情 報 学 基 礎 47－14
（1997. 9. 12）

ランキング文書検索におけるスコア合成法の評価

小川 泰嗣* 松田 透**
(株) リコー *ソフトウェア研究所 **情報通信研究所

文書のランキング検索の検索精度に影響を与える項目の一つとして、スコア合成法がある。スコア合成法とは、複雑な検索要求に対して、複数の処理単位の重要度から文書スコアを合成するための計算方式のことである。本稿では、算術和・確率和・確率積・最大値・最小値・P-NORM・P-CONORM の７種類の合成方式を、情報処理学会が作成・配布している情報検索用ベンチマークである BMIR-J1 を用いて評価した。

Evaluating Score Combination Methods in Ranking Retrieval

OGAWA Yasushi* MATSUDA Toru**
*Software Research Center **Information and Communication Reaserch Center
RICOH Co., Ltd.

One asspect that affects the effectiveness of ranking retrieval is a score combination method that calculates a single document score using weights of indexing units extracted from the user's query. Although many combinataion methods have been proposed, their effectiveness has not been studied yet in Japan. Thus, we have evaluated the following seven methods — sum, probabilistic and/or, minimum, maximum, p-norm and p-conorm — using the BMIR-J1 benchmark developed by IPSJ-SIGDBS.

1 はじめに

文書検索においては、対象データが自然言語で記述されているので、単純な boolean 方式よりもランキング方式が有効性とされている [2][11]。ランキング検索では、検索条件に対する文書の適切さ（以下、文書スコアと呼ぶ）を検索文書ごとに計算し、検索文書の順序付けを行う。ランキングのための文書スコア計算には多くの要素が絡み合っているが、特に以下の項目が検索精度に与える影響が大きい。

1. 処理単位の選択方式

処理単位とは、文書スコアを計算するうえでの基本単位のことであり、以下ユニットと呼ぶ。英語では単語をユニットとすることが一般的であるが、日本語では、英語のように単語がわかち書きされていないので、文書中のどんな要素をユニットとするかが問題となる [3][8]。形態素解析を利用して単語をユニットとする方式（単語索引）、文書中の連続する n 文字である n-gram をユニットとする方式（n-gram 索引）などがある。

2. ユニットの重要度計算方式

ある文書のスコアは、検索条件に含まれるユニットの、その文書における重要度から計算される。重要度はユニットに関する統計情報

[0]E-mail: *yogawa@src.ricoh.co.jp,
**matsuda@ic.rdc.ricoh.co.jp

などを用いて算出するのが一般的であり、具体的な計算方式には tf*idf 方式など様々なものがある [2]。

3. スコア合成方式

単一のユニットで構成されている検索条件であれば、そのユニットの対象文書における重要度をそのまま文書スコアとすればよい。しかし、複数のユニットから構成される複雑な検索条件の場合、それらユニットの重要度を合成し、文書スコアを得なければならない。合成方式としては、算術和・確率和・確率積など様々なものが提案されている [6]。

日本語文書のランキング検索に関しては、これまでユニットの選択方式・重要度計算方式については様々なバリエーションの比較検討が行われてきたが、スコア合成方式に関しては行われていない。そこで、本稿は合成方式に焦点をあて、様々な合成方式を評価する。検索精度にはスコア合成方式を含む上記三つの要因が複雑に絡まっているので、全ての組み合わせについて評価を行う必要があるが、そのような網羅的評価の実施は困難である。そこで、本稿では、ユニットとしては n-gram (n-gram 索引) を、重要度算出方式としては確率モデルを採用することとして、評価を行った。

以下、2章で n-gram 索引、3章で確率モデルに基づく重要度計算方式、4章でスコア合成方式について説明する。評価実験の方法・結果・考察は5章に示す。

2 N-gram 索引

本稿では、n-gram をユニットとする n-gram 索引を採用したが、n-gram 索引においては n が検索精度に大きく影響する [8]。したがって、本実験では以下の四つの方法を試みた。

2.1 Uni-gram 索引

Uni-gram (単一文字: $n = 1$) をユニットとする方式である。登録時には、文書中の全ての uni-gram の出現を索引ファイルに登録する。検索時には、(1) 検索要求文中の全ての uni-gram を抽出する、(2) 抽出された uni-gram の少なくとも一つを含む文書を検索する、(3) 検索文書全てについて、uni-gram の頻度情報を用いて文書スコアを計算する、という手順で処理が行われる。例えば、「アジアの熱帯雨林」からは「ア」「ジ」「ア」「の」「熱」「帯」「雨」「林」がユニットとして抽出される。

2.2 Bi-gram 索引

Uni-gram 索引では二文字以上から構成される単語が uni-gram に分割されるため、単語としての意味的な情報が失われ、検索精度の点で不利である。これに対し、bi-gram (二文字組: $n = 2$) をユニットとする bi-gram 索引では、単語と同等あるいはそれ以上の精度が得られる [8]。この方式では、部分的に重なり合う隣接する bi-gram もそれぞれ独立に抽出する。したがって、さきの例文からは「アジ」「ジア」「アの」「の熱」「熱帯」「帯雨」「雨林」が抽出される。

2.3 Combi 索引

日本語には単一文字で構成される単語があるので、単一文字語による検索もできなければならない。単純な bi-gram 索引でも、登録時には文書から全ての bi-gram を抽出するので、検索語として指定された単一文字からはじまる全ての bi-gram を OR 結合した検索条件を生成することで、単一文字語による検索を実施することができる。しかし、JIS で定義されている文字数は約 7000 であるため、これら文字の全てが実際に出現するわけではないことを考慮しても、検索速度の著しい低下が懸念される [7]。

この問題を解決する方法として、uni-gram 索引と bi-gram 索引の組み合わせが考えられる。以下、この方式を **combi 索引**と呼ぶ。検索精度の面でも、Uni-gram 索引には再現率を向上させる効果があるので [3]、combi 索引の方が bi-gram 索引よりも有利である [8]。この場合、さきの例文からは「ア」「ジ」「ア」「の」「熱」「帯」「雨」「林」「アジ」「ジア」「アの」「の熱」「熱帯」「帯雨」「雨林」が切り出される。

2.3.1 文字種考慮型 Combi 索引

日本語には、平仮名・片仮名・漢字など複数文字種があり、それらの使われ方には大きな違いがある。したがって、テキストから単純に n-gram を切り出すよりも文字種を考慮して、異なる文字種を跨がることがないように n-gram を抽出する方がよいと考えられる。

この場合、さきの例文からは「ア」「ジ」「ア」「の」「熱」「帯」「雨」「林」「アジ」「ジア」「熱帯」「帯雨」「雨林」が切り出される。文字種を考慮しない場合と比較して、「アの」「の熱」が抽出されない点が異なる。

3 確率モデルによるユニット重要度計算方式

本稿では、Robertson が提案した確率モデル [9] を採用した[1]。このモデルでは、検索条件 Q に対するユニット U_i の文書 D_j における重要度を以下のように計算する

$$w_{ij} = \frac{\log(N/df_i)}{\log(N)} \cdot \frac{qf_i}{Kq + qf_i} \cdot \frac{tf_{ij}}{Kd + tf_{ij}} \quad (1)$$

ここで、df_i は U_i を含む文書数（以下文書頻度）、qf_i は Q における U_i の出現頻度（要求内頻度）、tf_{ij}は D_j における U_i の出現頻度（文書内頻度）、Nはデータベース中の全文書数である。また、Kq, Kd は要求内頻度、文書内出現頻度の正規化パラメータである。

4 文書スコア合成方式

文書スコアの合成方式としては、これまで様々なものが提案されている [6]。本稿では、以下の合成方式を評価対象とする。

1. SUM:算術和

   Robertson らの提案したランキングモデルでは、検索条件が複数ユニットから構成される場合には、算術和を用いて文書スコアを算出としていた [9]。この場合、文書 D_j の文書スコア r_i は下式で計算される。

   $$r_j = \sum_{U_i \in Q} w_{ij} \quad (2)$$

2. OR:確率和

   複数の独立な事象の確率からそれら事象の少なくとも一つが起こる確率を求めるための計算方式である。

   $$r_j = 1 - \prod_{U_i \in Q} (1 - w_{ij}) \quad (3)$$

   導出ネットワークモデル [12] などにおいて、OR 演算子に対応した重要度の合成に使用されている。

3. AND:確率積

   複数の独立な事象の全てが起こる確率を求めるための計算方式である。

   $$r_j = \prod_{U_i \in Q} w_{ij} \quad (4)$$

   確率和の双対として、AND 演算子に用いられる。

4. MAX:最大値

   ファジィモデル [4] で採用されている計算方式であり、OR 演算子に用いられる。

   $$r_j = \max_{U_i \in Q} w_{ij} \quad (5)$$

5. MIN:最小値

   MIN の双対であり、AND 演算子に用いられる。

   $$r_j = \min_{U_i \in Q} w_{ij} \quad (6)$$

6. P-NORM

   幾何的距離の計算式を定義域・値域が [0,1] となるように修正したもので、拡張 Boolean モデル [10]・ファジィモデルにおいて OR 演算子に対応して使用されている。

   $$r_j = \left(\frac{\sum_{U_i \in Q} w_{ij}^p}{n} \right)^{\frac{1}{p}} \quad (7)$$

[1] 正確には、オリジナルの計算式に若干の修正を加えてある [8]。

パラメータ $p(1 \leq p)$ によって特性が変化するが、特に $p = 1$ は算術平均[2]と一致し、$p \to \infty$ では MAX となる。

7. P-CONORM

P-NORM の双対であり、$p = 1$ は P-NORM と同様に算術平均に一致し、$p \to \infty$ では MIN となる。

$$r_j = 1 - \left(\frac{\sum_{U_i \in Q} (1 - w_{ij})^p}{n} \right)^{\frac{1}{p}} \quad (8)$$

5 評価

5.1 評価方法

前章で示した合成方式について、検索精度による評価を行った。P-NORM/P-CONORM のパラメータ p としては、2, 3, 5 の三つの値を試みた。

評価用のデータとしては、(株) 日本経済新聞の協力によって、(社) 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）BMIR-J1 を利用した [5]。BMIR-J1 の規模は表 1 に示す通りである。なお、対象文書である記事にはあらかじめキーワード等の補足情報が付与されているが、今回の実験ではタイトルと本文のみを対象文書として使用した。

検索精度の評価指標としては、平均適合率 [11] を使用した。平均適合率とは、まずランキングごとの再現率・適合率から再現率が $0.0, 0.1, \cdots, 1.0$ における適合率を求め、つぎにこれらの再現率での適合率の平均をとったものである。なお、再現率は正解文書を洩れなく検索できる能力、適合率は正解でない文書を検索しない能力を表すもので、以下の式で計算される。

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{正解文書数}} \quad (9)$$

[2]今回は算術平均は比較の対象に含めていない。それは、演算子が入れ子になる複雑な検索条件は今回の実験対象としていないため、算術平均・算術和のいずれを用いても、同一検索条件に対するランキング結果が一致するからである。

表 1: BMIR-J1 の規模

	検索要求	対象文書
件数	60	600
平均文字数	11	703
最小文字数	2	102
最大文字数	28	3802
合計サイズ	—	872KB

$$\text{適合率} = \frac{\text{検索された正解文書数}}{\text{検索文書数}} \quad (10)$$

Robertson モデルでは、式 (1) に含まれるパラメータが検索精度に影響する。そこで、文書内頻度の正規化パラメータ Kd については $0.0, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0$ のように変化させ、要求内頻度の正規化パラメータ Kq については BMIR-J1 の検索要求が比較的短いので 0.0 に固定することとした。そして、スコア合成方式ごとに、全ての Kd について平均適合率を測定し、そのなかの最大値を最終的な評価結果とした。

5.2 評価結果

平均適合率の測定結果を、索引方式ごとにまとめたものが表 2〜表 5である。これらの表の比率の欄は、本来の Robertson モデルに相当する SUM をベースラインとし、その値に対する残りのスコア合成方式の平均適合率の増減比を示している。また、索引方式ごとに平均適合率が最も高いものには＊を付与してある。

スコア合成方式による検索精度の相違については次節で詳しく検討することとし、ここでは索引方式の傾向を見ておく。AND, MIN を例外とし、それ以外の合成方式に共通なこととしては、uni-gram, bi-gram, combi, 文字種考慮型 combi の順で検索精度が向上していることがあげられる。例えば、SUM 合成方式で見ると、uni-gram に対する向上率は 29.0%, 31.5%, 33.3% となっている。以上の結果から、(n-gram に基づく) 索引方式のなかでは、文字種考慮型 combi 索引が最も優れていると言えよう。

表 2: Uni-gram 索引に対する平均適合率

合成方式	平均適合率		比率 (%)
AND	0.2137		-41.2
MIN	0.1856		-48.9
P-CONORM ($p=5$)	0.3506		-3.5
P-CONORM ($p=3$)	0.3615		-0.0
P-CONORM ($p=2$)	0.3615		-0.0
SUM	0.3635		—
P-NORM ($p=2$)	0.3712	*	+2.1
P-NORM ($p=3$)	0.3645		+0.0
P-NORM ($p=5$)	0.3509		-3.5
MAX	0.2834		-22.0
OR	0.3647		+0.0

表 3: Bi-gram 索引に対する平均適合率

合成方式	平均適合率		比率 (%)
AND	0.1132		-75.8
MIN	0.1128		-75.9
P-CONORM ($p=5$)	0.4553		-2.9
P-CONORM ($p=3$)	0.4600		-1.9
P-CONORM ($p=2$)	0.4658		-0.1
SUM	0.4688	*	—
P-NORM ($p=2$)	0.4681		-0.0
P-NORM ($p=3$)	0.4505		-3.9
P-NORM ($p=5$)	0.4319		-7.9
MAX	0.3906		-16.6
OR	0.4656		-0.2

表 4: Combi 索引に対する平均適合率

合成方式	平均適合率		比率 (%)
AND	0.1543		-67.7
MIN	0.1545		-67.7
P-CONORM ($p=5$)	0.4635		-3.0
P-CONORM ($p=3$)	0.4683		-2.0
P-CONORM ($p=2$)	0.4733		-0.9
SUM	0.4780		—
P-NORM ($p=2$)	0.4810	*	+0.6
P-NORM ($p=3$)	0.4631		-3.1
P-NORM ($p=5$)	0.4414		-7.6
MAX	0.4090		-14.4
OR	0.4798		+0.1

表 5: 文字種考慮型 Combi 索引での平均適合率

合成方式	平均適合率		比率 (%)
AND	0.1133		-76.6
MIN	0.1128		-76.7
P-CONORM ($p=5$)	0.4664		-3.7
P-CONORM ($p=3$)	0.4730		-2.4
P-CONORM ($p=2$)	0.4780		-1.3
SUM	0.4845		—
P-NORM ($p=2$)	0.4905	*	+1.3
P-NORM ($p=3$)	0.4695		-3.1
P-NORM ($p=5$)	0.4463		-7.8
MAX	0.4009		-17.3
OR	0.4876		+0.7

5.3 考察

以下では、スコア合成方式による検索精度の相違について考察する。

まず気がつくことは、AND および MIN は他の方式と比較して極端に検索精度が悪いことである。AND・MIN 以外の方式では、検索条件から抽出された複数のユニットのうちの一つでも文書中に出現していれば、その文書のスコアは非 0 となり、ランキング結果に含まれるので、検索洩れは発生しにくい。これに対し、AND・MIN では、抽出ユニットのうちの一つでも出現していないものがあれば、文書スコアは 0 となり、検索されない。その結果、検索洩れが多くなり、検索精度が低くなる。実際、平均検索件数を比較すると、uni-gram 索引の場合で AND・MIN では 8.5 件、それ以外では 548 件と大きな差があった。Bi-gram 索引では 1.5 件・313 件と、差はさらに大きい。

つぎに、残りの合成方式のなかでは MAX が劣っていることがわかる。これは、MAX では文書スコアが最大の重要度のみによって決定され、残りのユニットの重要度が無視されるという single operand dependancy [6] が原因と考えられる。

残りの SUM, OR, P-NORM, P-CONORM はいずれも高い検索精度を示しており、スコア合成方式として有効と考えられる。P-NORM, P-CONORM に関しては、パラメータ $p=1$ では実質的に SUM と等しくなるので、p が小さい場合（特に $p=2$）に SUM と同等（ないしそれ以上）の検索性能を示すのであろう。逆に p を大きくすると検索精度が低下しているのは、特性が MAX あるいは MIN に近づくことが原因である。

OR に関しては、重要度が小さい範囲では SUM とほぼ一致し、重要度が大きくなるにしたがって MAX に特性が近づく。Uni-gram と bi-gram を比較すると、文書頻度は前者の方が大きい傾向にあるので、式 (1) からわかるように重要度は小さくなる。このことが、bi-gram において OR が

SUM よりも若干ではあるが劣っていることの理由と考えられる。

6 おわりに

本稿では、ランキング検索の有効性に影響を与える項目の一つであるスコア合成方式について、BMIR-J1 を用いて評価を行った。算術和・確率和・確率積・最大値・最小値・P-NORM・P-CONORM の7種類の合成方式を比較した結果では、P-NORM が最も優れていたが、算術和・確率和もほぼ同等の検索精度を示すことが分かった。

今後の課題としては、以下の項目があげられる。

- スコア合成方式の特性をより詳細に検討することで、索引方式ごとの合成方式の優劣を決定している要因をつきとめる。
- 今回は n-gram 索引と Robertson 重要度計算方式を用いて評価実験を行ったが、それ以外の索引方式（単語索引など）・重要度計算方式（tf*idf 方式など）を用いて評価を行う。
- P-NORM が優れているということは、英語を対象とした評価結果 [1][6] とも一致する。しかし、これらの実験では AND/OR 演算子を用いて構造化した検索条件を用いている。今回使用した単純な検索条件よりも構造化検索条件の方が高い検索精度を期待できるので、BMIR-J1 の検索要求に対する構造化検索条件を作成し、構造化条件を使用した場合のスコア合成方式の評価を行う。

参考文献

[1] E.A. Fox and J.A. Shaw. Combination of multiple searches. In *Proc. of 2nd TREC*, pp. 243–252, 1994.

[2] W.B. Frakes and R. Basza-Yates. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, New Jersey, 1992.

[3] H. Fujii and W. B. Croft. A comparison of indexing techniques for Japanese text retrieval. In *Proc. of 16th ACM SIGIR Conf.*, pp. 237–246, 1993.

[4] 日本ファジィ学会. 講座ファジィ. ファジィ・データベースと情報検索. 日刊工業新聞, 1993.

[5] 芥子 育雄. 情報検索システム評価用ベンチマーク Ver.1.0(BMIR-J1) について. 研究会報告 *DBS-106*, pp. 139–145. 情報処理学会, 1996.

[6] J.H. Lee. Analyzing the effectiveness of extended Boolean models in information retrieval. Technical Report TR95-1501, Cornell University, 1995.

[7] 小川 泰嗣. 日本語文書検索のための頻度情報を用いた効率的部分文字列索引の提案. 情報処理学会論文誌, Vol. 37, No. 10, pp. 114–120, 1996.

[8] Y. Ogawa and T. Matsuda. Overlapping statistical word indexing: A new indexing method for Japanese documents. In *Proc. of 20th ACM SIGIR Conf.*, pp. 226–234, 1997.

[9] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. of 17th ACM SIGIR Conf.*, pp. 232–241, 1994.

[10] G. Salton, E.A. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, Vol. 26, No. 12, pp. 1022–1036, 1983.

[11] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[12] H. Turtle and W.B. Croft. Inference networks for document retrieval. In *Proc. of 13th ACM SIGIR Conf.*, pp. 1–24, 1990.