情報検索のためのキーワード抽出実験*

3Z－03

武田善行　梅村恭司†
豊橋技術科学大学　情報工学系‡

1. はじめに

本研究では，情報検索のための索引語抽出において，索引語境界の特定の問題と情報検索に有効な索引語の特定の問題を考える．

情報検索では一般に検索質問と文章中の索引語が一致することに基づいて検索を行うが，言語によって索引語ないし語の境界は明確ではない．分かち書きの手段として形態素解析は一般的であるが，原則として処理対象となる文書に対して十分な語彙を持つ辞書が必要であり，そういった辞書を用意するコストは高い．

さらに語は1以上の形態素からなることや，代表的な索引語である専門用語の85%は複合語である事を考えると，出来合の辞書を用いた形態素解析がすべての応用に対して最適な手段とは言い切れない．たとえば複合語や連語の扱いは応用によって異なる．本研究のように情報検索目的であるなら，検索要求に適合する文書を特定する情報の量で扱いを決めるべきである．

また，検索質問中に含まれるすべての語は索引語になり得ない．情報検索の際に不要語を除くことは一般的であるし，検索質問に含まれるそれぞれの語が持つ検索要求を示す量は本来均一ではない．

本研究では，語境界の明確でない言語において自由文以外の知識を一切必要としない索引語の抽出法を使う[1]．また，この索引語の抽出法が使う*adaptation*の特徴に注目した[2]．*adaptation*は語が持つ内容量を示す特徴量として使えることが報告されていることから，抽出した索引語を使うことによって情報検索性能を向上させることができるのではないかと考えた．本研究では，NTCIR2テストコレクションを用いて情報検索性能評価を行い[3]，その結果を分析することによって，文献[1]の索引語抽出の枠組が情報検索の性能向上に有用であることを明らかにする．

2. *adaptation*の定義

多くの語は文書に繰り返し出現する傾向にあり，その度合いを示す特徴量は語の内容量に関わることが報告されている．*adaptation*はある文書に語 w が1回以上含まれている条件で，ある文書に語 w が2回以上含まれる条件付き確率である．文書が語 w を含む事象を $e_1(w)$，文書が語 w を2回以上含む事象を $e_2(w)$ とすると，*adaptation*は次のように定義される．

定義 1

$$adaptation(w) = p(e_2(w) \mid e_1(w)) = p(e_2(w)) / p(e_1(w))$$

また，語 w が文書に出現する確率と，語 w が文書に2回以上出現する確率は，文書集合全体で考えると文書頻度を用いて推定することができる．文書集合全体で語 w を含む文書の数を $df(w)$，語 w を2回以上含む文書の数を $df_2(w)$ とすると，*adaptation*の推定は次のように定義される．

定義 2

$$adaptation(w) = p(e_2(w)) / p(e_1(w)) \approx df_2(w) / df(w)$$

3. 索引語の抽出法

本研究で示す索引語抽出の手続きは2段階である．始めに文字列を索引語らしい文字列の連続に分割し，次に分割された文字列を評価し索引語らしいものを抽出する．分割に使う索引語らしさに*adaptation*を用いたが，*adaptation*は極端な高頻度文字列や低頻度文字列に対して推定が不安定であるため，文書集合からの学習を元に制限を加えた[1]．定義を次に示す．

定義 3

$$Score(w) = \begin{cases} -\infty & \cdots & df_2 < 3 \\ \log 0.5 & \cdots & df_2 \geq 3, df / N > 0.5 \\ \log df_2(w) / df(w) & \cdots & df_2 \geq 3, df / N \leq 0.5 \end{cases}$$

索引語抽出元の文字列において，文字を最小要素と

---

* Experiment of Keyword Extraction for Information Retrieval
† Yoshiyuki Takeda, Kyoji Umemura
‡ Toyohashi University of Technology Dept. of Information and Computer Sciences

するすべての分割に付いて評価し，それぞれの文字列の索引語らしさが最大になる分割を行う．分割された文字列の集合をWとして，定義を次に示す．

定義 4

$$words = \arg\max_{W}\left(\sum_{W=\{w_1,w_2,\cdots,w_n\}} Score(w_i)\right)$$

分割されたそれぞれの文字列に対する索引語らしさには，頻度計数元である文書集合における存在範囲を元に制限を加えた文書頻度と*adaptation*を用いた．また，文字は索引語でないとみなした．語wの文字数を$l(w)$として，定義を次に示す．

定義 5

$$keyword = \begin{Bmatrix} w \mid 0.00005 < df(w)/N < 0.1, \\ 0.1 < df_2(w)/df(w), l(w) > 1 \end{Bmatrix}$$

4. 情報検索性能評価

本研究で示した索引語抽出法が情報検索において有効であるか確認するために，索引語抽出を行わず検索質問をそのまま文書との一致に使った場合と，抽出した索引語を文書との一致に使った場合の情報検索性能を比較する．

本研究における情報検索性能の評価にはNTCIR2テストコクレションを用いた[3]．用いた検索質問は短い記述であり．自由文である．

情報検索システムには文字単位の*bigram*インデックスを用いた．検索質問において2文字の部分文字列をすべて抽出し，文書との一致をとる．それぞれの一致に対して語の重みを加算することによって検索質問と文書をスコアリングし，そのランクキングを検索結果とする．

語の重みには$TF \cdot IDF$の一種を用いた[4]．$tf(w,d)$を文書dに語wが含まれる頻度として，次に定義する．

定義 6

$$TF \cdot IDF = (1+\log tf(w,d))\left(1+\log\frac{N}{df(w)}\right)$$

本研究で示した索引語抽出法を用いて抽出した索引語を太字にして，検索質問の例を表1に示す．

NTCIR2テストコレクションには検索質問に完全に一致した文書のみを正解とした厳密な正解判定と，部分的に適合した文書も正解とする緩やかな正解判定の二つが含まれる．両方を用いて判定を行った結果を表2に示す．

表 1　検索質問の例

番号	検索質問
0104	**モノクローナル抗体**を利用した**肺小細胞癌**の**診断**と**治療**
0108	**XML**を用いた**自然言語処理**に関する**論文**
0109	**TCP**を**高速**化するために**改良**した**論文**
0110	**情報の可視化**を用いることで**情報検索**を**支援**するシステムについて**論じた文献**はないか。

表 2　検索質問毎の勝敗の割合(11 点平均精度)

正解判定	提案法が勝った数	提案法が負けた数
厳密	39	10
緩やか	42	7

仮説検定を用いて提案法の有効性を確認する．n個のうちm個以上の検索質問に対して提案法を用いた検索が高い性能を持っていた場合，提案手法が勝つ確率が低い$(p \le 1/2)$という仮説を立てる．二項分布を用いて考えればこの仮説が成立する確率は次のようになる．

$$P(X \ge m) = \frac{1}{2}\int_0^{1/2}\sum_m^n {}_nC_m p^m (1-p)^{n-m}\,dp \le \sum_m^n {}_nC_m (1/2)^n$$

表2より，厳密な正解判定の場合は危険率1.92×10^{-5}で，緩やかな正解判定の場合は危険率1.81×10^{-7}で，それぞれ仮説は棄却される．

5. まとめ

本研究では自由文以外の知識を一切必要としない索引語抽出法を用いて情報検索を行い，その結果情報検索性能を有意に向上させることができることを明らかにした．

References

[1] 武田善行，梅村恭司(2001)，キーワード抽出を実現する文書頻度分析，計量国語学，vol. 23, no. 2, pp. 65--90

[2] Kenneth W. Church(2000), Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2, *Coling*, pp. 173--179.

[3] Noriko Kando(2001), Overview of the Japanese and English IR Tasks at the Second NTCIR Workshop(Draft), *Proceedings of the Second NTCIR Workshop Meeting*, pp.4-37--4-60.

[4] Ian H. Witten, Alistair Moffat and Timothy C. Bell, Managing Gigabytes Compressing and Indexing Documents and Images, MORGAN KAUFMANN PUBLISHERS.