日本語テキストに対する統計的検索手法の性能比較
ーテストコレクションによる実証ー

岸田和明

Empirical examination on performance of some statistical methods for Japanese text retrieval by using large test collection

Kazuaki KISHIDA

Abstract

The paper reports some findings from an empirical study on comparison of retrieval performance between some statistical methods: vector space and probabilistic models. A large Japanese text test collection provided by the NACSIS was used, which consists of about 330,000 records of scientific proceedings. Each statistical method was testified using three kinds of indexing techniques for Japanese text: (1) longest matching against entries in a dictionary, (2) tokenizing by change of kind of characters, (3) a simple bi-gram method. Almost no statistically significant difference among the methods was observed, but it seems that probabilistic method based on logistic regression model indicates relatively better performance than other methods.

1 はじめに

ベクトル空間型モデルや確率型モデルなどの統計的な情報検索手法の性能に関する実証研究が、近年、米国における大規模検索実験プロジェクト TREC（http://trec.nist.gov/）等において試みられ、多くの知見が得られつつある。日本においても、1998年から99年にかけて学術情報センター（現：国立情報学研究所）によって同様の大規模テストコレクションが準備され、検索実験がおこなわれた（名称は NTCIR、http://www.rd.nacsis.ac.jp/）。このテストコレクションによって、日本語テキストに対する各手法の検索性能の実証的比較が可能になったわけであり、本稿は3つの代表的な統計的検索手法に対するその結果を報告するものである。

2 統計的検索手法と索引作成

2.1 代表的な統計的検索手法

本研究が取り上げる統計的な検索手法は次のとおりである。なお、以下の算式は1件の検索質問に対する1件の文献についてのものであり、結果として算出されるvによって、各文献が順位づけられて出力されることとする。

①ベクトル空間型モデル[1]

$$v = \sum_{j=1}^{M} d_j q_j \Big/ \sqrt{\sum_{j=1}^{M} d_j^2 \times \sum_{j=1}^{M} q_j^2}$$

$$d_j = \log tf_j + 1.0, \quad q_j = \left(\log qf_j + 1.0\right) \times \log\left(N/n_j\right)$$

②確率型モデル(1)：Okapi型[2]

$$v=\sum_{j=1}^{M}\left(\frac{2.2\times tf_j}{1.2\times(0.25+0.75dl/avdl)}\times qf_j\times\log\frac{N-n_j+0.5}{n_j+0.5}\right)$$

③確率型モデル(2)：ロジスティック回帰型[3]

$$v=-3.51+\frac{1}{\sqrt{L}+1}\Phi+0.0929L,$$

$$\Phi=37.4\sum_{j=1}^{L}\frac{qf_j}{ql+35}+0.330\sum_{j=1}^{L}\log\frac{tf_j}{dl+80}-0.1937\sum_{j=1}^{L}\log\frac{F_j}{F}$$

ここで、tf_jは文献中の語jの出現頻度、qf_jは検索質問内の出現頻度、n_jは語jが出現する文献数、Nは全文献数、Mは語の異なり総数、dlは文献の長さ、$avdl$はdlのデータベース中での平均、qlは検索質問の長さ、Fはデータベース中のすべての語の延べ出現総数、F_jは語jのデータベース中での延べ出現総数、Lは文献と検索質問とで共通する語数。

2.2 日本語テキストに対する索引作成

日本語テキストに対する索引作成の方法としては、最近では自然言語処理技法を応用したものなど様々な手法が提案されているが、本研究では、比較的単純な以下の4つの方法をそれぞれ別個に利用する。

① 辞書との最長照合（辞書は茶筅のもの（http://cl.aist-nara.ac.jp/lab/nlt/chasen/）を使用）
② 単純な字種切り（カタカナと漢字の間にも切れ目を入れ、ひらがなのみの索引語は除去）
③ 字種切り＋複合語分解（カタカナと漢字との間には切れ目を入れず、字種の切れ目によって識別された複合語に対して独自の分解アルゴリズムを適用）
④ バイグラム（2文字ずつ機械的に重複して切り分ける）

このうち、③の複合語分解のアルゴリズムとは、辞書との照合によって、複合語の構成要素を識別する単純な方法である（辞書との照合によって識別された構成要素に従って複合語を切り分け、それらも索引語として採用する）。ただし、1文字のみから成る構成要素は強制的に除去し、元の複合語については検索質問におけるその重みを機械的に2倍にした。

3 データと実験環境

3.1 テストコレクションの概要と処理

利用したテストコレクションは学術情報センターによる“ntc1-je1”（ad hocタスク用）であり、様々な学会の会議報告の標題や抄録等が339,483件収録されている。これらの標題と抄録（日本語のもの）のみを索引作成の対象とした。検索質問は“0031”から“0083”までの53件を用いた（詳細は学術情報センターのWebページを参照）。なお、今回利用したのは<narrative>フィールドと<concept>フィールドを含めた「長い（long）」検索質問文である。これらのファイル（適合判定のファイルを含む）の提供を学術情報センターより受け、IBM Aptiva H-55（150MHz、メモリ96MB、OSはLinux）上で、C/C++を使って上記の索引作成および検索をそれぞれ実行した。なお、各検索質問の適合文献の平均は約36文献（最少で6、最多で581）であった。

3.2 パラメータ

上記の検索手法のパラメータのうち主なものを表1に示す（Nは339,483）。なお、あまりに数多くの文献に出現する語は索引語として機能しないと仮定し、今回は3,000件以上に出現

する語は索引語として採用しないという方針を採った。

表1 主なパラメータ

	最長照合	単純字種切り	字種切り+分解	バイグラム
avdl	63.347	49.574	109.755	198.319
F	21,505,229	16,829,687	37,259,843	67,325,789

4 比較実験の結果

4.1 平均精度による比較

TRECにおいて標準的に用いられている尺度である、非補間の平均精度（average precision）を表2に示す。統計的手法3通り、索引作成方法4通りであるから、実行結果は全部で12通りとなるが、表2からわかるように、それらの間にはそれほど顕著な差はない。例えば、表2中で最も平均精度の高い「ロジスティック回帰型・字種切り＋分解」と（.299）、最も低い「ベクトル空間型・バイグラム」（.204）との間には有意水準95%で統計的な差があるものの（大標本法に基づく平均値の差の検定）、これら12通りの実行結果のペアの大部分には大きな差は見られない。ただし、全体的に、統計的手法としては「ロジスティック回帰型」が良く、索引作成方法としては「字種切り＋分解」が優れているようである。

表2 平均精度（非補間）：53検索質問に対する平均値

	最長照合	単純字種切り	字種切り+分解	バイグラム
ベクトル空間型	.205	.210	.274	.204
Okapi型	.227	.221	.294	.228
ﾛｼﾞｽﾃｨｯｸ回帰型	.242	.250	.299	.298

4.2 再現率－精度グラフによる比較

TRECにおける標準的な補間の方法を使って描いた再現率－精度グラフを図1に示す。なお、凡例については、「ベクトル空間型」をvec、「Okapi型」をokapi、「ロジスティック回帰型」をlogistと略記した。また、索引作成方法については、表1あるいは表2の表頭で左から順に通し番号（1～4）を付けて表した。4.1での結果と同様に、各手法間でそれほど顕著な差は見られない。

4.3 相関行列による分析

53の検索質問に対する各手法の平均精度のデータ（53行×12列）を使って相関分析を試みた。これは、手法間の性能の優劣が検索質問によって異なるかどうかをマクロ的に見るためである。最も相関が高かったのは、バイグラムを使った場合のベクトル空間型とOkapi型で.96であった。逆に最も低かったのは、最長照合によるベクトル空間型とバイグラムによるOkapi型で.54であった。

5 おわりに

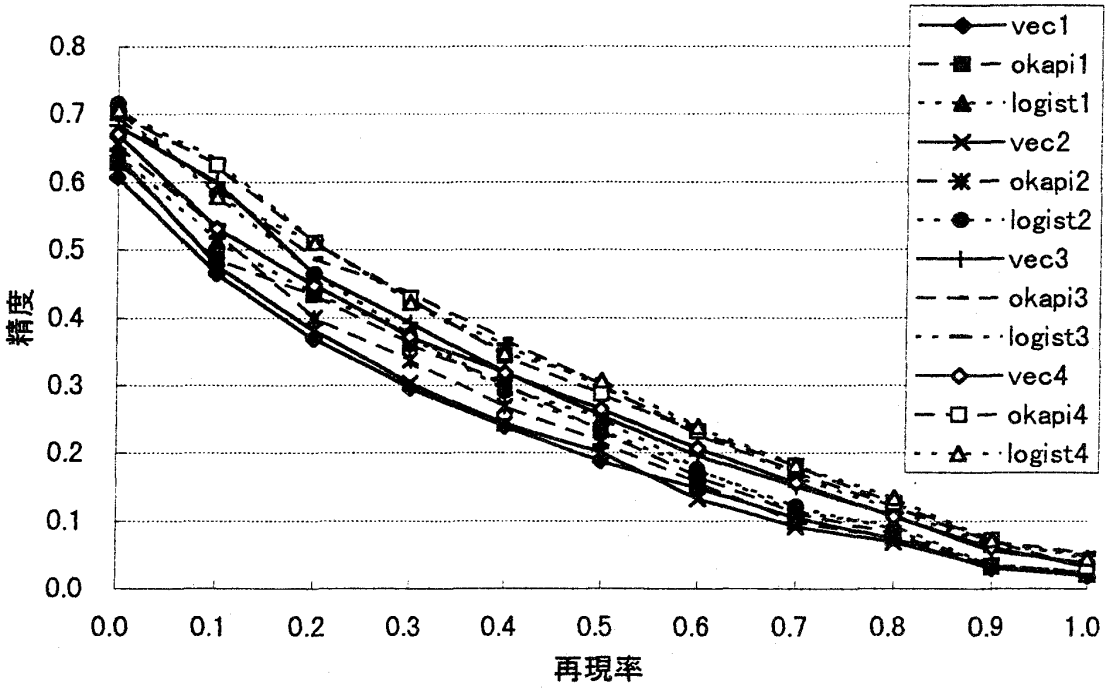
本研究では、学術情報センターの大規模テストコレクションを使って、3つの著名な統計的検索手法の日本語テキストにおける性能比較を実証的に試みた。索引作成方法としては比較的

単純な4種類の方法を用いた。その結果、手法間に顕著な差はなかったものの、統計的手法ではロジスティック回帰型モデル、索引作成方法では字種切りによって複合語を抽出し、さらにそれを辞書を使って分解する方法が優れていることが明らかになった。今後の課題は、

①どのような検索質問の場合にどの手法が優れているかを明らかにするためのミクロ的分析

②より高度な自然言語処理技法に基づく索引作成方法を使った場合の性能評価

の2点である。



謝辞

貴重なテストコレクションを準備され、研究目的の使用を認めていただいた学術情報センターの皆様に深く感謝いたします。

参照文献

[1]Buckley, C., et al. "Automatic query expansion using SMART: TREC 3," NIST Special Publication 500-226: Overview of the Third Text REtrieval Conference (TREC-3). (http://trec.nist.gov/)

[2]Robertson, S.E. et al. "Okapi at TREC-4," NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4). (http://trec.nist.gov/)

[3]Cooper, W., et al. "Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression," NIST Special Publication 500-215: The Second Text REtrieval Conference (TREC-2). (http://trec.nist.gov/)

岸田和明（駿河台大学 文化情報学部）

Kazuaki KISHIDA (Faculty of Cultural Information Resources, Surugadai Univ.)