1C4-04

類似概念判別に基づく情報検索システムの検索性能の評価

Evaluation of an information retrieval system with Concept-based Vector space Model.

阿部 仁志[*1]
Satoshi Abe

湯川 高志[*1]
Takashi Yukawa

[*1] 長岡技術科学大学
Nagaoka University of Technology

An information retrieval system with concept-based vector space model was evaluated and compared with Boolean model and probability based model to clarify its performance and application domain. The experiments were conducted using NTCIR-1, Japanese test collection for information retrieval. The performance and problems of the system are shown, and improvement methods are also discussed.

1. はじめに

近年、電子的ドキュメント、WWW および電子メールが広範囲の使用されるようになり、大規模なテキスト情報ベースから情報を検索し、抽出し、管理する方法が必要となってきている。検索においては、的確な検索要求をユーザが示すこと、人為的に付与したキーワードを含まないような大量の情報から必要な情報を検索することは非常に困難である。このような状況から、簡潔なユーザーインターフェイス、少ない検索要求から効率よく必要な情報を得たいという要求は日々ますます強いものとなっている。

上述の要求に応えるものとして、概念類似判別に基づいた情報検索モデルが提案されている[Kato 99]。このモデルは、概念ベースという語に関する知識を用いている。それにより、入力された検索語の幅を広げ、あいまいな検索要求でも的確な検索が可能であると考えられている。

しかしながら、これまでその有効性を広範に確認できる評価は行われていなかった。加藤らによって日本語情報検索テストコレクション BMIR-J2 を用いた評価結果が報告されているが、小規模な検索対象であった。そこで本稿では、概念類似判別に基づいた情報検索システムの検索性能について、NTCIR-1（本格版）を用いて、既存の方式である Boolean 検索、確率モデルによる検索と広範に比較した。また、評価結果に対し分析・考察を行った。それに基づき、複合語やドキュメント全体で頻度の高い語を検索語に用いると適合率が低くなるという問題がわかった。それを改善する方法の提案を行った。

2. 比較対象とした情報検索モデル

2.1 概念類似判別に基づいた情報検索モデル

概念類似判別に基づいた情報検索モデルは、ベクトル空間モデルのバリエーションであり、検索対象に基づいた概念ベースを用いる。概念ベースの張る概念ベクトル空間に、概念の頻度で特徴付けられた検索対象ドキュメントを配置し、ドキュメントベクトルと検索要求ベクトルとの余弦係数によってランク付けされる検索方式である。

連絡先：長岡技術科学大学, 〒940-2188 新潟県長岡市上富岡町 1603-1, Tel: 0258-47-5143, greens@stn.nagaokaut.ac.jp

概念ベースは、検索対象から概念のベクトル空間を張るために、検索対象における語の共起を、統計的な処理をして生成される。具体的には、検索対象ドキュメントの中で頻度の高い N 語（N は数千～一万程度）に対して、語と語の共起マトリックス（N×N）を生成し、それを特異値分解（SVD）によって 100～200 次元程度に圧縮する。ドキュメントベクトルは、そのドキュメントに含まれる概念について、それらの概念ベクトルの合成として生成される。検索要求のベクトルは、ドキュメントベクトルと同様に検索要求に含まれる概念ベクトルの合成で求められる。

2.2 Boolean モデル

Boolean モデルは、ドキュメント内のすべての単語を検索対象とし、検索要求の論理式が真となるドキュメントを検索結果とする検索モデルである。検索要求の論理式は単語を and、or などで結合したものである。

2.3 確率モデル

確率モデルは、ドキュメントが検索要求に適合する確率を推定し、ドキュメントの検索要求に対する適合度を計算してドキュメントをランク付けする検索方法である。

検索要求に適合するドキュメント R とし、あるドキュメント d が検索要求に適合する確率を $P(R \mid d)$、適合しない $\overline{P}(R \mid d)$ とすれば、その比 $g(d)$ は以下の式で求められ、ベイズの定理により

$$g(d) = P(R \mid d) / \overline{P}(R \mid d)$$
$$= P(d \mid R) / \overline{P}(d \mid R) \times P(R) / \overline{P}(R)$$

となる。上式に、$g(d)$ の大小関係を維持したまま、索引語がドキュメントに含まれている確率で表現するために、定数項を無視して対数をとり、式を整理すると以下の式となる[Robertson 76]。

$$g(d) \sim \sum \log P(i) / \overline{P}(i) + \sum \log (1 - P(i)) / (1 - \overline{P}(i))$$

ここで、$P(i)$ は検索要求に適合するドキュメントに索引

語 $t(i)$ が付与されている確率、$\overline{P}(i)$ は検索要求に適合するドキュメントに索引語 $t(i)$ が付与されていない確率である。

上記の式より、ドキュメントをランク付けするには、$P(i)$ を知る必要がある。索引語 $t(i)$ と検索質問 Q に対して、以下の式で最尤推定ができる。

$$P(i) = r/Nr \quad , \quad \overline{P}(i) = (n-r)/(N-Nr)$$

ここで、N は全ドキュメント数、Nr は検索要求 Q に適合するドキュメント数、n は索引語 $t(i)$ が付与されているドキュメント数、r は索引語 $t(i)$ が付与されている適合ドキュメント数である。Nr, n, r は、経験に基づく適当な初期値を設定する。また、検索結果をユーザに判定させて、その結果をフィードバックし、パラメータの更新を行うことを繰り返してパラメータの精度を改善することも可能である。

3. 実験

3.1 実験方法

日本語情報検索テストコレクション NTICR-1[1] の検索対象ドキュメントは、国内の 65 学会が主催する研究会、全国大会などの発表論文の著者抄録約 33 万件である。検索課題は、title , description , narrative , concept , field で構成され、title に検索要求の簡単で主要な概念、description に検索要求の記述、narrative に検索要求の判定基準、検索の目的、背景知識などの詳細な検索要求説明、concept に検索要求に関連した概念および、その類義語・上位語・下位語、field に分野、がそれぞれ記述されている。

今回の実験では、概念ベースモデルを用いた検索システムでは概念類似判別に基づいた情報検索システム、Boolean モデルを用いた検索システムでは Namazu[2]、確率モデルを用いた検索システムでは ruby-ir[3] [内山 01]を用いて比較を行った。検索対象の NTCIR-1 のデータセットは人為的に付与したキーワードを含んでいるが、そのようなキーワードがない検索対象での評価を行うため、このキーワードは除いた。

NTCIR-1 に収録される検索課題 Topic0001-0030 を用いて実験を行った。検索要求は NTCIR-1 の検索課題の構成要素から、簡単な検索要求での適合率の評価のために、検索要求が記述される description のみを用いて実験を行った。概念ベースシステム及び Namazu では、description を形態素解析し、その名詞句（3〜7 単語）のみを検索語とした。また、Namazu ではその単語を全て and で結合する論理式を検索要求として検索を行った。ruby-ir では、検索システム内で形態素解析されるため、description そのものを検索要求として入力した。また ruby-ir ではフィードバックは行わないものとした。

3.2 実験結果

各検索モデルでの検索課題 Topic0001-0030 おける検索結果の再現率 70%までの平均適合率 P-ave を表 1 に示す。再現率が 70%に満たない場合は、平等に比較を行うために、再現率の低い検索モデルの再現率までの平均適合率とした。また、Boolean モデルは再現率が低いために検索結果で得られた再現率までの平均とした。

表 1. 各検索モデルにおける平均適合率 P-ave

	概念ベースモデル	確率モデル	Boolean モデル
	P-ave	P-ave	P-ave
topic0001	19.9%	51.9%	70.1%
topic0002	11.7%	58.9%	0.0%
topic0003	14.1%	7.6%	0.0%
topic0004	5.3%	35.7%	0.0%
topic0005	6.3%	7.2%	0.0%
topic0006	5.7%	25.6%	0.0%
topic0007	4.7%	4.5%	0.0%
topic0008	0.0%	100.0%	0.0%
topic0009	12.6%	4.9%	0.0%
topic0010	3.7%	39.0%	84.2%
topic0011	0.2%	62.4%	0.0%
topic0012	0.7%	50.4%	0.0%
topic0013	0.2%	15.4%	100.0%
topic0014	57.2%	64.0%	100.0%
topic0015	2.0%	23.9%	0.0%
topic0016	0.3%	71.0%	0.0%
topic0017	11.0%	35.6%	100.0%
topic0018	15.5%	18.7%	25.0%
topic0019	11.3%	77.9%	97.6%
topic0020	2.2%	48.3%	100.0%
topic0021	48.3%	16.6%	21.0%
topic0022	7.3%	41.0%	75.0%
topic0023	21.9%	40.0%	63.9%
topic0024	31.1%	31.3%	43.4%
topic0025	0.0%	97.6%	99.6%
topic0026	2.0%	67.3%	100.0%
topic0027	14.0%	36.9%	0.0%
topic0028	44.5%	7.6%	0.0%
topic0029	29.6%	27.0%	36.1%
topic0030	0.2%	5.7%	0.0%

この中で概念ベースモデルが他の検索モデルに比べて、高い適合率を示した topic は 30 件中 6 件であった。その中で特に高い適合率を示した topic0021「機械翻訳における構造処理能力の評価」の recall-precision 図を図 1.に示す。

図 1.では、概念ベースモデルは、どの再現率の時も確率モデル、Boolean モデルに比べて適合率が高い結果となった。

[1] NACSIS Test Collection for IR System 1 (NTCIR-1)
http://research.nii.ac.jp/ntcir/

[2] Namazu project 開発 (Namazu)
http://www.namazu.org/

[3] 通信総合研究所 (ruby-ir)
http://www2.crl.go.jp/jt/a132/members/mutiyama/software.html

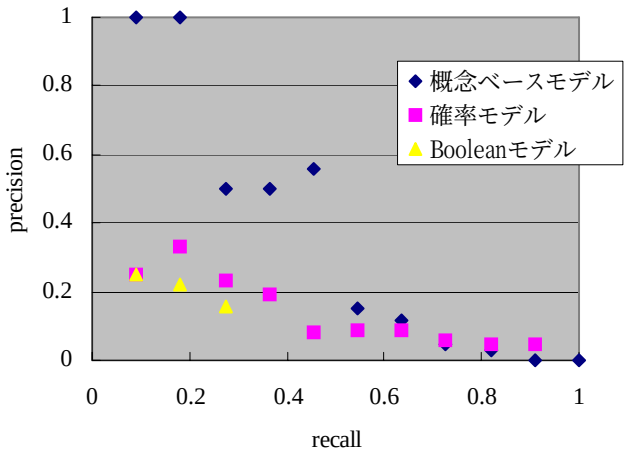

図 1.recall-precision topic0021
「機械翻訳における構造処理能力の評価」

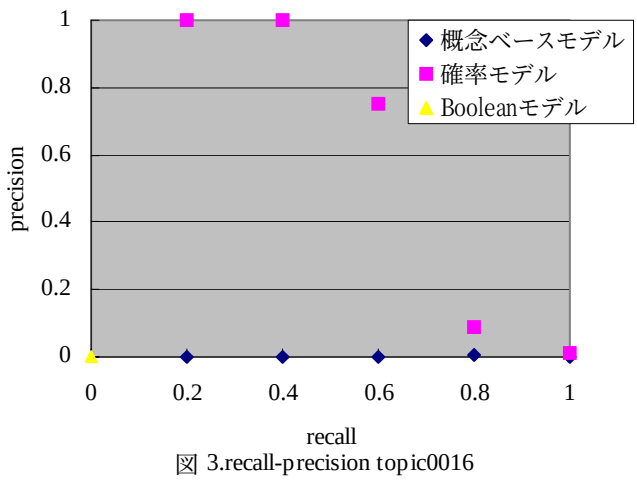

図 3.recall-precision topic0016
「最大共通部分グラフ問題について」

また、この中で概念ベースモデルが他の検索モデルに比べて、特に低い適合率を示したものとして、topic0004「モデルベースの文書画像理解について述べた文献」の recall-precision 図を図 2 に、topic0016「最大共通部分グラフ問題」の recall-precision 図を図 3 にそれぞれ示す。

図 2. 3.では、概念ベースモデルは、再現率に関わらず、適合率が確率モデルに劣っている。しかし、Boolean モデルに比べて適合率は高い結果となった。

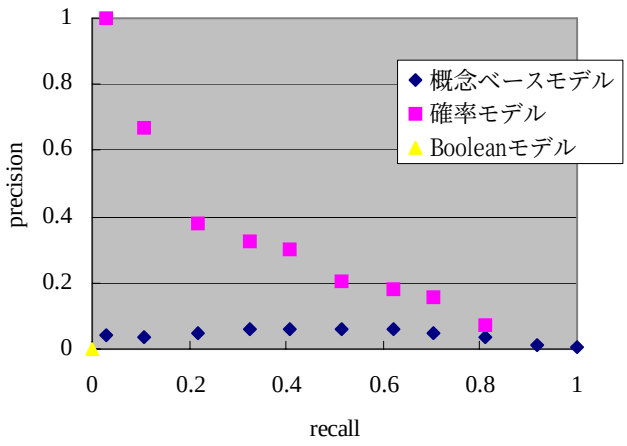

図 2.recall-precision topic0004
「モデルベースの文書画像理解について述べた文献」

4. 分析・考察

実験結果において、特に他の検索モデルに比べて、平均適合率の低かった検索要求について分析して、以下の二つことが分かった。

（1）複合語のような、いくつかの名詞句で構成される単語が多く使われる学術的、技術的な分野での検索対象において複合語が分割される場合、概念ベースモデルは他の検索モデルに比べて適合率が劣る。このような複合語に関する問題は、確率モデルや Boolean モデルでも同様であると考えられる。しかし、概念ベースモデルは、複合語が分割されると、それぞれの語と複合語に含まれない他の語の共起がベクトル空間の形成を大きく変化させるために、その他のモデルに比べて影響が大きいと考えられる。

具体的には、複合語は検索対象から概念ベースを作成するときに、形態素解析によって名詞句に分割される。このため、実験結果に示したような図 3 の topic0016 の検索要求「最大共通部分グラフ問題」のような複合語の場合、形態素解析の結果「最大」、「共通」、「部分」、「グラフ」、「問題」のように分割される。この形態素解析されたそれぞれの語は、この複合語を構成する語以外にも様々な語と共起している。そのため、合成されたドキュメントベクトルは、そのドキュメントの特徴を的確に示すベクトルとは異なるものになっている。その結果、適合するドキュメントと異なるドキュメントが検索要求に適合すると判断され、適合率を下げたと推測される。逆に、概念ベースモデルが他の検索モデルに比べて、高い適合率をして示した、図 1 の topic0021 のような例を考える。この例は、検索要求が「機械翻訳」、「構造処置能力」、「評価」と複合語を正しく形態素解析され、概念ベースに配置している。このような場合、概念ベースモデルは、確率モデル、Boolean モデルに比べて高い適合率を示している。

以上のことから、概念ベースモデルは、複合語を正しく概念ベースに配置し、ドキュメントベクトルを生成することができれば、他の検索モデルに比べて高い適合率を示すと推測される。

（2）全ドキュメントにわたって頻度の非常に高い単語が検索語に含まれる場合、適合率が著しく低下する。概念ベースシステムでは document frequency（DF）は考慮されていない。そのため、ドキュメント全体で頻度の高い単語が検索語に含まれることが、検索の適合率を下げた推測される。

また、この両方の特徴を併せ持つ検索要求に関しては、特に適合率が低い。図 2.図 3.はこの場合に該当する。改善のための方式を以下に提案する。

5. 改善のための提案

上述の分析により、検索の適合率を下げる問題が明らかになった。これら二つの問題を解決する方法として、以下の二つを提案する。

（1）概念ベースの生成時に、形態素解析によって分割された複合語は、その連結して生起する割合に基づいて概念ベースの高次元空間に、一つの概念として配置する。ドキュメントベクトル、検索要求ベクトルを求める際に、その特徴を示す重要な要素となり、複合語が多く使われる学術的、技術的な分野での検索対象でも高い適合率となると推測される。

（2）検索要求に頻度の高い単語を使用する場合には TF×IDF により重み付けの処理を行った後に、検索要求ベクトルを作成することによって改善する方法を提案する。これにより概念ベースに、より的確な検索要求ベクトルを作成でき適合率が上がると推測される。なお、加藤らは概念ベースによる検索モデルで得られたスコアと、TF×IDF 法によって得られたスコアを線形結合する方式を提案している[Kato 99]が、本提案の方が、DF 値の多きい語の検索ベクトルへの影響をより小さく抑えられると考えられる。

6. まとめ

本稿では、概念類似判別に基づいた情報検索システムの検索性能について比較評価を行った。日本語情報検索テストコレクション NTCIR-1（本格版）で、人為的に付与されたキーワードを含まない検索対象を用いて、数語程度の検索要求での適合率の評価を示した。概念ベースシステムでは、複合語や DF が高い語が、検索要求にある場合に適合率が低かった。今回の比較結果の分析・考察により、概念ベースモデルの改善法を二つ提案した。一つは、概念ベース作成する時に、複合語を正しく概念ベースに配置する。二つ目は IDF の考慮した検索要求ベクトルの生成である。この二つの改善方法を、インプリメントし、評価していくことが、今後の課題である。

7. 謝辞

概念類似判別に基づく情報検索システムは、NTT コミュニケーション科学基礎研究所社会情報研究部においてインプリメントされたプログラムを共同研究契約に基づき貸与を受け利用した。開発者の同研究部吉田仙氏（現 NTT 西日本研究開発センタ）、桑原和弘氏（現 ATR 知能ロボティクス研究所）に感謝する。Boolean モデルおよび確率モデルを用いた検索システムは、インターネット上に公開されている Namazu（Namazu project 開発）および ruby-ir（通信総合研究所内山将夫氏開発）を利用させていただいた。また、評価用データコレクションとして、国内学会の提供する学会発表データベースの一部を利用して国立情報学研究所によって作成された NTCIR-1（本格版）を使用許諾に基づき利用した。開発者・作成者諸氏に感謝する。

参考文献

[Kato 99] Kato, T., Shimada, S., Kumamoto, M., and Matsuzawa, K.:Idea-deriving information retrieval system. Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition. pp. 187-193 (1999).

[Baeza-Yates 99] R. Baeza-Yates, B. Ribeiro-Neto :Modern Information Retrieval, Addison Wesley, (1999)

[Robertson 76] S. E. Robertson and K, Sparck Jones, :Relevance weighting of search terms. Journal of the American Society for Information Sciences, 27(3):129-146, (1976).

[内山 01] 内山将夫, 井佐原均 :情報検索パッケージの実装, 情報処理学会研究報告, FI-63-8 pp. 181-184 (2001).