ソーシャルブックマーク数を正解とした検索ランキングの学習

数原 良彦[*1] 植松 幸生[*1] 戸田 浩之[*1] 井上 孝史[*1] 片岡 良治[*1]
Yoshihiko Suhara Yukio Uematsu Hiroyuki Toda Takafumi Inoue Ryoji Kataoka

[*1]日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

Many information retrieval methods that utilize social bookmarking systems can only search web pages that have been bookmarked and assigned tags by users. Therefore, they cannot deal with web pages with no tagging information. In this research, we propose the method to learn ranking function which target not only bookmarked web pages but also non-bookmarked web pages by using social bookmarking information. Experimental result showed that our proposed method can approximate social bookmark number ranking better than conventional ranking methods such as Okapi BM25 and PageRank.

1. はじめに

本研究では，ソーシャルブックマークサービスにおけるユーザのタグ付与履歴を用いて，検索ランキングを行うランキング関数を学習する手法を提案する．

ソーシャルブックマークとは，多数のユーザがコンテンツをブックマークし，コンテンツに対して自由記述のタグ付与を行うことで情報管理，情報共有を行う仕組みのことであり，他のユーザとブックマーク情報を共有することで，個人の嗜好を反映した，発見的な情報検索を行うことができる仕組みのことである．ソーシャルブックマークサービスとしては，世界最大規模の del.icio.us [*1]や，国内最大規模のはてなブックマーク [*2]などが挙げられる．これらのサービスでは，ユーザ，コンテンツ (ウェブページ)，タグ，時刻という 4 つ組でユーザのブックマーク情報を保持している．

ソーシャルブックマークを対象にした研究は，近年盛んに行われており，特にユーザのタグ履歴を用いたソーシャルブックマーク独自の様々な情報検索手法が提案されている [Wu 06][Bao 07][Yanbe 07][Heymann 07]．しかしながら，それら研究の多くは，ユーザにブックマークされたコンテンツを対象にしているため，検索対象がウェブ全体におけるごく僅かなコンテンツに絞られてしまう．

本研究では，ソーシャルブックマークにおいて，コンテンツに付与されたタグをクエリと見なし，ブックマーク件数がより多いものを検索結果上位とする正解データを用いることで，既存の順序付き機械学習の枠組みでランキング関数の学習を行う手法を提案する．これにより，全てのウェブページを対象にしたソーシャルブックマーク件数に基づくランキングを可能とする．

提案手法によって，ソーシャルブックマーク件数によるランキングを正しく近似しているかを従来手法との比較実験を行い，提案手法の有効性について検証を行った．

2. ランキング関数の学習

本説ではランキング関数について述べた後に，本研究で用いる Ranking SVM によるランキング関数の学習について概要を述べる．

一般的なウェブ検索システムでは，ユーザによってクエリが入力された際，クエリが含まれるウェブページを対象に，当該ページの重要度と，クエリに対するページの適合度の 2 種類のスコアを算出し，これらのスコアを足し合わせて得られる検索スコアによって検索ランキングを実現している．前者のスコアはクエリ非依存スコア，後者はクエリ依存スコアとも呼ばれる．具体的には，ページの重要度の算出手法として PageRank [Brin 98]，クエリに対するページの適合度の算出手法としては，TF-IDF や Okapi BM25 [Robertson 94] などが挙げられる．これらのスコアを入力として，ランキングに用いられる検索スコアを出力する関数をランキング関数と呼ぶ．一般的にランキング関数は線形関数であり，入力される各スコアに対して重み付け和を計算する．

次に，Ranking SVM を用いたランキング関数の学習について概要を述べる．先ほど述べた各スコアを素性と見なし，n 個の素性によって表現される特徴空間 $X \in R^n$ において，各事例に対して q が順位を表すラベル $Y = \{r_1, r_2, \ldots, r_q\}$ が付与されているとする．更に $\succ$ が適合度の相対順序を表す順序 $r_q \succ r_{q-1} \succ \cdots \succ r_1$ があたえられた際，以下を満たす関数 $f \in F$ を求める．

$$\vec{x}_i \succ \vec{x}_j \Leftrightarrow f(\vec{x}_i) > f(\vec{x}_j) \tag{1}$$

複数存在する f のうち，与えられた事例集合に対してあらかじめ定められた損失関数を最小化する関数 f^* を選択する．f を線形関数とすると，$f(\vec{x})$ は重みベクトルとの内積 $f_(\vec{x}) = \vec{w}^T\vec{x}$ と表現でき，式 (1) より，

$$\vec{x}_i \succ \vec{x}_j \Leftrightarrow \vec{w}^T(\vec{x}_i - \vec{x}_j) > 0 \tag{2}$$

を得る．$y_1 \succ y_2$ の時 $z = +1$，$y_2 \succ y_1$ の時 $z = -1$ とすれば，スラック変数 ξ_i を導入して二次計画問題に変形することで，SVM と同様に解くことができる．

$$\text{minimize} : \frac{1}{2}||\vec{w}||^2 + C\sum_i \xi_i \tag{3}$$

$$\text{subject to} : \xi_i \geq 0,\ z_i\vec{w}^T(\vec{x}_i^{(1)} - \vec{x}_i^{(2)}) \geq 1 - \xi_i$$

以上より，クエリ・文書ペア集合と，各ペアの適合度の相対順序が与えられれば，Ranking SVM を用いてランキング関数を求めることができる．例えば Joachims は，あるクエリに対す

連絡先: 数原良彦，日本電信電話株式会社 NTT サイバーソリューション研究所，神奈川県横須賀市光の丘 1-1，suhara.yoshihiko@lab.ntt.co.jp

∗1 http://delicious.com/
∗2 http://b.hatena.ne.jp/

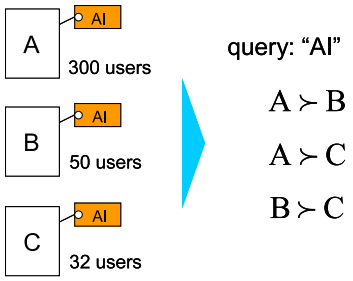

図 1: ソーシャルブックマーク件数に基づく正解順序

る検索結果において，クリック回数がより多いページが相対的に好まれるという仮定に基づいて，クエリ・文書ペアにおける相対順序を定義している [Joachims 02].

3. 提案手法

本研究では，ソーシャルブックマーク件数に基づくランキングを実現するために，コンテンツに付与されたタグをクエリと見なし，コンテンツのブックマーク件数の多い順序でランキングされたものを正解データとして用いる．

クリックログでは，入力されたクエリと，クリックされたページの情報がわかっているが，ソーシャルブックマークサービスにおいて，ユーザがブックマークしているコンテンツにどのような方法で到達したか判断することができない．

そこで我々は，ユーザがコンテンツに付与したタグが，ユーザの情報要求に基づいたものであると考え，ユーザが付与したタグによって当該コンテンツを検索したものと仮定する．

図 1 に提案手法による正解データの生成方法を示す．この図では，タグ “AI” が付与された 3 つのコンテンツをブックマーク件数順に並べている．コンテンツ A は 300 ユーザ，B は 50 ユーザ，C は 32 ユーザによってブックマークされている．この情報から図の右側に示すように，クエリ “AI” によって検索された際に $A \succ B, A \succ C, B \succ C$ という正解順序を生成する．これによってクエリ・文書ペアと，それらの相対順序が得られるため，前節で述べた枠組みでランキング関数の学習を行うことができる．

この際，ブックマーク件数自体を利用しない理由を述べる．[Joachims 02] におけるクリック数と同様，多数のユーザによってブックマークされるコンテンツは，その他多数のユーザの目に触れる機会が多くなり，より多くブックマーク件数を獲得する可能性が高いと考えられ，ブックマーク件数自体ではなく，ランキングの相対順序で扱った方が妥当だと考えたためである．

最初に述べたように，提案手法ではソーシャルブックマークには存在しないウェブページも対象にするため，ページの特徴を表す素性には，ソーシャルブックマーク固有の情報は用いず，一般的なウェブページから抽出可能なものを選択する [Liu 07]．これにより，全てのウェブページを対象にした，ソーシャルブックマーク件数に基づくランキング関数を生成することができる．

4. 評価実験

提案手法の有効性を確認するため，評価実験を行った．

表 1: 実験で使用したデータセットにおける頻出タグ

rank	tag name	tag frequency
1	design	933704
2	blog	662221
3	software	651480
4	tools	629757
5	programming	565651
...	...	...
996	robot	8170
997	ecology	8162
998	colors	8158
999	ia	8155
1000	safety	8130

4.1 データセット

ソーシャルブックマークサービス del.icio.us の 2008 年 11 月時点の 76,357 ユーザについて，ユーザ，コンテンツ，付与したタグのデータを取得した．クエリセットは，データセット中における頻出タグ上位 1,000 件を選択した．表 1 に実験で使用したクエリ (タグ) とデータセットにおけるタグ頻度を示す．各タグが付与されているコンテンツについて，それぞれタグが付与されている件数を集計し，タグ付与件数で降順に並び替え，上位 100 件のコンテンツを評価データ候補とした．検索対象は，全文検索システムにインデクスされることを想定しているため，バイナリファイルを除外し，本文に少なくともタグを 1 個を含むページに限定し，ウェブページの収集を行った．1,000 個のタグに対して合計 46,427 件のウェブページを取得することができた．

4.2 実験

ランキング関数の学習アルゴリズムには，Ranking SVM を選択し，実装には svm_rank [*3] を用いた．Ranking SVM のカーネルには線形カーネルを用いた．

提案手法で用いた素性は以下のとおりである．

1. クエリ依存の素性
   - query_in_title: タイトルにクエリを含む
   - bm25: Okapi BM25 スコア (BM25)
   - log_bm25: bm25 の対数
   - raw_tf: クエリの出現頻度 (TF) ($freq(q,d)$)
   - norm_tf: 文書長正規化 TF $\left(\frac{freq(q,d)}{length(d)}\right)$
   - log_norm_tf: norm_tf の対数
   - idf: $\left(\log \frac{N}{df(q)}\right)$
   - tf_idf: norm_tf × idf
2. クエリ非依存の素性
   - pagerank_score: PageRank スコア
   - is_index_page: インデクスページである
   - is_cgi: ページが CGI である
   - url_length: URL 長

*3 http://www.cs.cornell.edu/People/tj/svm_light/svm_rank.html

表 2: ランキング手法ごとのケンドールの順位相関係数

method	Kendall's τ
proposed	0.2537
BM25	-0.0019
PageRank	0.1098

- title_length: タイトル長
- link_number: 出力リンク数
- log_link_number: link_number の対数
- insite_link_ratio: サイト内リンクの比率
- outsite_link_ratio: サイト外リンクの比率

BM25 のパラメータは一般的に用いられる $k_1 = 1.5, b = 0.75$ とした．raw_tf における $freq(q, d)$ は，単語 q が文書 d に出現する頻度，norm_tf における $length(d)$ は文書 d の長さ，idf における $df(q)$ は単語 q を含む文書数，N は総文書数を表している．PageRank スコアは，Google Toolbar [*4]を用いて 2009 年 4 月時点のスコアを取得した．インデクスされていないページについてはスコアを求めることができないため，スコアを 0 とした．

link_number はページにおける出力リンクのうち，a タグによるものに限定した．同一サイト内へのリンクと異なるサイトへのリンクを区別し，全リンク中のそれぞれの比率を insite_link_ratio，outsite_link_ratio とした．

PageRank スコアによるランキング (PageRank) ，BM25 によるランキング (BM25) ，提案手法 (proposed) によるランキングの 3 手法について比較を行った．

評価指標は，各手法による検索結果上位 k 件 ($k = 1, 5, 10, 15, 20, 30, 50$) の結果が，ブックマーク件数に基づいた実際のランキングの上位 k 件の結果と一致している割合を表す上位 k 件の一致率，各手法と正解ランキングの相関を表すケンドールの順位相関係数 τ を用いた．なお，提案手法については正解データを 10 分割し，うち 9 個を訓練データ，残り 1 個を評価データとして学習と評価を行い，異なる組合せについて 10 回繰り返す 10 点交差法を用いることで，データセット全てのクエリについて評価を行った．

4.3 結果

実験によって得られたケンドールの順位相関係数の結果を表 2，上位 k 件の一致率を図 2 に示す．

表 2 において，提案手法によるランキングがソーシャルブックマーク件数によるランキングと最も高い相関を示している．一方，BM25 によるランキングは，ほぼ無相関であることがわかる．

図 2 では，各手法の上位 k 件における一致率が描かれている．$k = 10$ において提案手法は 0.4 程度の値を示していることから，提案手法によるランキングの上位 10 件中 4 件は，ブックマーク件数によるランキングの上位 10 件に含まれることがわかる．この図より，提案手法が全ての k 件において，最も高い一致率を示していることがわかる．

データセット全てを用いて学習を行った際の各素性の重みを表 3 に示す．素性の重みの絶対値が大きいものは，最終的なランキングに与える影響が大きいことを意味している．正の重みを持つ素性が正の値，負の重みを持つ素性が負の値を取るほど，検索スコアが高くなる．表 3 より，最も影響の大きい素性 is_index_page と pagerank_score である．また，query_in_title と bm25，log_tf が大きい負の重みを持っていることから，これらの素性が 0 または，小さな値を持つ際にページが高いランキングをされることがわかる．raw_tf や norm_tf に対する重みがほぼ 0 であることから，ソーシャルブックマーク件数におけるランキングには，単純な単語頻度が意味を持たないと考えられる．

次に，実験結果においてケンドールの順位相関係数が高いクエリ，低いクエリと，データセットにおけるこれらのクエリの平均ブックマーク数を表 4 に示す．表 4 における τ と平均ブックマーク数の相関係数は 0.300 であり，平均ブックマーク数が多いタグ (クエリ) の方がより高い精度でランキング関数を学習できることがうかがえる．

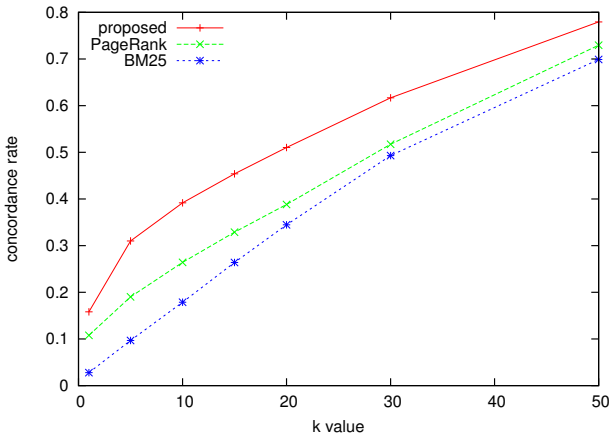

図 2: 上位 k 件における一致率 (proposed, PageRank, BM25)

表 3: 全データ学習時の各素性の重み

weight	feature
0.23701	is_index_page
0.23608	pagerank_score
0.09318	log_link_number
0.00133	title_length
0.00064	raw_tf
0.00004	link_number
-0.00016	norm_tf
-0.00018	log_norm_tf
-0.00269	url_length
-0.00784	insite_link_ratio
-0.00895	outsite_link_ratio
-0.01394	tf_idf
-0.02160	is_cgi
-0.02425	log_bm25
-0.11032	log_tf
-0.14442	bm25
-0.20303	query_in_title

*4 http://toolbar.google.com

表 4: クエリ毎のケンドールの順位相関係数

tag name	Kendall's τ	ave. bm. num.
temp	1.000	85.7
humour	1.000	207.7
googlemaps	0.733	110.5
projectmanagement	0.667	138.8
opensource	0.667	374.1
...	...	...
China	-0.103	77.0
military	-0.104	37.6
pda	-0.126	132.1
rubyonrails	-0.170	235.5
forex	-0.201	82.5

5. 関連研究

教師付き機械学習を用いたランキング関数の学習については，近年盛んに研究されており [Liu 07]，Microsoft Research によって順序付き機械学習の評価用データセットである LETOR Dataset や各手法のベンチマーク情報などが公開されている*5．

評価実験で用いた Ranking SVM は，Herbrich らによって提案された [Herbrich 00]．

Joachims は，クリックログデータを元に，Ranking SVM を用いて検索結果の再ランキング手法を提案している [Joachims 02]．

Bao らは，ソーシャルブックマークにおけるユーザ，コンテンツ，タグの 3 部グラフ構造に PageRank アルゴリズムを適用し，ユーザ，コンテンツ，タグの重要度を算出する SocialPageRank と，クエリとタグの類似度をモデル化した SocialSimRank を提案し，これを素性として Ranking SVM を用いてランキング関数を学習している [Bao 07]．

また，Richardson らは，ランキング関数の素性として PageRank を利用せずとも，高い検索精度を実現できることを示している [Richardson 06]．

6. おわりに

本研究では，ソーシャルブックマークサービスにおいて，付与されたタグをクエリ，ブックマーク件数による順序を正解と見なすことで，順序付き機械学習の枠組みでランキング関数の学習を行う手法を提案した．

評価実験より，提案手法がベースライン手法である BM25 に基づくランキング，PageRank に基づくランキングに比べ，上位 k 件における一致率，ケンドールの順位相関係数において，より正確にソーシャルブックマーク件数によるランキングを近似していることが示された．

しかしながら，提案手法によってソーシャルブックマーク件数に基づくランキングを十分に近似出来ているとは言い難い．今後はまず，使用する素性を検討とすると共に，タグ付与件数に含まれる多数ユーザの異なる観点を考慮するなど，正解データ構築の工夫を行うことで，より正確にソーシャルブックマーク件数によるランキングの近似を検討したいと考えている．また本手法について，順序相関などの近似精度だけではなく，一般ウェブページを対象にした検索結果についてユーザによる適合性評価を行い，有効性を検証したい．

*5 http://research.microsoft.com/en-us/um/beijing/projects/letor/

参考文献

[Bao 07] Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., and Yu, Y.: Optimizing web search using social annotations, in *Proceedings of the 16th international conference on World Wide Web (WWW '07)*, pp. 501–510 (2007)

[Brin 98] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, in *Proceedings of the seventh international conference on World Wide Web 7 (WWW7)*, pp. 107–117 (1998)

[Herbrich 00] Herbrich, R., Graepel, T., and Obermayer, K.: Large Margin Rank Boundaries for Ordinal Regression, in *Advances in Large Margin Classifiers*, pp. 115–132, Cambridge, MA (2000), MIT Press

[Heymann 07] Heymann, P., Koutrika, G., and Garcia-Molina, H.: Can social bookmarking improve web search?, in *Proceedings of the international conference on Web search and web data mining (WSDM '08)* (2007)

[Joachims 02] Joachims, T.: Optimizing search engines using clickthrough data, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02)*, pp. 133–142 (2002)

[Liu 07] Liu, T.-Y., Qin, T., Xu, J., Xiong, W., and Li, H.: LETOR: Benchmark dataset for research on learning to rank for information retrieval, in *LR4IR 2007, in conjunction with SIGIR 2007* (2007)

[Richardson 06] Richardson, M., Prakash, A., and Brill, E.: Beyond PageRank: machine learning for static ranking, in *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, pp. 707–715 (2006)

[Robertson 94] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M.: Okapi at TREC-3, in *Proceedings of the Third Text REtrieval Conference (TREC 1994)* (1994)

[Wu 06] Wu, X., Zhang, L., and Yu, Y.: Exploring social annotations for the semantic web, in *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, pp. 417–426 (2006)

[Yanbe 07] Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K.: Can social bookmarking enhance search in the web?, in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07)*, pp. 107–116 (2007)