ率を示している。全般にわたり、方法 (1) を用いた場合、検索もれが最も少ない。ついで方法 (2) を用いた場合が少なく、最も悪いのが方法 (3) である。一方、図 3は閾値の 0 から 1.0 までの値に対する誤検索率を示しており、閾値が 0.65 から 1.0 の範囲では、方法 (1) を用いた場合に不適切なものの割合が低く、閾値が 0 から 0.6 の範囲では方法 (1) と (2) の間には大きな差はない。また、閾値が 0 から 0.2 の範囲以外では、方法 (3) は他の 2 つの方法よりも明らかに劣っている。

以上は比率に着目して検討したものであるが、場合によっては比率ではなく、検索件数自体が問題になることがある。図 4は閾値の 0 から 1.0 までの値に対する検索もれの実数と誤検索の実数を示している。それぞれ、検索もれの実数は検索できなかった適切な文書の数、誤検索の実数は検索してしまった不適切な文書の数を表す。この場合には、なんらかの係数による両者の荷重和が検索の失敗の指標となる。

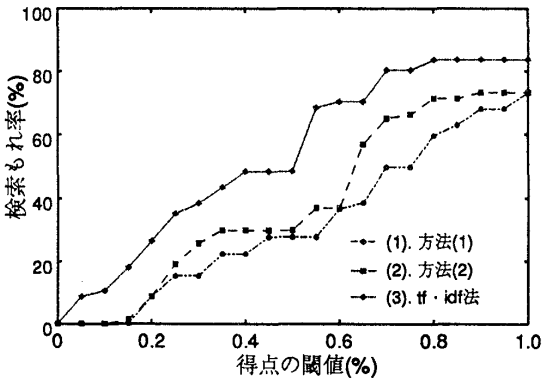

図 2. 各順位付けにおける検索もれ率の比較

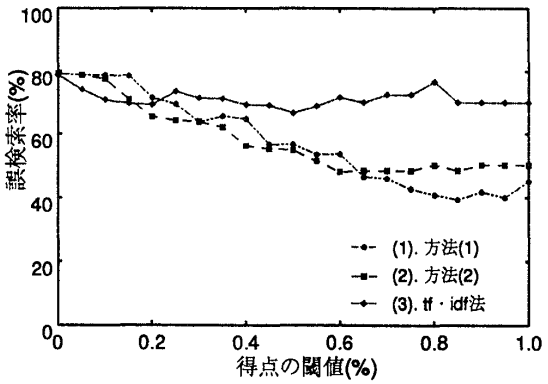

図 3. 各順位付けにおける誤検索率の比較

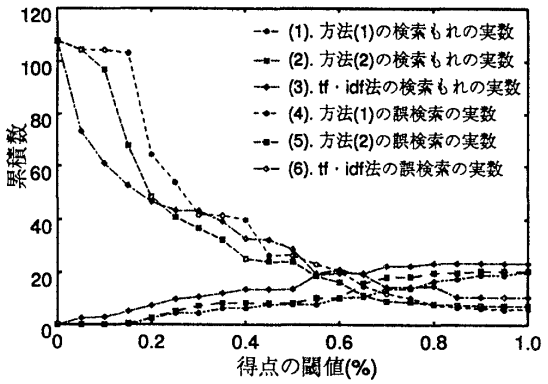

図 4. 実数による評価

5. おわりに

本稿では、検索式を拡張し、検索結果の順位付けにより検索もれが少なく、かつ適切な結果をユーザに提示する手法を述べた。この場合の順位付けにおいては提案手法が有効であることを示した。

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://els.nacsis.ac.jp/nacsis-els-j.html.