2O4-3

クリックログの正解尤度の推定と検索への適用

Estimation of correct likelihood of click log and its application to search

別所 克人[*1] Katsuji Bessho
大塚 淳史[*1] Atsushi Otsuka
西田 京介[*1] Kyosuke Nishida
浅野 久子[*1] Hisako Asano
松尾 義博[*1] Yoshihiro Matsuo

[*1] 日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation

The click log accumulated in the retrieval system is useful for improving retrieval accuracy, but the clicked document is not necessarily correct answer for the retrieval query. In this paper, we propose a method to estimate the likelihood that the clicked document is correct for the retrieval query in the click log. We also show that retrieval accuracy improves by extending the retrieval target document with the corresponding retrieval query (with the estimated correct likelihood) in the click log.

1. はじめに

FAQ 検索等、検索対象文書数が比較的制限されている検索においては、入力キーワードの文字列一致検索では、検索クエリと意味的に適合しているにも関わらず入力キーワードを含んでいない文書が検索されず、意味的に適合する文書が十分得られないことが多発する。単語分散表現の一つである単語概念ベクトルを用いて、検索対象文書とクエリをともにベクトルで表現し、ベクトル間のコサイン類似度をとることにより検索する概念検索[Schutze 94][別所 08]は、入力キーワードの有無に関わらず、意味的に適合している文書を検索できるため、上記課題を解決するのに有用である。

しかしながら、意味的に適合しても表現に大きな乖離があるクエリと文書(例;クエリ:「クレカで決済できる？」, 文書:「安心サービスの支払い方法を教えてください。」)に対しては、ベクトル間の類似度は不当に低くなり、検索精度に問題があった。これを解決するための方法の一つとして、検索の事前処理において、文書に、対応するクエリを付加する文書拡張法が有効と推察される。検索時に、付加したクエリと意味的に類似した新規のクエリ(例;上記例で「カードで払える？」)が入力されると、新規クエリのベクトルと、拡張後の文書のベクトルとは、値が近くなるからである。

各検索対象文書に対し、対応するクエリのリストを作成するのは多大なコストを要する。検索システムにおいて蓄積されるクリックログには、ユーザが入力したクエリと、その検索結果において、ユーザがクリックした文書の情報が含まれている。クリックした文書が、対応するクエリの正解文書であると見做し、該文書に該クエリを付加するようにすれば、クエリ作成コストを無くすことができる。このようにコスト削減のために、クリックログを活用することは有効と考えられる。だが、ユーザがクリックする文書が、必ずしもクエリと意味的に適合しているとは限らない。ノイズとなるクエリを付加することにより、精度上の課題も発生すると考えられる。

本稿の一つの目的は、概念検索において、各検索対象文書を、クリック関係にあるクエリで拡張することによる検索精度の変化を検証することである。本稿のもう一つの目的は、クリックログにおけるクエリとクリック文書との対応関係が正解である(意味的に適合している)尤度を推定する方式を提案し、該方式を用いた、対応関係が正解であるか否かの 2 値分類の精度を検証する。そして、各検索対象文書を、クリック関係にあるクエリで、推定した正解尤度付きで拡張する方式を提案し、その検索精度を検証する。

以下、2 節で関連研究、3 節で提案手法を述べる。そして 4 節で評価実験を示し、5 節でまとめを述べる。

連絡先:別所克人, 日本電信電話株式会社 NTT メディアインテリジェンス研究所, 神奈川県横須賀市光の丘 1-1, bessho.katsuji@lab.ntt.co.jp

2. 関連研究

Agichtein らは、クリックログにおける文書のクリック回数、文書閲覧時間、クエリと文書との共有単語数等の素性を抽出し、ランキング学習を行う手法を提案している[Agichtein 06]。本稿の提案手法では、それらとは異なる素性を使用する。

Xue らは、各検索対象文書を、クリックログにおける対応クエリで拡張する文書拡張の手法を提案している[Xue 04]。Xue らの手法では、同一の文書に対応づいているクエリ同士の類似性と、同一のクエリに対応づいている文書同士の類似性により、直接対応づいていないクエリ・文書間の関連性を導出する。本稿の提案手法もクリックログを用いた文書拡張の手法であるが、クリック関係にあるクエリ・文書ペアに対する正解尤度を導出するものであり、また、その導出過程は異なっている。

Ricard は、クリックログにおけるクエリを、対応するクリック文書のタームを用いてベクトル化し、クエリ集合をクラスタリングする。クエリとその対応するクリック文書とのスコアを、該クエリが属するクラスタ内のクエリと該文書との間のクリック回数分増大させる[Ricard 05]。後述する本稿の提案手法も、対象クエリの近傍内の他のクエリの存在を考慮する点は共通しているが、クエリベクトルやスコア算出過程は異なり、また、対象文書の近傍内の他の文書の存在も考慮する点が異なる。

3. 提案手法

3.1 前提条件

本稿の提案手法で前提とする検索システムは、図 1 のように、ユーザが検索クエリを入力すると、その検索結果が提示されるものであり、ユーザは検索結果中で、クエリに該当すると思った文

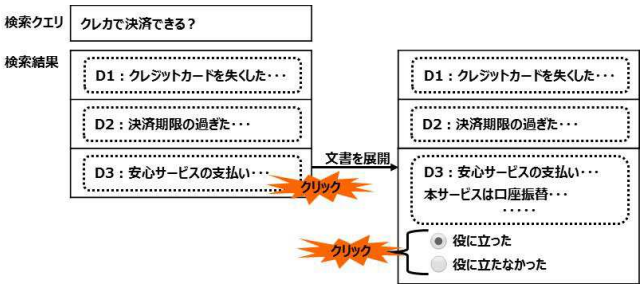

図 1 :検索システムの画面例

●クエリIDの構成情報
・ユーザID
・クエリ発行日時
・クエリ文字列

●文書IDの付属情報
・文書文字列

●クエリID・文書IDのクリックペアの付属情報
・「役に立った」，「役に立たなかった」，「いずれでもない」のいずれか一つ

図 2 :クエリ ID と文書 ID に関する情報

書をクリックする。場合によっては、さらに、クリックした文書が役に立ったか否かを選択するボタンがついていて、ユーザがそれを選択できるものとする。

クリックログ中のクエリはクエリ ID によって一意に識別され、クエリ ID は、図 2 のように、ユーザ ID、クエリ発行日時、クエリ文字列から構成されるものとする(ユーザ ID が取得できない場合は、任意のユーザに対し同一のユーザ ID を割り当てる)。検索対象文書は文書 ID で一意に識別され、各文書 ID に文書文字列が付属する。クリック関係にあるクエリ ID と文書 ID のペアをクリックペアと呼ぶこととすると、一つのクリックペアに対し、役に立ったか否かを選択するボタンの最終的な押下状況(「役に立った」,「役に立たなかった」, 左記の「いずれでもない」)が付属しているものとする。

3.2 正解尤度付きでない文書拡張法

本稿の提案手法における、正解尤度付きでない文書拡張法について述べる。本文書拡張法では、各文書 ID に対し、該文書 ID の文書文字列と、該文書 ID とクリックペアをなすクエリ ID (複数ありえる)のクエリ文字列のそれぞれを、半角空白を間に挟んだ上で連結する。連結後の文字列を単語分割し、その中の各内容語(名詞、動詞、形容詞等)を単語概念ベクトル[別所08]に変換し、単語概念ベクトルを加算し長さ 1 に正規化した概念ベクトルを、該文書 ID の概念ベクトルとする。検索時は、入力文字列を同様の処理で概念ベクトルに変換し、入力概念ベクトルと各文書概念ベクトルとのコサイン類似度を算出し、文書 ID をコサイン類似度の降順にランキングする。

3.3 クリックペアの正解尤度推定

本稿の提案手法における、クリックログ中の各クリックペアが正解である(意味的に適合している)尤度を推定する方式について述べる。各文書 ID、及び、クリックログ中の各クエリ ID に対し、その文字列を 3.2 節と同様の処理で概念ベクトルにマッピングしておく。任意の概念ベクトルを x=(x1,x2,･･･,xn)としたとき、x の半径 r (≧0)の近傍を、{(y1,y2,･･･,yn) | xi – r ≦ yi ≦ xi+r

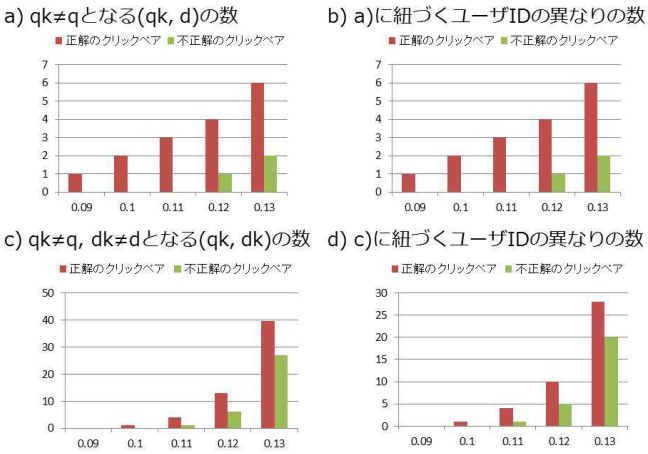

図 3 :仮説の検証結果

1) qk≠qとなる近傍内のクリックペア(qk, d)の数
2) qk≠q, dk≠dとなる近傍内のクリックペア(qk, dk)の数
3) 1)～2)それぞれに対し、該当ペアに紐づくユーザIDの異なりの数
4) 1)～3)それぞれに対し、その値の全体に占める割合
5) 対象クリックペア(q, d)が「役に立った」か否か
6) 対象クリックペア(q, d)が「役に立たなかった」か否か
7) qk≠qとなる近傍内の「役に立った」ペア(qk, d)の数
8) qk≠q, dk≠dとなる近傍内の「役に立った」ペア(qk, dk)の数
9) qk≠qとなる近傍内の「役に立たなかった」ペア(qk, d)の数
10) qk≠q, dk≠dとなる近傍内の「役に立たなかった」ペア(qk, dk)の数
11) 7)～10)それぞれに対し、該当ペアに紐づくユーザIDの異なりの数
12) 7)～11)それぞれに対し、その値の全体に占める割合

図 4 :クリックペア(q,d)の正解尤度推定のための素性

(1≦i≦n)}と定義する。クリックペア(q,d)の近傍に、クリックペア(qk,dk)があるとは、q の概念ベクトル Vq の近傍に qk の概念ベクトル Vqk があり、d の概念ベクトル Vd の近傍に dk の概念ベクトル Vdk があることを意味するものとする。クリックペアに関し、以下の仮説(1)(2)を置く。

・仮説(1):対象クリックペアが正解である場合、近傍内の他のクリックペアは多く、不正解である場合、少ない。

・仮説(2):対象クリックペアが正解である場合、近傍内の他のクリックペアに紐づくユーザの異なりの数は多く、不正解である場合、少ない。

仮説(1)は、対象クリックペアを支持する他のクリックペアが多ければ、対象クリックペアの正解尤度は高いという考えに基づいている。すなわち、対象クエリと意味の近いクエリに対し、対象文書と意味の近い文書をクリックした事例が多ければ、対象事例の正解尤度は高いという考えである。

仮説(2)は、対象クリックペアを支持するユーザが多ければ、対象クリックペアの正解尤度は高いという考えに基づいている。すなわち、近傍内の他のクリックペアが N 個のとき、紐づくユーザの異なりの数が少数の場合(例えば 1 人が N 回、入力とクリックを行ったような場合)は、少数のユーザが恣意的にクリックしている可能性もあり、そのような場合は対象クリックペアの信頼性は低く、逆に、紐づくユーザの異なりの数が多数の場合(例えば N 人が 1 回ずつ、入力とクリックを行ったような場合)は、対象クリックペアの信頼性は高いという考えである。

仮説を検証するため、正解か否かのラベルを付与したクリックペアのリストを用意し、正解の各対象クリックペア(q,d)に対し、その近傍内の他のクリックペアに関する図 3 の a)～d)の数をカウントし、中間値をとった。不正解の対象クリックペアに対しても同

様の中間値をとった。図 3 は、横軸に半径をとり、半径ごとの中間値を示したものである。a)～d)のいずれの値も、正解のクリックペアの方が、不正解のクリックペアより多い傾向があり、仮説は正しい傾向にある。したがって、a)～d)の数は、クリックペアが正解である尤度を決定する因子となりうる。

クリックペア(q,d)の正解尤度を決定しうる因子として、近傍の半径を固定した上で、a)～d)の数を含む図 4 に挙げた素性が考えられる。図 4 の 4)の素性は、クエリ ID 数や文書 ID 数に依存しない比率に相当するもので、4)の「全体」とは、(qk,d)の数に対してはクエリ ID 数を、(qk,dk)の数に対しては、クエリ ID 数×文書 ID 数を、ユーザ ID の異なりの数に対してはユーザ ID 数を指す。5)以降の素性は、クリックペアの付属情報として、「役に立った」,「役に立たなかった」が存在する場合に追加されうるものである。5),7),8)は正解尤度を上げ、6),9),10)は正解尤度を下げると推察される。

提案手法は、これらの素性を用いた機械学習により、クリックペアの正解尤度を推定する。すなわち、学習フェーズにおいて、正解か否かのラベルを付与したクリックペアのリストを入力とし、クリックペアごとに、図 4 中の指定した素性の値を算出し素性ベクトルを生成する。素性ベクトルとラベルの組のリストから、分類モデルを生成する。推定フェーズでは、ラベル無しのクリックペアのリストを入力とし、クリックペアごとに、学習時に使用したのと同じ素性の値を算出して素性ベクトルを生成し、素性ベクトルと分類モデルとから正解尤度を算出する。このように提案手法では、比較的少量のラベル付きデータから分類モデルを生成し、その後、大量のラベル無しデータに対し正解尤度を推定する。

3.4 正解尤度付き文書拡張法

本稿の提案手法における、正解尤度付き文書拡張法について述べる。本文書拡張法では、3.3 節記載の推定フェーズで用いたラベル無しのクリックペアのリストを入力とし、各文書 ID の概念ベクトルを次のようにして生成する。各文書 ID に対し、該文書 ID の文書文字列の重みを 1 とし、該文書 ID とクリックペアをなすクエリ ID のクエリ文字列の重みを該クリックペアの推定正解尤度とする。該文書 ID に対し、重みが閾値 α 以下のクエリ文字列は除外した上で、各文字列中の各内容語の概念ベクトルに、所属する文字列の重みを乗じた概念ベクトルを、全テキストの全内容語にわたって加算し長さ 1 に正規化した概念ベクトルを、該文書 ID の概念ベクトルとする。このようにすることにより、クエリ文字列の概念を正解尤度の分だけ、文書 ID の概念ベクトルに反映させることができる。検索時の処理は、3.2 節と同様である。

4. 評価実験

4.1 節にて評価実験データについて述べた後、4.2 節で正解尤度推定方式の評価実験結果を述べ、4.3 節で正解尤度有り無し双方の文書拡張法の評価実験結果を述べる。

4.1 評価実験データ

スマートフォン関連の 278 個の検索対象文書を用いて、概念検索と BM25 スコアによる検索を組合せた検索システムを構築した。検索結果の各文書には、役に立ったか否かを選択するボタンも表示した。複数のユーザそれぞれにユーザ ID を割り当てた上で、スマートフォン関連のクエリを入力してもらい、検索結果上位 10 件の中から該当する文書をクリックしてもらった。役に立ったか否かを選択するボタンは、押下してもしなくても自由とした。

収集したクリックログにおけるクエリ ID を、正解尤度推定の学習用、推定用と、検索時の検索クエリ用の 3 グループに分割した。但し、同一のユーザ ID を含むクエリ ID は、異なるグループ間で影響し合わないように、同一グループとなるようにした。各グループには、所属するクエリ ID を含むクリックペアのリストが対応づく。

学習用データは、クエリ ID 数:2282, ユーザ ID 数:236, クリックペア数:2172（内、「役に立った」:1590,「役に立たなかった」:255,「いずれでもない」:327）となった。各クリックペアに学習のため、正解か否かのラベルを付与した。この結果、正解数:1210, 不正解数:962 となった。

推定用データは、クエリ ID 数:7991, ユーザ ID 数:826, クリックペア数:8233（内、「役に立った」:5828,「役に立たなかった」:875,「いずれでもない」:1530）となった。各クリックペアに正解尤度推定結果の評価のため、正解か否かのラベルを付与した。この結果、正解数:4492, 不正解数:3741 となった。

検索クエリ用データは、クエリ ID 数:843 となった。各クエリ ID に検索結果の評価のため、全文書 ID から正解の文書 ID を選び出し対応付けた。正解文書 ID ののべ総数は 3424 となった。

単語概念ベクトルは、QA サイト等から取得した、様々なジャンルの約 616 万記事と、スマートフォン関連の約 44 万記事を合わせたコーパスをもとに、[別所 08]の手法を用いて生成した 591437 語の 1000 次元の単語概念ベクトルを使用した。

4.2 正解尤度推定の評価実験

学習用のクリックペアリストで学習処理を行い、推定用のクリックペアリストで推定処理を行った。図 4 の 1)～4)の計 8 個の素性のみを使用する「役に立った」情報を利用しないケースと、図 4 の 1)～12)の計 26 個の素性を使用する「役に立った」情報を利用するケースの双方を評価した。素性値算出にあたり、近傍半径は 0.11 とした。学習及び推定は LIBLINEAR を用いた[NTU 16]。正解尤度を推定した結果、正解尤度 0.5 以下を不正解とした場合の、推定結果の正解率を表 1 に示す。提案方式はいずれのケースも、ベースラインとなる、全クリックペアを正解とする方式より高精度となり、有意水準 1%で有意差を確認した。

表 1 :推定結果の正解率
(*):正解尤度 0.5 以下を不正解とした場合

方式	正解率
全クリックペアを正解とする方式	54.6%
提案手法(「役に立った」情報を利用しない)*	67.3%
提案手法(「役に立った」情報を利用する)*	66.4%

4.3 文書拡張法の評価実験

検索精度評価にあたり、以下の方式を比較対象として採用した。

A)BM25 検索方式:各文書 ID に対し、クエリ中の内容語と該文書 ID との BM25 スコアの総和を、該文書 ID のスコアとしてランキングする方式。

B)通常の概念検索方式:各文書 ID に対し、該文書 ID の文書文字列のみから概念ベクトルを生成する方式。

C)理想方式:各文書 ID に対し、推定用データ中のクエリ ID の中でペアとして正解であるもの全てで該文書 ID の文書文字列を拡張して、該文書 ID の概念ベクトルを生成する方式。

提案方式としては、以下の方式を検証した。

D)正解尤度付きでない文書拡張法(「役に立った」情報を利用しない):推定用のクリックペアリストを用いた 3.2 節記載の方式。

E) 正解尤度付き文書拡張法(「役に立った」情報を利用しない, α=0.4):学習用及び推定用のクリックペアリストを用いた、4.2 節記載の「役に立った」情報を利用しない正解尤度推定と文書拡張を行う方式。文書概念ベクトル生成時の閾値 α を 0.4 とする。

F) 正解尤度付きでない文書拡張法(「役に立った」情報を利用する):推定用のクリックペアで「役に立った」もののみを用いた 3.2 節記載の方式。

G)正解尤度付き文書拡張法(「役に立った」情報を利用する, α=0.3):学習用及び推定用のクリックペアリストを用いた、4.2 節記載の「役に立った」情報を利用する正解尤度推定と文書拡張を行う方式。文書概念ベクトル生成時の閾値 α を 0.3 とする。

E), G)の閾値 α は、α を様々に変化させた上で検索精度が最も高かったときの値である。

検索結果の評価尺度として、N(≧1)を任意に固定し、検索結果上位 N 件中に正解文書 ID を含む検索クエリの割合である正解含有クエリ比率をとる。各方式の 5 以下の N に対する正解含有クエリ比率は、表 2 のようになった。

表 2:各方式の正解含有クエリ比率

方式＼N	1	2	3	4	5
A	43.1%	52.1%	56.0%	58.4%	60.9%
B	44.5%	57.2%	62.6%	65.7%	70.1%
C	63.5%	74.4%	80.2%	83.9%	85.9%
D	56.1%	69.3%	75.4%	79.8%	82.4%
E	56.7%	70.0%	77.1%	80.3%	82.7%
F	57.7%	70.5%	76.9%	80.4%	82.3%
G	57.7%	71.1%	77.2%	80.4%	83.0%

B)は A)より高精度であり、検索対象文書数が比較的制限されている検索においては、キーワードの文字列一致検索よりも概念検索の方が優位である傾向がある。

D)は B)より高精度であり、N=3 のとき 12.8 ポイント向上する。クリック関係にあるクエリで文書拡張することにより、概念検索の精度が大幅に向上するといえる。

E)は D)より精度が微増する。N=3 のとき 1.7 ポイント向上し、有意水準 5%で有意差があった。正解尤度を考慮することにより、検索精度への効果がわずかだがある。

F)は D)より精度がほぼ上回った。推定用のクリックペアの中で正解のものは 4492 個で 54.6%を占めるのに対し、推定用のクリックペアで「役に立った」もののみの中で正解のものは 3780 個で 64.9%を占めていた。正解のクリックペアの個数は D)の方が多いものの、ノイズの割合は F)の方が低いことが、F)の方が優位である原因と考えられる。

G)は F)より精度が同じか微増する。G)は D)同様、推定用のクリックペアを全て使用するが、正解尤度を考慮しているため、このような結果となったと考えられる。

C)は E), G)より高精度であるが、各文書 ID に対し、C)では、E), G)にはないペアとして正解のクエリ ID が追加されていることが精度差になっているものと思われる。C)の精度と 100%の差は、検索クエリデータ中のクエリで、推定用データ中のいずれのクエリとも類似度が低く、推定用データ中のクエリではカバーしきれていないものの割合を意味している。

図 5 は、推定用データ中のクエリ ID を増やしていった場合の、対応するクリックペアリストを用いた方式 E)の正解含有クエリ比率を、N の値別にプロットしたものである。クエリ ID が増えていくにつれ、精度の増加率は下がっていくものの、精度はほぼ単調増加していく。クリックログが増加していくにつれ、クリックログ中のクエリが様々な表現をカバーしていき、漸次的に検索精度が向上していくといえる。

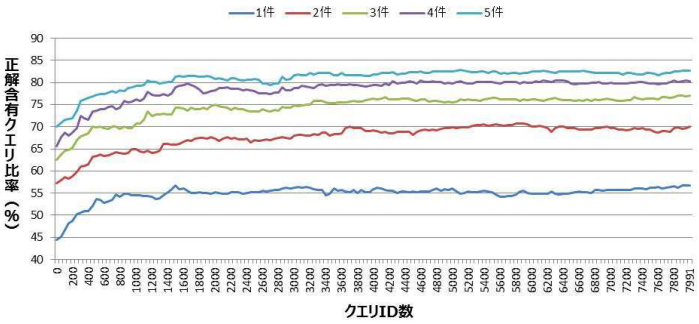

図 5 :クエリ ID 数ごとの正解含有クエリ比率

5. まとめ

本稿では、クリックペアの正解尤度を推定し、検索対象文書を、クリックペアをなすクエリで推定正解尤度付きで拡張することにより、検索精度が微増することを示した。

今後の検討事項として、学習用データから生成した分類モデルを、全く別の検索対象文書と正解尤度推定の推定用データに適用した場合に、どれだけの効果があるかを見ていくことが挙げられる。また現状、文書に対し、クリックペアをなすクエリのみを追加するため、クリック関係にないが、文書と意味的に適合するクエリが文書の概念ベクトルに反映されないという問題がある。クリック関係にないクエリ ID と文書 ID のペアに対しても、適切に正解尤度を推定し、検索に反映させていくことも今後の課題である。

参考文献

[Schutze 94] H. Schutze, and J. Pedersen: A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, Proc. of RIAO'94, pp.266-274, 1994.

[別所 08] 別所克人, 内山俊郎, 内山匡, 片岡良治, 奥雅博: 単語・意味属性間共起に基づくコーパス概念ベースの生成方式, 情報処理学会論文誌, Vol. 49, No. 12, pp.3997-4006, 2008.

[Agichtein 06] E. Agichtein, E. Brill, and S. Dumais: Improving Web Search Ranking by Incorporating User Behavior Information, Proc. of SIGIR, pp.19-26, 2006.

[Xue 04] G. R. Xue, H. J. Zeng, Z. Chen, Y. Yu, W. Y. Ma, W. S. Xi, and W. G. Fan: Optimizing Web Search Using Web Click-through Data, Proc. of CIKM, pp.118-126, 2004.

[Ricard 05] R. Beaza-Yates: Applications of Web Query Mining, Proc. of ECIR'05, pp.7-22, 2005.

[NTU 16] Machine Learning Group at National Taiwan University: https://www.csie.ntu.edu.tw/~cjlin/liblinear/