4W−3

ユーザによる文書ランキングの調整が可能な対話的WWW検索支援手法の提案

仲川こころ　木下 敦史　高田 喜朗　関 浩之

奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

WWW (World Wide Web) の普及とそこに流通する文書の増加にともない，数多くの WWW 検索サービスが提供されるようになった．しかし未だ問題点も多く，検索作業におけるユーザの負担は増大しているのが実情である．従来検索の効率や精度（再現率・適合率）に注目した研究が多くなされているが，本研究では特に検索作業中に生じる心理的負担に着目し，これを軽減することを目的としている．

現在のキーワード検索サービスは，検索要求のたびに，データベース中の各文書に対してクエリー（ユーザが入力したキーワードや検索式）との適合度を計算し，その値がしきい値よりも大きい文書だけを適合度の降順に並べたリストをユーザに提示する．ユーザにとって適合度が高い文書がランキングの下位に配置されてしまうような場合もしばしば生じるが，このような場合にユーザはランキングを変更・修正するすべを持たないため，不適切なランキングに我慢を強いられることになる．上記の問題について，ユーザの検索目的や特徴を自動的に推測し，クエリーを適応させていくような研究がある（適合フィードバック [2]）．しかし，著者らの知る限り，ユーザがランキングやスコア計算規則そのものを直接制御・修正できるような手法はほとんど研究されていない．

本稿では，提示されたランキングをユーザが自分の検索目的に沿うように調整し，システムはその調整結果を基に文書のスコア計算規則を修正する，という手法を提案し，上記問題の解決を図る．

2 システム設計の方針

2.1 動作手順の概要

提案手法による動作手順は以下のようになる：

(1) ユーザがキーワードを入力する．

(2) システムは，データベース中の各文書ごとにキーワードに対する適合度を計算し，その降順に並べた文書リスト（初期ランキング）を出力する．

(3) ユーザは，不当に順位が低いと感じる文書を上位に上げるか，または自分にとって不適合と感じる文書を下位に下げる操作を行う（図 1）．

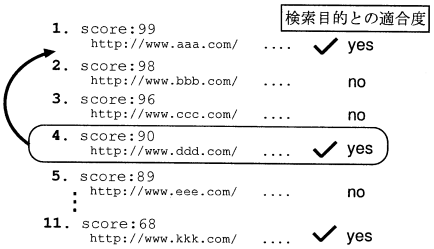

図 1: ユーザによるランキングの調整

(4) システムは，ユーザの調整操作（文書順位の上げ／下げ）からユーザの調整意図を推測し，スコア計算規則を変更する（3 節）．

(5) システムは，手順 (4) で推測した新しい計算規則を用いて，手順 (2) で抽出した各文書のスコアを再計算し，新しいランキングをユーザに提示する．

一回の調整作業ごとにランキング中の全文書のスコアが再計算されるため，ユーザが直接調整しなかった文書も順位が変化することになる．ユーザにとっての適合度を手順 (4) でうまく推測できれば，新しいランキングでは適合文書が（調整前よりも）上位に集まると期待できる．

以下，本手法の各要素について説明する．

2.2 ランキングの作成と調整

2 つのベクトル $\boldsymbol{v}_1$ と $\boldsymbol{v}_2$ のなす角を θ とする時，ベクトル間の類似度 $\mathrm{sim}(\boldsymbol{v}_1, \boldsymbol{v}_2)$ を以下のように定義する．

$$\mathrm{sim}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{(\boldsymbol{v}_1, \boldsymbol{v}_2)}{|\boldsymbol{v}_1||\boldsymbol{v}_2|} = \cos\theta \quad (|\boldsymbol{v}_1| \text{ は } \boldsymbol{v}_1 \text{の長さ}) \quad (1)$$

ユーザが手順 (1) で入力したキーワードの集合を $KW = \{kw_1, kw_2, \ldots, kw_p\}$ とすると，文書 d の KW に対する縮約特徴ベクトル $\mathrm{rfv}(d)$ は以下で定義される $p+m$ 次元ベクトルである．

$$\mathrm{rfv}(d) = (\underbrace{c_{kw_1,d}, c_{kw_2,d}, \ldots, c_{kw_p,d}}_{p}, \underbrace{c_{f_1,d}, \ldots, c_{f_m,d}}_{m}) \quad (2)$$

ここで，$c_{kw_u,d}$ $(1 \le u \le p)$ はキーワード kw_u と文書 d の関連度（$tf*idf$ [4] など），$c_{f_v,d}$ $(1 \le v \le m)$ は単語以外の文書の特徴を示す値である（例えばリンク密度や更新の新しさなど）．以降簡単のため，$\mathrm{rfv}(d)$ を $\boldsymbol{d}$ と表記する．

システムが手順 (4) で推測するスコア計算規則とは，縮約特徴ベクトルの各要素に対する重みを要素に持つベクトル

Adjustable Ranking of Documents – An Interactive Method for Supporting WWW Retrieval

Kokoro Nakagawa, Atsufumi Kinoshita, Yoshiaki Takata and Hiroyuki Seki

Graduate School of Information Science, Nara Institute of Science and Technology

$\boldsymbol{k}=(k_1,k_2,\ldots,k_{p+m})$（クエリーベクトルと呼ぶ）である．$\boldsymbol{k}$ の x 番目の要素 k_x は，$\boldsymbol{d}$ の x 番目の要素に対する重みを示している．文書 $\boldsymbol{d}$ のスコアを，クエリーベクトル $\boldsymbol{k}$ との類似度 $\mathrm{sim}(\boldsymbol{d},\boldsymbol{k})$ で定義する．システムが手順 (2) で抽出した文書集合を D，ユーザの i 回目の調整操作に応じてシステムが作るクエリーベクトル $\boldsymbol{k}^{(i)}$ とすると，D の $\boldsymbol{k}^{(i)}$ によるランキングとは，以下の不等式を満たす文書リスト $d_1,d_2,\ldots,d_{|D|}$ である．

$$\mathrm{sim}(\boldsymbol{d}_1,\boldsymbol{k}^{(i)})\geq\mathrm{sim}(\boldsymbol{d}_2,\boldsymbol{k}^{(i)})\geq\cdots\geq\mathrm{sim}(\boldsymbol{d}_{|D|},\boldsymbol{k}^{(i)})$$

ただし，初期ランキングの作成（手順 (2)）には $\boldsymbol{k}^{(0)}=(\underbrace{1,1,\ldots,1}_{p},\underbrace{0,0,\ldots,0}_{m})$ で表される初期クエリーベクトル[1]を用いる．$\boldsymbol{k}^{(i)}$ $(i\geq 1)$ の作成方法を 3 節で述べる．

3 クエリーベクトルの推測

3.1 線形計画法に基づく手法

2.1 節手順 (3) において，ユーザが文書 d_h を d_{l-1} と d_l $(1\leq l<h\leq|D|)$ の間に移動する調整操作を行った場合，ユーザは次のような考えを持っていると仮定する．

- d_h は $d_1,d_2,\ldots,d_{l-1}$ よりも検索目的に適していない．
- d_h は $d_l,d_{l+1},\ldots,d_{h-1}$ よりも検索目的に適している．

上記の条件を満たすランキングを生成するために，下記の不等式を満たすクエリーベクトル $\boldsymbol{k}=(k_1,\ldots,k_{p+m})$ を求める．

$$\mathrm{sim}(\boldsymbol{d}_h,\boldsymbol{k})\leq\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{k})\quad\text{for } 1\leq j\leq l-1\qquad(3)$$

$$\mathrm{sim}(\boldsymbol{d}_h,\boldsymbol{k})\geq\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{k})\quad\text{for } l\leq j\leq h-1\qquad(4)$$

式 (3)，式 (4) を満たす解は一般に複数存在するため，$k_1,\ldots,k_{p+m}$ に対する別の線形制約式として

$$k_1+\cdots+k_{p+m}=k_1^{(i)}+\cdots+k_{p+m}^{(i)}\qquad(5)$$

を考える．ここで $\boldsymbol{k}^{(i)}=(k_1^{(i)},\ldots,k_{p+m}^{(i)})$ は前回のクエリーベクトルである．以上の制約式 (3)，(4)，(5) を満たし，$\boldsymbol{k}^{(i)}$ と $\boldsymbol{k}$ の類似度が最大になるようなベクトル $\boldsymbol{k}$ を線形計画法 [1] によって求め，新しいクエリーベクトル $\boldsymbol{k}^{(i+1)}$ とする．

3.2 E 尺度に基づく手法

3.1 節と同様に，ユーザが文書 d_h を d_{l-1} と d_l $(1\leq l<h\leq|D|)$ の間に移動する調整操作を行った場合，本手法では 2 つのベクトル $\boldsymbol{g}$ (good) と $\boldsymbol{b}$ (bad) を定義する．

$$\boldsymbol{b}=(\boldsymbol{d}_l+\boldsymbol{d}_{l+1}+\cdots+\boldsymbol{d}_{h-1})/(h-l),$$

$$\boldsymbol{g}=(\boldsymbol{d}_1+\boldsymbol{d}_2+\cdots+\boldsymbol{d}_{l-1}+\boldsymbol{d}_h)/l$$

∗1 手順 (1) でユーザがキーワード以外の要素についても指定できるように実装すると，後半の m 成分についても 0 以外の数値を持った初期クエリーベクトルになる．

ユーザの調整意図を反映した新しいクエリーベクトル $\boldsymbol{k}$ を生成するため，本手法では E 尺度 [3] の概念を用いる．E 尺度は検索結果を評価するための指標の 1 つで，検索結果の適合率 P と再現率 R の重み付き調和平均から求められる．また，値が小さいほど良い検索結果であることを意味する．

$$E=1-\frac{(\beta^2+1)PR}{\beta^2P+R}\qquad\beta\text{は非負定数}\qquad(6)$$

β は適合率と再現率のどちらを重視するかを設定するための非負定数で，両者を同等に見る場合は $\beta=1$ である．直感的に，新しいクエリーベクトル $\boldsymbol{k}$ は，出来るだけ $\boldsymbol{g}$ に類似し，かつ，出来るだけ $\boldsymbol{b}$ に類似していない事が求められる．本手法はこの関係を再現率と適合率のトレード・オフに見立てることで，$\boldsymbol{k}$ の最適化問題を E 尺度の最小化問題と定式化する．ここでベクトル間の非類似度 $\mathrm{dsim}(\boldsymbol{v}_1,\boldsymbol{v}_2)$ を以下のように定義する．

$$\mathrm{dsim}(\boldsymbol{v}_1,\boldsymbol{v}_2)=\sqrt{1-\left(\frac{(\boldsymbol{v}_1,\boldsymbol{v}_2)}{|\boldsymbol{v}_1||\boldsymbol{v}_2|}\right)^2}=\sin\theta\qquad(7)$$

次に，(6) 式の P,R を $\mathrm{sim}(\boldsymbol{g},\boldsymbol{k})$, $\mathrm{dsim}(\boldsymbol{b},\boldsymbol{k})$ で置き換えた式

$$E=1-\frac{(\beta^2+1)\,\mathrm{sim}(\boldsymbol{g},\boldsymbol{k})\,\mathrm{dsim}(\boldsymbol{b},\boldsymbol{k})}{\beta^2\mathrm{sim}(\boldsymbol{g},\boldsymbol{k})+\mathrm{dsim}(\boldsymbol{b},\boldsymbol{k})}$$

を最小にする $\boldsymbol{k}$ を求める．簡単のため $\beta=1$ とすると，与えられたベクトル $\boldsymbol{g}$ と $\boldsymbol{b}$ に対して，以下の r が最大になるような $\boldsymbol{k}$ を求めれば良い．

$$r=\frac{\mathrm{sim}(\boldsymbol{g},\boldsymbol{k})\,\mathrm{dsim}(\boldsymbol{b},\boldsymbol{k})}{\mathrm{sim}(\boldsymbol{g},\boldsymbol{k})+\mathrm{dsim}(\boldsymbol{b},\boldsymbol{k})}\qquad(8)$$

(8) 式 r を最大にする長さ 1 のベクトル $\boldsymbol{k}$ は，次式 (9) で求めることが出来る．

$$\boldsymbol{k}=\frac{c\boldsymbol{g}-\boldsymbol{b}}{|c\boldsymbol{g}-\boldsymbol{b}|}\quad\text{ただし}\quad c=\frac{1+\sin\theta}{\cos\theta}\qquad(9)$$

4 おわりに

ユーザによる調整が可能な文書ランキングの提供による WWW 検索支援について，その目的と設計案を述べた．上記の設計を基に試作したシステムを用いて本手法の評価実験を行い，現在は結果集計中である．今後本手法の有効性の評価，既存の WWW 検索手法との比較分析等を行う予定である．

参考文献

[1] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P.: *Numerical Recipes in C*, pp.430–444, Cambridge University Press, 2nd edition, 1992.

[2] Rocchio, J.: Relevance Feedback in Information Retrieval, in Salton, G. ed., *The SMART Retrieval System : Experiments in Automatic Document Processing*, pp.313–323, Prentice Hall, 1971.

[3] Rijsbergen, van C.: *Information Retrieval*, Butterworths, 2nd edition, 1979.

[4] Witten, I. H., Moffat, A. and Bell, T. C.: *Managing Gigabytes: Compressing and Indexing Documents and Images*, Von Nostrand Reinhold, New York, 1994.