1E2-01

Transductive学習による最小文書判定からのクエリ拡張

Query Expansion with the Minimum Judgment

岡部正幸[*1] 梅村恭司[*2] 山田誠二[*3]
Masayuki Okabe Kyoji Umemura Seiji Yamada

[*1]豊橋技術科学大学 情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology

[*2]豊橋技術科学大学 情報工学系
Information and Computer Science, Toyohashi University of Technology

[*3]国立情報学研究所
National Institue for Informatics

Query expansion techniques generally select new query terms from a set of top ranked documents. Ideally, all of those documents should be known as relevant by a user's manual judgment. However it is difficult to get enough feedback from users in practical situations. In this paper we propose a query expansion technique which performs well even in the case where only a relevant document and a non-relevant one are known and relevance of other documents are unknown. To overcome the lack of relevance information, our technique prepares some tentative relevant documents using transductive learning. Terms for expansion are selected from those documents by a standard scoring formula with a canonicalization function. Several functions are tested for the canonicalization. Experimental results show that our technique outperforms some traditional methods in standard precision and recall criteria.

1. はじめに

クエリ拡張とは，情報検索分野において検索結果を改善させるテクニックの一つであり，これまでに多くの手法が提案されている．中でも，ユーザからのフィードバック情報を用いる方法は良い性能を示すことが知られており，標準的な手法として多くの研究で用いられている．この手法は，一般にユーザからより多くの文書判定情報が与えられるとそれだけ検索性能を向上させる良いクエリを選択できるが，通常の検索場面においてユーザから十分なフィードバック情報を得ることは極めて難しい [1].

そこで本研究では，ユーザフィードバックが極めて少ない場合，特に適合文書と非適合文書が一つずつのみ与えられた場合においても有効なクエリ拡張方法を提案する．この設定において一番の問題は既知の適合文書が極めて少ないため，追加単語候補のスコア計算が適切に行えないということである．我々は与えられた文書の適合性情報の不足を補うため，Transductive 学習と呼ばれる機械学習の方法を用いて，他の適合である可能性の高い文書を利用することを試みる．フィードバックを利用して検索性能を向上させる方法はこれまでにいくつも提案されているが，本研究のようにユーザから得られる文書の適合性情報を最小限に設定するというアプローチは，ユーザから十分な情報が与えられているという従来研究の仮定と比較してより現実的な仮定であり，ユーザフィードバックに要するコストと検索性能の向上の新たなトレードオフ点を探る試みであるともいえる．

2. クエリ拡張

これまで多くのクエリ拡張方法が提案されてきた．中には特定分野の文書を訓練データとして利用するドメイン特化型のクエリ拡張方法もあるが [2, 3]，多くの方法は手動で用意された既知の適合文書または擬似的に適合である見なした文書を利用する．特に Robertson の方法 [4] と呼ばれる手法は，適合文書集合を利用したクエリ拡張の標準的手法として多くの研究で用いられている [5, 6]．本論文で提案するクエリ拡張方法も，この Robertson の方法をベースとする．この手法では，与えられた適合文書の中に出現する各単語について下記のスコア計算を行い，このスコアの高いものから順に追加単語とする．

$$wpq_t = \left(\frac{r_t}{R} - \frac{n_t - r_t}{N - R}\right) * \log \frac{r_t/(R - r_t)}{(n_t - r_t)/(N - n_t - R + r_t)} \quad (1)$$

ここで，r_t は単語 t を含む既知の適合文書数，n_t は単語 t を含む文書数．R は既知の適合文書数，N はコレクション内の文書数である．(1) 式の第 2 項は，Robertson/Spark Jones の重みと呼ばれ，Okapi 検索システムの単語の重み付けにも利用されている．(1) のスコア計算式は元々次に示す式から導き出されたものである．

$$wpq_t = (p_t - \bar{p_t}) \log \frac{p_t(1 - \bar{p_t})}{\bar{p_t}(1 - p_t)} \quad (2)$$

p_t は単語 t が適合文書中に現れる確率，$\bar{p_t}$ は単語 t が非適合文書に現れる確率である．これを見てわかるように，p_t と $\bar{p_t}$ の推定精度がスコア計算に影響を与える．(1) 式の場合では，p_t を $\frac{r_t}{R}$ で，$\bar{p_t}$ を $\frac{N_t - R_t}{N - R}$ によって推定している．このように推定する場合，ある程度の適合文書数が必要となる．pseudo フィードバックにより検索結果の上位 n 文書[*1]を適合文書とみなす方法もあるが，この pseudo フィードバックは初期検索性能に大きく依存するという問題がある．

本研究では，文書の適合性情報はユーザから与えられると仮定するが，p_t を精度よく推定できるだけの適合文書がユーザによって示されることは，現実の検索場面においてほとんど期待できない．そこで我々は，ユーザから最低限のフィードバックが与えられるという状況を考える．具体的には，ユーザフィードバックによって適合文書と非適合文書が一つずつ分かっているという仮定をおく．しかしこの設定で (1) 式を利用

連絡先: 岡部正幸，豊橋技術科学大学 情報メディア基盤センター，〒 441-8580 豊橋市天伯雲雀ヶ丘 1-1，0532-44-6639，okabe@imc.tut.ac.jp

*1 通常は $n = 30$ 前後

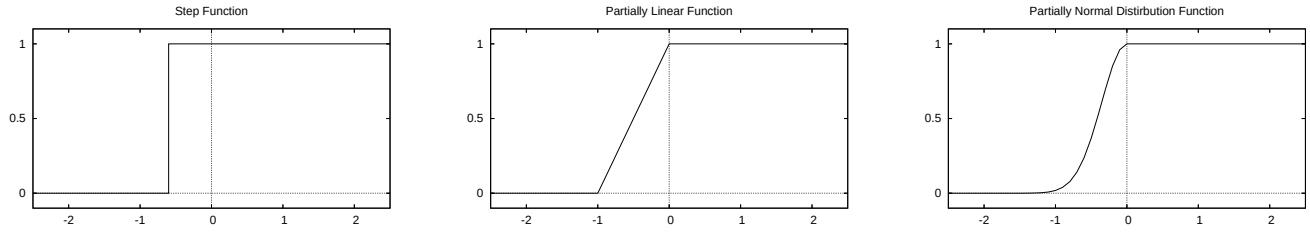

図 1: 変換関数 $f(x)$

すると，r_t と R が小さすぎて，p_t の推定をうまく行うことができない．そこで本研究では，機械学習によって適合文書である可能性の高い文書を選び出し，暫定的に適合文書数を増やすことを考える．機械学習方法には，訓練文書数が少ない場合に有効とされる，Transductive 学習を用いる．

3. Transductive Learning

Transductive 学習は，transduction と呼ばれる推論方法に基づいており，訓練データからラベル付け関数を生成することなしに直接テストデータのラベル付けを行う [7].

学習タスクは n 点からなるデータ集合 X 上で定義される．X は訓練データ集合 $L = (\vec{x}_1, \vec{x}_2, ..., \vec{x}_l)$ とテストデータ集合 $U = (\vec{x}_{l+1}, \vec{x}_{l+2}, ..., \vec{x}_{l+u})$ からなる．Transductive 学習では $l \ll u$ となるデータ集合を扱う．学習の目的は L 中の各データと各データがもつラベル情報を用いて，U 中の各データにラベルを割り当てることである．

Transductive 学習を実現するアルゴリズムは，これまでにいくつか提案されており [8, 9, 10]，様々なタスクにおいて訓練データ数が少ない場合での優位性を示している．本研究では，これらの中から高い性能を持つとされる “*Spectral Graph Transducer*”(SGT) [11] アルゴリズムを用いる．SGT は U へのラベル割り当てを制約付き ratiocut 問題として定式化し，この緩和問題を解くことによって近似解を導き出す．

SGT をクエリ拡張に適用する場合，X はヒットリストの上位 n 文書とする．L は既知の適合文書，非適合文書の 2 文書とする．このとき U は X 中の L 以外の文書であり，U の各文書が適合か非適合であるかは未知である．SGT は U の各文書に対して適合文書と判断した場合は γ_+ に近い値を，非適合文書と判断した場合には γ_- を割り当てる．ここで，$\gamma_+ = +\sqrt{\frac{1-p}{p}}$，$\gamma_- = -\sqrt{\frac{p}{1-p}}$ である．p は全データ中に占める適合文書数の割合であるが，通常は正確な値は分からないので適当に決めてやる必要がある．

SGT はいくつか重要なパラメータを持つ．我々が行うクエリ拡張タスクにおいて特に重要なのは，γ_+ と γ_- の推定値 $\hat{\gamma}_+$ と $\hat{\gamma}_-$ である．先に述べたように我々は訓練データが極めて少ない[*2] という仮定をおいているため，この 2 つのパラメータの設定次第で学習性能に大きな影響を与える．次章では，これらのパラメータへの依存性を緩和する方法について述べる．

4. Transductive 学習を利用したクエリ拡張

4.1 学習結果に基づく適合文書の可能性の推定

2 章で述べたように，追加単語を選ぶ際の各単語のスコア計算において重要なのは，ある単語 t が適合文書に出現する確率 p_t を求めることである．Robertson の方法では，既知の適合文書の中で単語 t が出現する文書の割合を p_t の値としている．この場合，適合文書は人手によりチェックが行われた本当の適合文書である．一方，我々の設定では学習アルゴリズムが適合と判断した文書も用いるため，各文書が本当に適合文書であるがどうかは保証されない．実際，SGT では各文書に割り当てられるのは実数値 z であり，$z > 0$ または $z < 0$ となるかでラベルの割り当てを行っている．学習結果は完全ではないので，$z < 0$ となるものでも実際には適合文書である場合も多い．そこで我々は，SGT が学習結果として各データに与える実数値 z を各文書が適合文書である可能性を示す値に変換する関数 $f(x)$ を用意し，この関数値を利用して p_t を計算することにする．$f(x)$ には下記の 3 つの関数を用いた．これらは $z < 0$ となる文書を適合文書とみなす方法が異なる．

1. Step function (SGT-step)

$$f(x) = \begin{cases} 1 & x \geq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

2. Partially Linear function (SGT-linear)

$$f(x) = \begin{cases} 1 & x \geq 0 \\ 1 + x & -1 \leq x < 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

3. Partially normal distribution function (SGT-ndist)

$$f(x) = \begin{cases} 1 & x \geq 0 \\ \exp(-2x^2) & -1 < x < 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

図 1 は各関数の形を示している．横軸は z，縦軸は $f(x)$ を示している．この関数の内から一つを用いて，次式により p_t を計算する．

$$p_t = \sum_{i:t \in d_i} f(z_i) / \sum_{i=1}^{n} f(z_i) \tag{6}$$

各単語 t について，t が出現する各文書 d_i について，各文書に与えられた学習結果 z_i を入力とした関数 f の値の和をとって，それが全文書の合計に占める割合を計算している．(6) 式の n は学習に用いた文書数である．つまり，各文書は単に適合・非適合を示す 1，0 のバイナリ値ではなく，適合文書である可能性を示す値 $f(z_i)$ を持っていると考える．この式は確率の推定値ではなく，単語 t が適合文書に出現する可能性を示す一つの値といえるが，範囲は 0 から 1 の間に収まっており，これを p_t の推定値として扱うことにする．

*2 適合文書と非適合文書が一つずつ，つまり $|L| = 2$

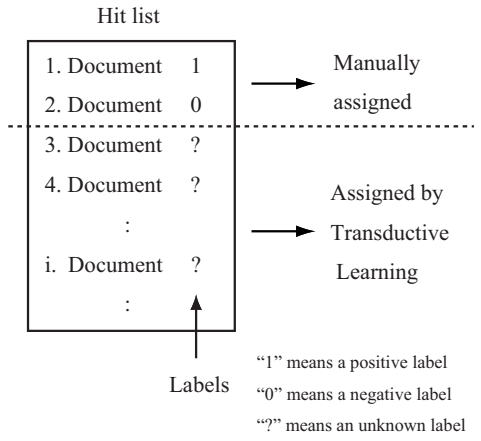

図 2: Transductive 学習による適合文書の発見

4.2 クエリ拡張手続き

ここで，Transductive 学習を利用したクエリ拡張手続きの全体的な流れについてまとめておく．

1. 初期検索：検索システムへの初期クエリを入力する．
2. 適合性の判定：初期検索に対するヒットリストの上位から適合文書と非適合文書を一つずつ見つけ出し，これを訓練データ用の文書とする．
3. Transductive 学習による暫定的適合文書の発見：図 2 に示すように，前のステップで発見された訓練データ文書を利用して他に可能性のある適合文書を SGT によって発見する．適合文書であると学習アルゴリズムが判断しただけで，真の適合文書であるわけではないので，ここでは暫定的適合文書と呼ぶ．これらの文書は適合文書である可能性を示す 0 から 1 の値を持つ．
4. クエリ拡張のための単語選択：(2) 式と (6) 式を用いて各単語のスコアを計算する．このスコアの高いもの上位 m 個を初期クエリに追加する単語として選択する．
5. 拡張クエリによる再検索：初期クエリと前のステップで選択された追加単語を拡張クエリとし，これを検索システムに入力するで，新しいヒットリストを受け取る．ここまでがクエリ拡張の 1 サイクルとなる．

5. 実験

5.1 設定

検索システムとして Okapi システム [12] を用いた．また検索文書データとして TREC-8 の adhoc タスク [13] で用いられた検索課題と正解集合を利用した．

文書データ数は約 520,000 で，各文書にはストップワードの除去とステミングを施した．またトピックには TREC-8 の adhoc タスクで用いられた 50 トピック（No.401-450）の内，初期検索結果の上位 10 位以内に適合文書または非適合文書が一つもランクされないトピック 8 つを除いたものを使用した．

5.2 比較手法

実験では，提案したクエリ拡張方法を次の方法と比較した．

表 1: 11 点平均適合率

	5	10	15	20
Normal	0.191	0.175	0.164	0.162
Pseudo	0.213	0.210	0.206	0.206
SGT-step-0	0.230	0.230	0.220	0.215
SGT-step-α	**0.245**	0.241	0.240	0.231
SGT-linear	0.238	**0.249**	**0.257**	**0.240**
SGT-ndist	0.246	0.248	0.245	0.239

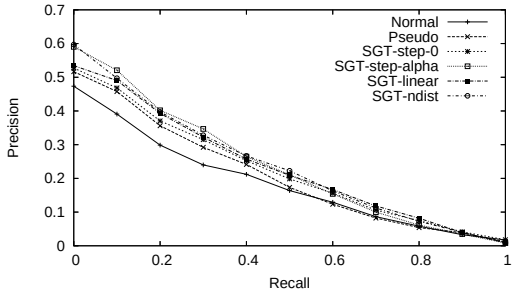

図 3: 再現率－適合率曲線

Normal ： 適合文書と非適合文書一つずつ利用したシンプルな Robertson の方法．

Pseudo ： この方法はいわゆる *pseudo relevance feedback* と呼ばれる方法で，検索結果の上位 n 文書を適合文書とみなす方法である．予備実験において $n = 30$ のとき最もよい性能を示したため，以後の実験でこの方法による結果は値は全て $n = 30$ とした場合の結果である．

提案したクエリ拡張方法は，4 章で説明した関数のどれを使用するかにより，4 つのバリエーションを試した．それぞれ，**SGT-step-0**, **SGT-step-**α, **SGT-linear**，**SGT-ndist** と呼ぶことにする．**SGT-step-0** と **SGT-step-**α はステップ関数における α の値が異なる．前者は $\alpha = 0$ とした方法である．後者は z_i の値が高いもの上位 30 個を含むような値，つまりこの方法は SGT が割り当てた値の大きいもの上位 30 文書を適合文書である可能性を持つ文書として扱う方法である．**SGT-linear** と **SGT-ndist** はそれぞれ部分線形関数と部分正規分布関数を用いた方法である．

5.3 実験結果

表 1 は追加単語数を 5，10，15，20 とした場合のクエリ拡張後に得られた検索結果の 11 点平均適合率を示したものである．再現率と適合率は適合性が既知の 2 文書を除いて計算してある．SGT に使用したデータ文書数は 100 である．また，全データに占める適合文書数の割合を示すパラメータ[*3]は 0.1 とした．この値はトピック全てにおいて適した値ではないが，Pseudo における $n = 30$ という値も同じことがいえるのでここでは全てのトピックについて同じ値を用いた．

結果を見て分かるように，SGT を用いた方法は Normal, Pseudo に比べて概ね良好な結果を示している．追加単語数による差は見られなかった．また，SGT を用いた方法の中では SGT-step-0 に比べて他の方法の結果が良いことから，バイナリの判定結果ではなく適合文書である可能性を示す値を用いた効果があったといえる．ただし，関数による差はあまり見られなかった．

∗3 SGT のパッケージでは "-p" パラメータ

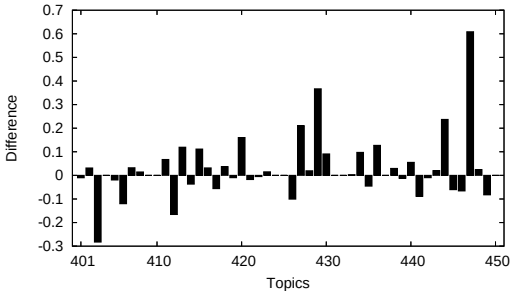

図 4: SGT-step-α と Pseudo の 11 点平均適合率の差

表 2: 上位 n 位における再現率

n	SGT only	SGT-based QE
20	0.083	0.185
40	0.155	0.257
60	0.214	0.294
80	0.297	0.331
100	0.346	0.363

図 3 は，5 単語を追加した場合の各手法の再現率－適合率曲線を示したものである．SGT を利用した方法の曲線が Normal, Pseudo と交差していないことからその優位性が確認できる．

6. 考察

図 4 追加単語数 5 の場合の SGT-step-α と Pseudo の 11 点平均適合率の差をトピックごとに示した棒グラフである．棒が上に出ていれば SGT-step-α の値の方が高いことを示す．図を見て分かるように，いくつかのトピックにおいては Pseudo に負けている．これは，クエリ拡張を行う際に用いた文書集合の中で本当に適合文書であったものの数と関係している．図 5 は SGT-step-α と Pseudo がクエリ拡張を行う際に用いた 30 文書に占める本当の適合文書数の差を示している．多くのトピックにおいて，図 5 に示す適合文書数が勝っていれば（棒が上に突き出ていれば），平均適合率も勝っている．適合文書数がひどく落ち込んでいるトピックも見受けられるが，これは先に述べた SGT に与える適合文書数の割合を示すパラメータ設定がうまくいっていないことが原因である．

本研究ではクエリ拡張を行うことを前提としているが，実際には SGT のみでも適合文書を発見することはできる．表 2 は，SGT が与える判定値を用いて文書を並び替えた場合とクエリ拡張 (SGT-step-α，5 単語追加) を行った場合の上位 n 位までの文書における再現率を比較した表である．この表から分かるように，SGT 単独では効率よく適合文書を発見することができない．また，100 位のところでもクエリ拡張を行った場合が勝っており，クエリ拡張を行うことによって初期検索の上位 100 位以下にランクする適合文書を発見できていることが分かる．

7. まとめ

本研究では，ユーザから得られる文書の適合性情報が最小の場合においても機能するクエリ拡張方法を提案した．適合性情報の不足を補うため，この方法では Transductive 学習の

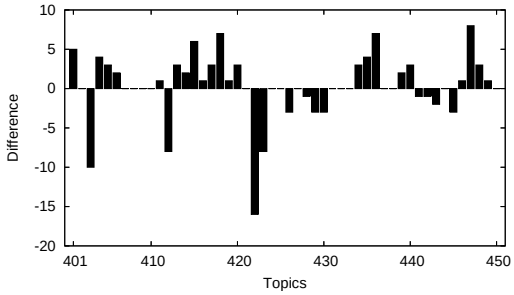

図 5: SGT-step-α と Pseudo の上位 30 文書における適合文書数の差

一種である SGT アルゴリズムを用いた．また，追加候補選択のための単語のスコア計算を学習結果を利用して計算する方法を示した．実験で従来手法と比較した結果，我々のクエリ拡張方法の優位性を示すことができた．SGT のパラメータをより適切に設定することにより改善できる余地は残されており，今後はそれをどのように実現するか検討していく必要がある．

参考文献

[1] S. Dumais and et al. Sigir 2003 workshop report: Implicit measures of user interests and preferences. In *SIGIR Forum*, 2003.

[2] G. W. Flake and et al. Extracting query modification from nonlinear svms. In *Proceedings of WWW 2002*, 2002.

[3] S. Oyama and et al. keysword spices: A new method for building domain-specific web search engines. In *Proceedings of IJCAI 2001*, 2001.

[4] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of SIGIR 2003*, pages 213–220, 2003.

[5] S. Yu and et al. Improving pseud-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of WWW 2003*, 2003.

[6] A. M. Lam-Adesina and G. J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of SIGIR 2001*, pages 1–9, 2001.

[7] V Vapnik. *Statistical learning theory*. Wiley, 1998.

[8] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of ICML '99*, 1999.

[9] X Zhu and et al. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of ICML 2003*, pages 912–914, 2003.

[10] A. Blum and et al. Semi-supervised learning using randomized mincuts. In *Proceedings of ICML 2004*, 2004.

[11] T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of ICML 2003*, pages 143–151, 2003.

[12] S. E. Robertson. Overview of the okapi projects. *Journal of the American Society for Information Science*, 53(1):3–7, 1997.

[13] E. Voorhees and D. Harman. Overview of the eighth text retrieval conference. 1999.