2A1-04

素性推定器を用いたランキング学習

Learning to Rank Based on Feature Estimation

数原 良彦[*1] 片岡 良治[*1]
Yoshihiko Suhara 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

Learning to rank has attracted great attention in information retrieval. Incorporating features extracted from click-through logs has been demonstrated to significantly improve the performance of ranking functions. However, many queries and documents have no clicks because of ranking bias so that learning-to-rank algorithms cannot rely strongly on features extracted from click-through logs. In this paper, we propose the learning-to-rank framework using feature estimators to solve this problem.

1. はじめに

ウェブ検索において，ユーザに適切な検索結果を提示するためのランキングは重要な要因であり，より高精度なランキング実現を目指して様々な手法が開発されてきた．

一般的なウェブ検索システムでは，ユーザに入力されたクエリについて，検索対象のウェブ文書群を保持したインデクスに対して検索処理を行い，クエリを含む文書に対して検索スコアを算出し，検索スコアの順に検索結果を提示することでランキングを実現する．具体的には，入力されたクエリと文書の類似度である BM25 スコア [Robertson 94] や，リンク解析に基づいたページ重要度である PageRank スコア [Brin 98] などのランキング素性 (以下，素性) を多数利用し，これらの素性を入力とするランキング関数によって検索スコアを算出する．このため，適切なランキング関数を用意することは検索ランキングにおいて重要な課題である．

最近では，教師あり機械学習を用いてランキング関数を生成するランキング学習 (leraning to rank) が盛んに研究されている．既存のランキング学習手法では，訓練データとして，人手による適合性評価やクリックログ (click-through log) が用いられてきた．人手による適合性評価は，クエリに対する検索結果に対して，検索意図に適合するかを複数の評価者によって付与したものである．ランキング学習では，人手による適合性評価を正解ラベルとして用いてランキング関数の最適化を行う．クリックログは，ユーザの入力クエリと検索結果に対するクリック履歴に関する検索システムのログである．ランキング学習におけるクリックログの利用方法は，(1) 素性として用いる方法 [Agichtein 06, Dupret 10, Gao 09] と， (2) 適合性評価として用いる方法 [Dou 08, Joachims 02] の 2 種類に分けられる．本稿では，(1) のクリックログを素性として用いるアプローチを扱う．先行研究により，クリックログを素性として用いることで高性能なランキング関数を生成可能なことが報告されている [Agichtein 06, Gao 09]

クリックログを素性として用いる場合，ランキング上位にクリックが偏るランキングバイアスや，クエリ頻度の小さいクエリには相対的にクリック数が少なくなる問題があり，クリックログの補正を行うことでより高精度なランキング関数が生成可能なことが報告されている [Gao 09]．また，クリックログを素性として利用したランキング学習を行うためには，人手による適合性評価が付与されたウェブページに対してクリックログが付与されている必要がある．このため新規に追加されたページや，クエリ頻度の低いクエリに対する検索結果には，クリックログを用いた素性を有効に用いることができない問題が挙げられる．

本稿では，このような性質を持つ素性を「不完全な素性 (incomplete feature)」と呼ぶ．ここで不完全な素性とは，クリックログのように，値に偏りが存在し，場合によっては一切付与されていないような素性のことを表す．なお，クリックログの場合には，値が適切に付与されているかどうかを判別することが困難ではあるものの，与えられた値によって学習に一定の効果が得られることがわかっている．

我々は，クリックログのような全てのページに付与されない素性を目標値とする素性推定器を生成し，この推定器の推定値を素性として加えた訓練データを用いてランキング学習を行う方法を提案する．クリックログの場合，素性推定器の学習には正解ラベルである人手による適合性評価が不要なため，クリックログが付与されたデータを素性推定器の学習に用いることができる．このため，本手法はラベルなしデータを利用した半教師あり学習と考えることもできる．

我々は以下の 2 つの検証を通じて提案手法の有効性を明らかにする．

1. 不完全な素性が全く与えられなかった場合において，素性推定器を用いることによって検索精度を向上可能なことを検証する．
2. 不完全な素性がある程度付与されている場合においても，大量のデータを用いて学習した素性推定器の推定値を用いることで，精度向上が可能であることを検証する．

本稿では，推定対象の素性をクリックログに限定し，評価実験を通じて提案手法の有効性を検証する．評価実験では，商用モバイルウェブ検索データとクリックログを用いた評価実験の結果より，素性推定器を用いたランキング学習による検索精度向上の効果を検証し，その結果を報告する．

2. ランキング学習

本節では教師あり機械学習を用いたランキング学習の概要について述べる．ランキング学習手法は，適合性評価が付与さ

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

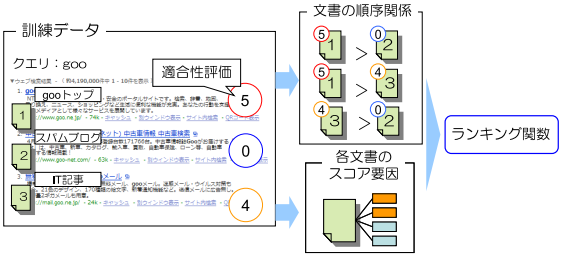

図 1: 人手による適合性評価を用いたランキング学習

れた訓練データに対してどのような目標関数を最適化するかという観点で分類される．本節では，適合性評価点数を順序に落とし込み，文書の順序誤りを最小化するペアワイズ手法 [Joachims 02] の例を用いて解説を行う．

図 1 に学習に用いる訓練データと，ランキング学習の処理の概要を示す．訓練データの例を図の左側に示す．この例ではクエリ goo に対して得られた 3 件の検索結果に対して，検索結果がクエリが表す検索意図に適合しているか，人手による適合性評価が 6 段階評価で付与されている．点数が高いほど，高い適合性を表す．一般的には評価の偏りを軽減するために，適合性評価は複数の評価者によって付与される．クエリ毎にユーザの検索意図が異なるため，同じ文書でも，クエリによって適合性評価が異なり，適合性評価はクエリ毎に作成されることに注意する．

この適合性評価を元に，評価点数に従って文書の順序関係を抽出する．同時に適合性評価が付与された文書それぞれについて，クエリ依存の素性，クエリ非依存の素性の抽出を行う．

ペアワイズ手法によるランキング学習では，適合性評価から得られた順序関係に対して順序の誤りを損失関数として設定し，訓練データに対して損失関数を最小とするランキング関数を生成する．直感的な解釈では，生成されたランキング関数は，これらの順序関係をできるだけ満たすように各素性の重みが設定されたものと捉えることができる．

ランキング学習では，大量のクエリの適合性評価が付与された訓練データを用いて学習を行うことで，未知のクエリに対しても高性能なランキング関数を生成することを目指す．このように，クエリに対する検索結果集合に適合性を表す指標が付与されていれば，ランキング学習が可能であることがわかる．

2.1 素性推定器を用いたランキング学習

本稿では，クリックログのように全ての文書に付与されない素性について，ランキング学習に用いる訓練データ以外のデータを用いて素性推定器の学習を行い，出力した推定値を用いて最終的なランキング学習を行う手法を提案する．

提案手法の処理の流れを図 2 にしたがって述べる．

(1) 素性推定器の学習に用いる訓練データの素性抽出を行う．この際，人手による適合性評価が付与された訓練データと素性空間が同一，または部分空間である必要がある．
(2) クリックログと抽出した素性を用いて，素性推定器の学習を行う．クエリ依存の素性が存在するため，通常のランキング学習手法をそのまま適用可能である．
(3) 人手による適合性評価が付与された訓練データの素性抽出を行う．
(4) (3) で得られた訓練データを入力とし，素性推定器を用いて推定値を出力する．出力された推定値を訓練データの新たな素性として追加する．

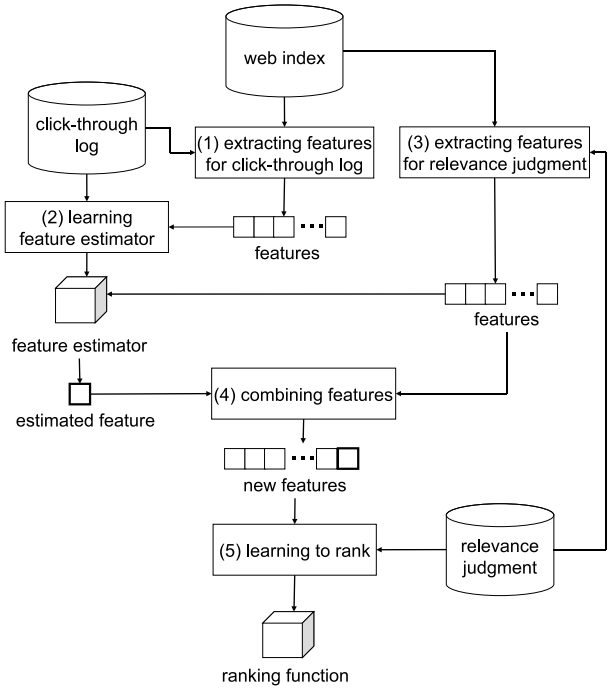

図 2: 素性推定器を用いたランキング学習の概要

表 1: 実験に用いたデータセットの概要

dataset	#query	#doc/#query
relevance dataset	48	331.9
click dataset	5,000	21.8

(5) 作成された新しい素性空間を持つ訓練データを元に，ランキング学習を行う．

人手による適合性評価の作成はコストが高いため，大量に作成することが難しい．一方でクリックログは比較的容易に格納しておくことが可能であり，人手による適合性評価が付与されていないクエリは，素性推定器の訓練データとして用いることができるため，大量のラベルなしデータを活用してランキング学習を行うことができる．このように我々の手法は，人手による適合性評価が付与されていない大量のラベルなしデータを用いた半教師ありランキング学習と見なすことができる．

3. 評価

提案手法の有効性を検証するために評価実験を行った．

3.1 データセット

評価実験には，商用モバイル検索エンジンの検索結果に対する人手による適合性評価と 3ヶ月分のクリックログを用いた．データセットの概要を表 1 に示す．relevance dataset はランキング学習に用いる人手による適合性評価の訓練データ，click dataset は素性推定器生成に用いる訓練データを表している．

人手による適合性評価は，商用モバイル検索エンジンのクエリログを元に頻度の高いクエリから選択した 48 件のクエリについて，約 300 件の検索結果を被験者 3 人が 4 段階の評価 (非常に適合，適合，部分適合，不適合) を付与したデータを用いた．本実験では，この 3 人の被験者の評価平均を適合性評価と見なし，小数点が発生する場合は四捨五入して 4 段階

に変換した．クリックログの素性として，当該訓練データに含まれるクエリ-文書ペアに対する 3ヶ月間のクリック数の合計を用いた．

素性推定器の学習に用いる訓練データは，クリックログの中から人手による適合性評価データに含まれないクエリを 5,000 件選択し，クリック数の上位約 20 件を訓練データとして用意した．この際，クリック回数をそのまま正解ラベルとした．

3.2 実験条件

データセットをクエリによって 5 分割し，そのうち 3 組を訓練データ，1 組を検証用データ，残りの 1 組をテストデータとして用いる 5-fold cross validation によって評価を行った．

ランキング学習手法には，既存手法である RankingSVM [Joachims 02] を用いた．RankingSVM の実装には `svm_rank` [*1] を用いた．実験では線形カーネルを用いた．また，マージンと訓練誤差のトレードオフパラメータ $c \in \{0.01, 0.001, 0.0001\}$ から検証用データにおいて順序誤り誤差を最小化する値を選択した．

本実験では，素性推定器の効果を明確に計測するために，ベースの素性としては BM25 スコアのみを用いた．比較のために BM25 スコアのみによるランキング (bm25) の評価も行った．

ベースラインと提案手法では BM25 スコアとそれぞれの素性を追加した訓練データを用いた．ベースライン手法としては，当該クエリにおけるクリック数を素性として扱う手法 (click_plain) と，クリック数に Good-Turing 法による頻度スムージング [Gao 09] を適用した手法 (click_smooth) を用いた．

クリックログが全く記録されていない状況において，素性推定器による推定値付与による精度向上を検証するために，BM25 スコアと素性推定器の推定値を素性とした手法 (proposed) を用意した．また，クリックログが記録されたクエリ-文書についても，素性推定器の効果があるかを推定するため，click_plain に素性推定器の推定値を追加した手法 (proposed + click_plain) を用意した．クリックログの素性推定器には RankingSVM を利用した．

3.3 評価指標

生成されたランキング関数の有効性を検証するため，情報検索において評価指標として広く用いられている Normalized Discounted Cumulative Gain (NDCG) [Järvelin 02] を用いて評価を行った．NDCG は NDCG@k のように各検索順位における評価指標として用いられ，検索結果上位 k 件において，理想的なランキングへの近さを表す評価指標と解釈できる．クエリ q に対する検索結果 k 位における DCG の値は，

$$\mathrm{DCG}_q@k = rel_1^q + \sum_{i=2}^{k} \frac{rel_i^q}{\log_2 i} \tag{1}$$

によって計算される．ここで rel_k^q は，クエリ q における k 番目の順位の適合性評価点数を表している．DCG@k は，順位が下がるにつれて対数によって重みを減衰させた評価点数の重み付け和と考えることができる．NDCG は，$\mathrm{DCG}_q@k$ を用いて

$$\mathrm{NDCG}@k = \frac{1}{M} \sum_{q=1}^{M} \frac{\mathrm{DCG}_q@k}{\mathrm{IDCG}_q@k} \tag{2}$$

によって計算される．$\mathrm{IDCG}_q@k$ は，クエリ q において評価点数の順番に並べられた理想的なランキングにおける $\mathrm{DCG}_q@k$ の値を表しており，理想的なランキングにおける NDCG 値が 1 になるように正規化を行っている．NDCG@k は，評価に用いられた M 件のクエリに対する $\mathrm{NDCG}_q@k$ の平均によって計算される．モバイルや PC ウェブ検索エンジンでは，検索結果を 5 件または 10 件表示するものが一般的であるため，評価実験では特に $k = 5, 10$ の NDCG 値に注目する．

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

表 2: 実験結果

method	NDCG@5	NDCG@10
bm25	.5447	.5544
click_plain	.5710	.5748
click_smooth	.5569	.5685
proposed	.5657	.5730
proposed + click_plain	**.5807**	**.5920**

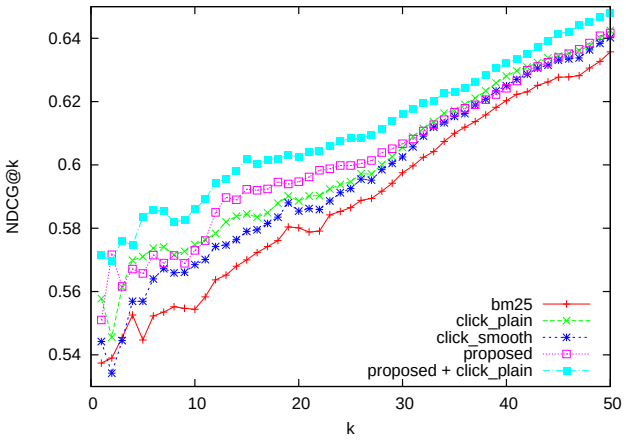

図 3: 各手法の NDCG 値

3.4 結果

実験結果を表 2 に示す．表 2 より以下の結果が得られた．

- bm25 に比べてクリックログを用いたベースライン手法が高い NDCG 値を示した．
- click_plain に比べて click_smooth が低い値を示した．
- proposed は，bm25 より高い NDCG 値を示したものの，ベースライン手法より低い値を示した．
- proposed + click_plain が，全ての手法に比べて高い NDCG 値を示した．

また，各手法について $k = 1, \ldots, 50$ の NDCG 値を示したものを図 3 に示す．図 3 より得られた結果を以下に示す．

- proposed + click_plain により，$k = 1, \ldots, 50$ において，ベースライン手法よりも高い値を示した．
- proposed が $k = 2, 15, \ldots, 30$ において，ベースライン手法に比べて高い値を示した．

3.5 考察

bm25 以外の手法が bm25 に比べて高い精度を示していることから，先行研究同様，クリックログを用いた素性の追加により，ランキング精度を向上可能であることを確認した．

click_smooth が click_plain よりも低い精度を示したのには 2 つの理由が考えられる．1 つ目は，クリック数に対するスムージングが適切に行われなかった可能性が挙げられる．この場合，スムージング手法を改善することによって解消されると考

えられる．2 つ目は，そもそもスムージングの必要がない可能性である．スムージングを適用することにより，真にクリック回数が 0 回のクエリ-文書ペアに対して，不適切にクリック回数を付与してしまったことが考えられる．

proposed が bm25 に比べて高い精度を示していることから，クリックログが全く付与されない場合において，大量のラベルなしデータを用いて素性推定器を生成することにより，精度向上が可能であることが検証された．

proposed + click_plain が click_plain に比べて高い精度を示していることから，クリック回数という素性に，異なる情報源から得られたクリック数推定値を追加することで，より高精度なランキング学習が実現可能であることを検証された．この結果より，素性推定器を利用することで，異なる情報源に含まれる情報を適切に抽出し，利用できることが示された．

4. 関連研究

クリックログを素性として用いるランキング学習は多数存在する．[Agichtein 06, Dupret 10] などは，ユーザの検索行動のモデル化を通じてクリックログから適切な情報の抽出を試みている．一方で，Gao ら [Gao 09] の研究では，クリックログはランキングバイアスやクエリ頻度の問題があるため，不完全な素性とみなし，これに対する補正を通じて検索精度向上が可能であると報告している．

半教師あり機械学習を用いたランキング学習手法も多数提案されており [Duh 08, Jin 08, Li 09]，様々な方法によってラベルなしデータを活用している．我々の知る限り，素性推定器を用いたランキング学習手法は存在しない．

素性推定器を用いた半教師あり学習という点では，我々の手法は Ando ら [Ando 05] と類似している．Ando らは，補助問題を定義し，あらかじめ定義した補助問題の学習によって，補助問題から対象課題に適した素性を得る方法を提案し，テキストのチャンキング課題に適用している．

不完全な素性を扱う場合には，機械学習における欠損値推定の問題 [Lakshminarayan 96] として解くことも考えられる．しかしながら，クリックログの場合には，クエリ頻度と検索結果に対するクリック数が著しく偏っているため，閲覧されてクリックされなかった場合 (クリック数が 0) と，ユーザに提示されていないためクリックを得ることができなかった場合 (欠損値) の区別が困難である．そのため何を欠損値として置き換えて，何をそのままの値を用いるか判断が困難であり，不完全な素性を扱う課題を欠損値推定の問題としては解くことは難しいと考えている．

5. おわりに

本稿では，クリックログのような不完全な素性に対して，大量のラベルなしデータを用いて素性推定器を学習し，素性推定器の推定値を素性として用いるランキング学習手法を提案した．

商用モバイル検索システムの実データを用いた評価実験を通じて，提案手法の有効性を検証した．本研究の貢献は，以下のとおりである．

- ランキング学習に用いる訓練データ以外のデータを用いて素性推定器を学習し，素性推定器による推定値を素性として用いるランキング学習手法を提案した．
- 不完全な素性が一切付与されていない文書に対して素性推定器による推定値を素性として用いることで検索精度を向上することを検証した．
- 不完全な素性がある程度付与されている場合でも，素性推定器による推定値によって更なる精度向上が可能であることを検証した．

今後は，クリックログ以外の素性を対象に素性推定器を生成することによる精度向上を検討したい．クリックログ以外の素性推定器を同時に用いることで，更なる精度向上が可能であると考えている．ランキング関数は検索精度に直接影響を与えるため，ランキング学習は実用の観点からも重要の課題と考えている．今後も引き続き多様な情報源を用いたランキング学習の研究に取り組みたい．

参考文献

[Agichtein 06] Agichtein, E., Brill, E., and Dumais, S.: Improving web search ranking by incorporating user behavior information, in *Proc. SIGIR '06*, pp. 19–26 (2006)

[Ando 05] Ando, R. K. and Zhang, T.: A high-performance semi-supervised learning method for text chunking, in *Proc. ACL '05*, pp. 1–9 (2005)

[Brin 98] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, in *Proc. WWW7*, pp. 107–117 (1998)

[Dou 08] Dou, Z., Song, R., Yuan, X., and Wen, J.-R.: Are clickthrough data adequate for learning web search rankings?, in *Proc. CIKM '08*, pp. 73–82 (2008)

[Duh 08] Duh, K. and Kirchhoff, K.: Learning to rank with partially-labeled data, in *Proc. SIGIR '08*, pp. 251–258 (2008)

[Dupret 10] Dupret, G. and Liao, C.: A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine, in *Proc. WSDM '10*, pp. 181–190 (2010)

[Gao 09] Gao, J., Yuan, W., Li, X., Deng, K., and Nie, J.-Y.: Smoothing clickthrough data for web search ranking, in *Proc. SIGIR '09*, pp. 355–362 (2009)

[Järvelin 02] Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques, *ACM Trans. Inf. Syst.*, Vol. 20, No. 4, pp. 422–446 (2002)

[Jin 08] Jin, R., Valizadegan, H., and Li, H.: Ranking refinement and its application to information retrieval, in *Proc. WWW '08*, pp. 397–406 (2008)

[Joachims 02] Joachims, T.: Optimizing search engines using clickthrough data, in *Proc. KDD '02*, pp. 133–142 (2002)

[Lakshminarayan 96] Lakshminarayan, K., Harp, S. A., Goldman, R. P., and Samad, T.: Imputation of Missing Data Using Machine Learning Techniques, in *Proc. KDD-96*, pp. 140–145 (1996)

[Li 09] Li, M., Li, H., and Zhou, Z.-H.: Semi-supervised document retrieval, *Inf. Process. Manage.*, Vol. 45, No. 3, pp. 341–355 (2009)

[Robertson 94] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M.: Okapi at TREC-3, in *Proc. TREC-3*, pp. 109–126 (1994)