4D-5

キーワード検索の精度・再現率の限界についての考察

和多 太樹 † 廣川 佐千男 ††
† 九州大学大学院システム情報科学府 †† 九州大学情報基盤センター

1 はじめに

Web 文書や特許情報のような莫大な数の文書から必要とする文書だけを効率よく見つけ出すための技術が必要とされている。情報検索における代表的な検索モデルとして、ブーリアンモデル、ベクトル空間モデル、確率モデルがある。検索方式の評価法として一般的に用いられる尺度として再現率と適合率がある。再現率は、適合する文書の何割が検索結果として得られたかを示し、適合率は、検索結果の中の何割が適合する文書であったかを示す。再現率と適合率はトレードオフの関係にあり、精度を上げようとすると再現率が下がり、再現率を上げようとすると精度が下がることが経験的に知られている。検索システムの評価や比較を行なうために、検索質問についての検索結果の評価をのため、横軸を再現率、縦軸を適合率としてプロットする再現率適合率グラフを描くと、右下がりのグラフになるのが一般的である。しかし、ユーザの求めているファイル、つまり正解集合が何であるのかは実際的には通常わからないものであり、テストコレクションの構築においては適合情報を作成する作業が最も困難といわれている [4]。本発表では、ランダムに生成した検索要求と正解集合について、適合率・再現率を求めてその限界を考察した。

2 ブーリアンモデルによる再現率・適合率の限界

ブーリアンモデルでは AND、OR、NOT の組み合わせにより柔軟にクエリを生成することができる。単純には AND(NOT) で検索範囲を狭め、OR で広げることが可能である。しかし、AND を多く組み合わせて、検索結果を厳選した場合、得られら結果が正解ばかりだとしても取りこぼしがあるかもしれない。一方、OR を用いて結果を広く得たとすると再現率は高い値が得られるかもしれないが、ノイズまで拾ってしまい適合率が下がる可能性がある。したがって、ブーリアンモデルを用いて、再現率・適合率ともに高い結果を得ようと思うと、AND、OR、NOT を組み合わせた複雑なクエリ生成が必要になってくる。また、ユーザが検索の度に何十ものキーワードを用いることは通常考えにくく、普通、検索を行う際に使用するキーワードには限度がある。そう考えると、精度の高い検索には、どのようなキーワードを選択するのかも重要な要素であると言える。キーワードの選び方次第では、どのようなブール式を生成しても再現率・適合率に限界が生じてしまうのではなかろうか。加えて、これら再現率・適合率の限界はドキュメントの集合にも依ることも自然に考えられる。ここまでのことをまとめると、人間が常識的に使うことのできるキーワード数の上ではブーリアンモデルによる検索には再現率・適合率には限界があり、その限界はドキュメント集合や使用するキーワード群の選び方にも依存すると考える。

本発表では、単純なキーワードの組み合わせについての網羅的実験に基づき、精度・再現率の限界を調べ、ドキュメントとキーワードの選び方がその限界にどのような影響を与えるかについて考察した。

3 実験概要

定義より、再現率・適合率は、(1) 単語文書行列、(2) 検索質問、(3) 正解集合の 3 つの要素から決まる。

表 1: 単語文書行列の例

	w_1	w_2	w_3	w_4	w_5	w_6
D_1	1	1	0	0	1	0
D_2	1	0	1	0	0	0
D_3	0	0	1	0	0	0
D_4	0	1	1	1	1	0
D_5	1	0	0	0	0	1
D_6	1	0	1	0	1	0
D_7	0	1	0	0	0	1

例えば、表 1 のような単語文書行列を考えたとき、検索質問を w_3 AND w_5、正解集合を $\{D_1, D_4, D_7\}$ とすると、この検索質問からブーリアンモデルで得られる検索結果は、$\{D_4, D_6\}$ となる。このとき、再現率および適合率は、再現率 $= 1/3 = 0.33$, 適合率 $= 1/2 = 0.5$ となる。また、(110010) というベクトルを検索質問として考えることもできるこれは論理式 $w_1 \wedge w_2 \wedge \overline{w_3} \wedge \overline{w_4} \wedge w_5 \wedge \overline{w_6}$ として表すこともできる。言葉で表現すると、w_1、w_2、w_5 だけを含み、他の単語は

Limit of Boolean Query for IR
† Taiki WADA(t-wada@i.kyushu-u.ac.jp)
†† Sachio HIROKAWA(hirokawa@cc.kyushu-u.ac.jp)
Graduate School of Information Science and Electrical Engineering, Kyushu University (†)
Computing and Communications Center, Kyushu University (††)

含んでいない文書の検索を意味し、検索結果は $\{D_1\}$、再現率および適合率は、再現率 $= 1/3 = 0.33$, 適合率 $= 1/1 = 1$ となる。

ユーザーが検索を行う際には何が正解であるかは事前にはわからない。そこで、単語文書行列に対し、検索質問と正解集合の対をランダムに生成し、その再現率・適合率を求める実験を行った。具体的には、まず、100×20 の単語文書行列をランダムを一つ作り、ランダムに生成する 100000 組の検索質問と正解集合の組に対し、再現率・適合率を求めた。

検索質問に関しては、「ベクトル表現」と「キーワードの AND 検索」の二種類の実験を行なったが、紙面の都合で後者についてだけ述べる。AND 検索で用いるキーワードの数を 1 〜 3 と変化させそれぞれについて再現率適合率グラフをプロットした。

4 実験結果と考察

検索質問を AND 検索と捉え、プロットしたものが図 1 である。再現率はほぼ変わらないのに対し、適合率はキーワード数を増やすにつれて良くなっていることがわかる。これは、キーワード数を増やすほど結果が厳選され、より目的に合致する文書が得られたからと考えられる。図 1 において、右上半部分には全然プロットされていない。これは単純な AND 検索では再現率、適合率の両方が同時に高い値となることがないことを意味する。

ともに 1.0 となる場合は、検索質問により正解文書集合が完全に特徴付けられていることを意味する。ベクトル表現の検索質問を考えるとこれは、形式概念束 [1, 2] における「概念」といえる。単語文章行列の行あるいは列のサイズを n とすると、正解集合も検索のベクトル表現も指数オーダー $O(2^n)$ となる。一方固定された単語文章行列について、概念はそれほど多くない。この結果、再現率適合率グラフの右上半部分は空白となると考えられる。さらに、ユーザーが与えた検索ベクトルに対し、概念束における上位概念、あるいは下位概念を使うことで検索拡張や検索の絞り込みを行なえば、再現率あるいは適合率の向上が考えられる。

5 まとめと今後の課題

再現率と適合率はトレードオフの関係にあり、精度を上げようとすると再現率が下がり、再現率を上げようとすると精度が下がることが経験的に知られている。ブーリアンモデルを用いて、再現率・適合率ともに高い結果を得ようと思うと、AND、OR、NOT を組み合わせた複雑なクエリ生成が必要になってくる。ユーザが使うキーワード数を高々 3 個と仮定して、単純な

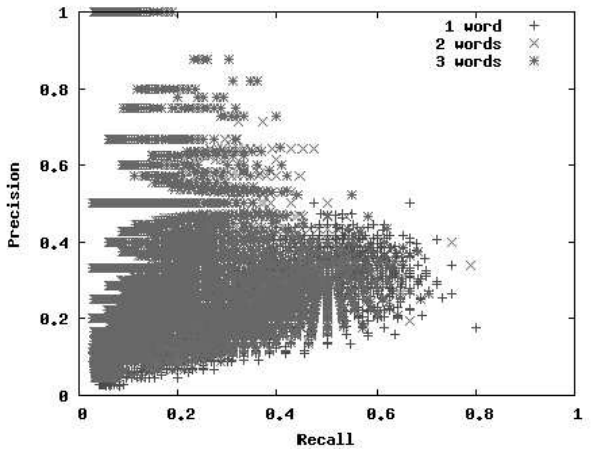

図 1: ランダムな AND 検索の再現率適合率グラフ

キーワードの AND の組み合わせについて、精度・再現率の限界をランダムな実験的により求めた。またその限界の理由を概念束の理論から考察した。

本発表では、検索質問として AND 検索で実験を行ったが、OR や NOT についての実験も必要である。検索質問の変形 [3] として概念束や概念グラフ [5] を利用する方式についても今後検討する予定である。

参考文献

[1] Bernhard Ganter, Rudolf Wille, C. Franzke, Formal Concept Analysis : Mathematical Foundations, Springer-Verlag, 1999

[2] Claudio Carpineto, Giovanni Romano, Concept Data Analysis : Theory and Applications John Wiley & Sons, 2004

[3] Masaharu YOSHIOKA, Makoto HARAGUCHI. "An Appropriate Boolean Query Reformulation Interface for Information Retrieval Based on Adaptive Generalization", International Workshop on Challenges in Web Information Retrieval and Integration, pp. 145-150, 2005.

[4] 栗山和子, 吉岡真治, 神門典子. "大規模テストコレクション NTCIR-2 の構築: 対話型追加検索と言語横断的プーリングの効果", 情報処理学会論文誌データベース, Vol.43, No. SIG2 (TOD13), pp.48-59, March 2002

[5] 廣川佐千男, 下司義寛, 和多太樹. 文書群からの概念グラフの構成, 情報処理学会第 169 回自然言語処理研究会,pp.79–84, 2005