

混合エキスパートによるブログ検索モデルの統合

Aggregating Blog Retrieval Models by Mixture of Experts

村里 創† 野口 智義† 関 和広† 上原 邦昭†
Sou Murasato Tomoyoshi Noguchi Kazuhiro Seki Kuniaki Uehara

あらまし

ウェブログ (ブログ) の普及にともない, 通常のウェブページを対象にしたウェブ検索モデルではなく, ブログに特化した検索モデル・手法に関する研究が広く行われている. ブログの特徴の 1 つとして, ブログの筆者は通常個人であり, 自身の興味あるいは生活などに関する記事が複数投稿され, 1 つのブログサイトを構成する点が挙げられる. 従来のブログサイト検索 (または「ブログフィード検索」) では, 1 サイトに属する記事の内容, あるいはクエリ適合度をどのように統合するかに着目することでフィード検索を行っている. これに対して本研究では, ユーザの情報要求に適合するブログサイトには何らかの共通した特徴が存在すると仮定し, この特徴を利用した機械学習によるブログフィード検索の枠組を提案する. より具体的には, 複数の分類器を統合する手法であるアンサンブル学習を応用し, 検索モデルを分類器と見なすことでモデルの効果的な統合を行う. 本研究では, 特に *Mixture-of-Experts* を利用することで, ブログサイト個々の特徴に応じた検索モデルの統合を目指す.

1. まえがき

ウェブ上で広く普及しているメディアの 1 つとしてウェブログ (ブログ) がある. ブログは, 個人の主観的意見が多く, かつリアルタイム性のある情報が発信されるという特徴を持っているため, 近年注目されている情報資源である. また, ブログのもう 1 つの特徴としてフィードの存在が挙げられる. フィードとは, 一般的にはブログ記事の概要もしくは全体を配信用に加工した文書であり, RSS や ATOM といったフォーマットがある. ユーザはお気に入りのブログサイトのフィードをフィードリーダーに登録することで, 最新の記事の通知を自動的に得ることができ, タイトル・概要などが興味を惹けば, 実際にブログサイトにアクセスすることでより詳しい情報を得ることができる. 上記のようにフィードリーダーに登録するためなど, 自分の興味に合致するブログサイトを検索によって見つけることを「ブログフィード検索」または「フィード検索」と呼ぶ. なお, フィード検索における「フィード」は, 「ブログサイト」と同義であることが多く, 以下, 本稿でもフィードをブログサイトの意で用いる.

フィード検索に関係する代表的な取り組みとして, Text Retrieval Conference (TREC) が挙げられる. TREC は, 情報検索関連の評価型ワークショップであり, 検索対象やタスクの類型 (検索, 質問応答など) から, いくつかの「トラック」が開設されており, フィード検索はブログトラックで継続的に取り組まれてきた課題である. 従来手法の多くは, 1 つのフィードが複数の記事から構成されることに利用し, それらの記事 (あるいは検索スコア) をどのように

統合するかに着目してきた. これに対して本研究では, ユーザの検索要求を満たすフィードには何らかの共通の特徴があり, またフィードによってそれに適した検索モデルは異なるという 2 つの仮定を基に新しいフィード検索の枠組を提案する.

2. 関連研究

本研究に関連する研究として, まずフィード検索について従来研究を概説する. その後, 本研究で提案する検索モデルの統合に関連して, 従来一般的な検索モデル統合の手法をまとめる.

2.1 フィード検索

従来のウェブ検索とフィード検索との大きな違いは検索の単位にある. すなわち, 前者では, フィード内の一つひとつの記事に対して検索を行うのに対し, 後者では, 複数の記事をまとめたフィードを単位として検索を行う. このことから, 多くの従来手法では, これら複数の記事 (あるいはその検索スコア) を何らかの方法で統合することでフィードを検索することが一般的である.

例えば, Seo と Croft [1] は分散型情報検索における資源選択の考えを応用し, フィード検索を行った. 分散型情報検索では, 複数のデータベースが分散して存在するため, まずユーザのクエリに対して適切なデータベースの選択, すなわち資源選択を行う必要がある. Seo と Croft は, 分散型情報検索におけるデータベースをフィードに置き換えることで, 資源選択の手法を利用した. 一方, Craig ら[2] は専門家検索 (expert search) を利用した手法を提案している. 専門家検索とは組織内の文書を用いて, 大規模な組織における専門家の検索を行うことであり, その一般的な手法は, 各人の専門性をその人の出版物や電子メールに基づいてプロフィールとして表現することである. Craig らは, フィードを専門家と見なし, それに属するブログ記事によってフィードのプロフィールを作成することでフィード検索を行った. また, Elsas ら[3] は, フィードの各記事の文章から *small document* と *large document* と呼ばれる 2 つの文書を構築し, それらを別々に検索して結果を統合することで, 検索精度が向上することを示した. 他にも, ヒューリスティックにフィード検索に関するモデルを構築する手法も提案されている. Lee ら[4] は, トピックとフィードとの適合度を 2 つの仮定に基づいて推定した. 1 つは, クエリに適合する記事数がフィードに属する記事数に対して高い割合であるフィードは, より適合性が高いという仮定である. 2 つ目は, クエリへの適合度が高い記事がそのフィードを表しているという仮定である. これらの仮定に基づいた検索スコアにより, Lee らは 2008 年におけるブログトラック参加者の中で最も高い検索精度を報告している.

2.2 検索モデルの統合

従来の（フィード検索ではない）情報検索において、検索モデルの統合が古くから試みられており、一般的に検索精度が向上することが知られている。この直感的説明として、「多くのモデルによって上位に順位付けられる文書はクエリに適合する可能性が高い」と言うことが出来る。本研究における提案手法は検索モデルの統合に関係しているため、以下に関連研究を教師あり・教師なしの観点からまとめる。教師なしの検索モデル統合の古典的な手法として、Lee ら[5]は複数の検索モデルにおける文章の検索スコアを正規化し、それらの和（CombSum）、最大値（CombMax）などによる統合手法の有用性を実験的に示している。他の方法として、ボルダ方式（Borda method）あるいはコンドルセ方式（Condorcet method）を利用した統合手法も提案されている。これらは、一般に選挙において用いられる候補者をランク付けする投票アルゴリズムである。ボルダ方式によるランキングの統合として、Borda fuse[6]が挙げられる。これは、各検索モデルにおける順位付けをボルダ得点によって与えることで、ある文書のボルダ得点による総合得点を算出し、再順位付けを行うものである。しかし、ボルダ方式やコンドルセ方式は、必要十分条件を考えれば起こりうる状況すべてにおいて最適であるとはいえない。そこで Farah と Vanderpooten [7]は、従来のボルダ方式やコンドルセ方式の問題点を解決したランキング統合手法を提案している。

一方、教師あり手法としては、Yue ら[8]による SVM-MAP が挙げられる。SVM-MAP は、文書の順位を推定するためのサポートベクトルマシンアルゴリズムの1つであり、教師付き事例を用いた学習によって、情報検索システムの評価指標としてよく使われる Mean Average Precision (MAP) を最大化する。ここでの教師付き事例とは、クエリに対する適合文書あるいは非適合文書である。SVM-MAP は、複数の検索モデルのスコアの分布を素性として、その最適な組み合わせ（重み）を訓練データから学習する。ブログトラックなどにおける従来手法は、いずれもフィードに属する記事、あるいはその検索スコアをどのように統合するかに着目した手法だと考えることが出来る。本研究では、次節で述べる別の着眼点から、アンサンブル学習を応用した検索モデルの統合による新しいフィード検索の枠組を提案する。

3. 提案手法

3.1 概要

本研究では、次の2つの仮定を基に新しいフィード検索の枠組を提案する。1つは、ユーザのクエリに適合するフィード（以降、「適合フィード」と呼ぶ）あるいは非適合フィードには何らかの特徴があるということである。分かりやすい例では、ユーザが神戸のスウィーツに興味があるのであれば、適合フィードに属する記事の多くが神戸のスウィーツに関する記述を含むものと考えられる。言い換えれば、「神戸 スウィーツ」というクエリに対する各記事の検索スコアは概して高いと考えられる。もう1つの仮定は、フィードによってそれに適した検索モデルは異なるということである。既存の検索モデルには多くの種類があり、一般に、どの文書コレクション、どのクエリに対しても最適なモデルは存在しない。

本研究では、上述のようなフィードごとの特徴を利用することでそれぞれに適切な検索モデルを選択（重み付け）し、より高精度なフィード検索を実現する。その際、検索モデルを機械学習における分類器と見なすことで、アンサンブル学習を応用する。アンサンブル学習とは、個々に学習した複数の分類器から1つの分類器を構築する手法であり、本研究では特に、事例ごとに適切な分類器を選択する手法である Mixture-of-Experts (ME) を応用する。

3.2 Mixture-of-Experts

アンサンブル学習とは、複数の分類器を統合することで予測能力の向上を図る学習手法であり、様々な手法が存在する[9]。単純には、複数の分類器の出力に対して、平均、和、最大値、最小値などを用いることで1つの分類器を構築する方法がある。また、訓練データを複数回サンプリングすることで分類器による予測を複数行い、それらの投票で最終的な予測を行うバギングと呼ばれる手法がある。さらに、ブースティングと呼ばれる逐次的に例題の重みを変えて学習し、分類器の予測を組み合わせる手法も提案されている。しかしながら、これらの手法は事例（ここではフィード）に応じて適切な分類器を選択するという本研究の狙いに適さない。そこで、入力事例の特徴量によって複数の分類器の重みを動的に変化させることが可能な ME を応用することで、検索モデルの統合を行う。

ME [10]はアンサンブル学習の1つであり、ゲートネットワーク（gating network）によって複数の分類器から適切な分類器を選択したり、適切な重み付けによって分類器の統合を行う。MEでは、入力事例 x を分類器 $D_1, \dots, D_i, \dots, D_T$ によって分類し、各出力に対してゲートネットワークによる重みを与え、重み付けされた出力のうち最大値を最終的な出力としている。

本研究では、入力事例をフィード、分類器を検索モデル、各分類器による出力を検索スコアと見なすことで、MEの枠組を検索モデルの統合に応用する。ここで、通常のMEと提案手法との大きな違いは、ゲートネットワークへの入力である。前者では、各学習器に入力する事例 x と同一の情報をゲートネットワークへの入力とする。しかしながら提案手法では、後述するフィードの他の特徴量を入力とする。なお、MEでは唯一の学習器を選択しているのに対し、提案手法では各検索モデルに関して入力フィードに応じた重みを付与することでモデルの統合を行う。図1に本研究におけるMEの概念図を示す。

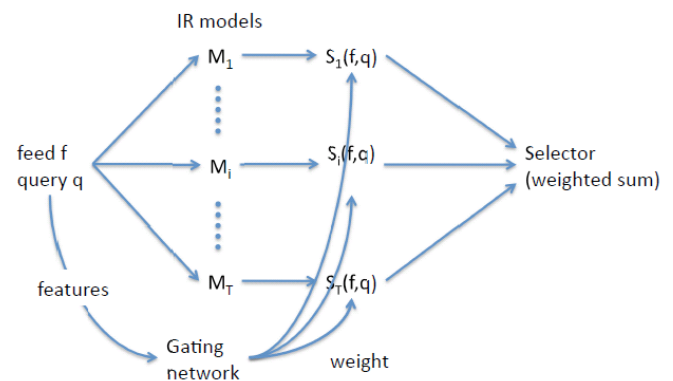


Fig. 1 Mixture-of-expert to combine multiple IR models

図1では、あるクエリ q と入力フィード f に関して、各検索モデル M_i による検索スコア $S_i(f, q)$ を算出し、さらに

ゲートネットワークから得た重み $w_i(f, q)$ によって線形結合を行うことでフィード f の最終的な検索スコア $S_{sum}(f, q) = \sum_i w_i(f, q) \cdot S_i(f, q)$ を出力する。ME による検索モデル統合の具体的な手順を以下に示す。

- (1) フィードの特徴量の算出
- (2) ゲートネットワークの構築
- (3) 検索モデルの統合

最初の2ステップが訓練フェーズであり、最後のステップが実際の検索に相当する。以下、それぞれについて詳しく説明する。

(1) フィード特徴量の算出

本研究では、クエリに適合する（あるいは不適合な）フィードには何らかの特徴があり、それらの特徴によって有効な検索モデルが異なると仮定し、複数の検索モデルの統合を行う。本研究で利用するフィードの特徴量について以下で述べる。なお、これらの特徴量は、関連研究などを考慮して恣意的に選択したものであり、その表現方法も含めて、必ずしも最適であるとは限らない。

・ フィードに属する記事の検索スコアの分布

ユーザのクエリに適合するフィードには、概して、そのクエリに適合する記事が多く存在すると考えられる。同様に、非適合フィードにはクエリに適合する記事が少ないと考えられる。これらから、同一フィードに属する記事の検索スコアの分布を特徴量として用いる。分布の表現方法は以下の通りである。まず、ある検索モデルについて、フィード全ての記事の検索スコアを降順に整列し、等間隔でスコアを一定数サンプリングする。（ここでは50個とする）これらのスコアを $K = \{k_1, \dots, k_b, \dots, k_{50}\}$ とする。 k_1 は検索スコアの最小値、 k_{50} は検索スコアの最大値、幅は $(k_{50} - k_1)/50$ となる。同数の要素からなるベクトル $H(f) = \{h_1, \dots, h_b, \dots, h_{50}\}$ によって検索スコアの分布を表現する。各要素 h_i の値は、フィード f に属する記事のうち、その検索スコア S_i が $S_i < k_i$ を満たす記事の数とする。また検索スコアの統計量である、平均、分散、最頻値、尖度も用いる。

・ フィードに属する記事数

記事が極端に少ないフィードは、その内容がユーザのクエリに適合するとしても、「お気に入り」としてフィードリーダに登録されることは考えにくい。そこで、フィードに属する記事数を特徴量として利用する。

・ フィードの文章量

記事数と同様に、文章の量も適合・非適合の判断に関係していると考えられるため、特徴量として利用する。文章量は、フィードを構成する記事の総単語数（延べ）で表現する。また、記事数の違いを考慮するため、記事ごとの総単語数も特徴量として用いる。

・ フィード内のクラスタ数

一般的に、ブログは様々な事柄（トピック）について書かれることが多い。例えば、日記のようなブログでは、趣味、仕事、生活、旅行、友人など、トピックは多岐に渡る。しかしながら、フィード検索によって見つけようとするブログは、検索に用いるクエリに関する記事が大多数で、他のトピックについて書かれている記事は少ない方が望ましい。言い換えれば、トピック数は少ない方が望ましい。ここでは、フィードに属する記事をクラスタリングして得られるクラスタ数をトピック数と見なし、フィードの特徴量として利用す

る。本研究では、クラスタリングに EM アルゴリズムを用いる。なお、Macdonald ら[2]により、同様のクラスタ数を利用することがフィード検索に有効であることが報告されている。

・ フィードへのインリンク数

他のページから自身へリンクしている数を特徴量とする。これは、他のページからのリンクはリンク先への「支持（endorsement）」であるという考えに基づく[11]。また、フィードから他のページへのリンク数も特徴量として利用する。

・ 更新頻度

適合フィードは、ユーザのクエリに適合するだけでなく、ある程度の頻度で更新されていることが望ましい。しかしながら、人手で行うことが不可能なほどの頻度で更新されているフィードは、スプログである可能性もある。そこで、フィードの記事の更新頻度を、最初と最後に書かれた記事の時間差（日数）を記事数で割ることで表現する。

・ クエリ語を含む記事の割合

同一フィードに属する記事集合のうち、クエリ語を含む記事の割合を特徴量とする。多くの記事がユーザのクエリを含む場合、適合フィードである可能性が高いと考えられる。

(2) ゲートネットワークの構築

次に述べる訓練データを入力とし、ニューラルネットワークによってゲートネットワークを構築する。入力ユニット数は、上記で述べた特徴量の数である。各検索モデルの重みを出力するため、出力ユニット数は統合を行う検索モデル数になる。中間ユニット数は (出力ユニット数+入力ユニット数)/2 とする。

訓練データには、適合フィードであれば最も高い順位を出力した検索モデルを、非適合フィードであれば最も低い順位を出力した検索モデルをラベルとする。そしてニューラルネットワークによって学習し、各検索モデルに与えられる確率値を重みとする。

(3) 検索モデルの統合

訓練データによるゲートネットワークの構築後、未知のクエリに対して検索モデルの統合によるフィード検索を行う。具体的には、手順(1)で述べた特徴量をいずれかの検索モデルによって検索された各フィードに対して抽出し、手順(2)で構築したゲートネットワークにより各検索モデル M_i の重み w_i をフィードごとに推定する。最後に、得られた重みによって検索モデルの検索スコアを統合する。なお、各検索モデルのスコアは、順位の逆数とす。

4. 評価実験

4.1 実験設定

提案手法の有効性を評価するため、本研究では、TREC ブログトラックで使用された Blog06 と呼ばれるコレクションを用いて評価実験を行った。このコレクションは、100,649 件のフィードに属するブログ記事約 3.5 百万件から成る。また、検索システム評価のために、2008 年のブログトラックで 50 件のクエリが作成されており、それぞれのクエリについて、適合するフィード、適合しないフィードがプーリングという手法により人手でラベル付けされている。本研究では 2008 年のデータを 2-fold cross validation により利用した。

統合の対象とする検索モデルとしては、2008年のTRECのblog distillation taskのパラダイムが異なる上位3つのモデル（cmuLDwikiSP, KLEDistFBT, UMassBlog3）を用いた。評価指標としては、ブログトラックで用いられた指標と同一のMean Average Precision (MAP)を用いた。直観的には、MAPはrecall-precision曲線の下面積を表しており、高い値を持つほど検索性能が高い。

4.2 実験設定

本研究で利用した3つの検索モデルについて、それぞれの検索精度 (MAP) を表1に示す。なお、評価に用いたフィード数は、検索結果上位100件である。

Table 1 retrieval performance in MAP

cmuLDwikiSP	KLEDistFBT	UMassBlog3
0.3056	0.3010	0.2711

まず、本研究で主張した「フィードによって適した検索モデルは異なる」という仮定が正しいかを確認するための実験を行った。各フィードについて表1の3つのモデルの中から最適な検索モデル（適合フィードであれば最も大きなスコアを、非適合フィードであれば最も小さなスコアを出力したモデル）を誤りなく選択した理想的なケースを考えた場合、検索精度 (MAP) は0.5962となった。各モデルの精度よりも大きく精度が向上しており、この仮定は正しいことが裏付けられた。

次に提案手法により表1の検索モデルの統合を行った結果、検索精度 (MAP) は0.2847となった。この結果は、利用した各検索モデルの検索精度と比較して向上しておらず、提案手法の有効性は確認できなかった。今回用いたフィードの特徴では各モデルの重みが適切に学習できていない可能性があり、今後の検討が必要である。

5. むすび

本研究では、ブログフィード検索問題に関して、検索モデルを分類器とみなすことで、アンサンブル学習の1つであるMEを応用した検索モデル統合の新しい枠組を提案した。提案手法では、ユーザのクエリに適合するフィードまたは適合しないフィードには何らかの特徴があると仮定し、各フィードの記事数や単語数で表現した。さらに、フィードごとに有効な検索モデルは異なるという仮定のもと、前述のフィードの特徴によって動的に検索モデルの重みを決定することで複数モデルの統合を行った。大規模なブログ記事コレクションを用いた評価実験の結果、フィードによって適した検索モデルは異なるという仮定の妥当性を示すことはできたものの、提案手法の効果は見られなかった。

今後、より適切な重みづけを行うために各フィードを表す特徴を再検討することで、フィード検索のさらなる精度向上を図っていく予定である。

文 献

- [1] Seo, J. and Croft, W.B.: Blog site search using resource selection, Proceeding of the 17th ACM conference on information and knowledge management, pp. 1053–1062 (2008).
- [2] Macdonald, C. and Ounis, L.: Key blog distillation: ranking aggregates, Proceeding of the 17th ACM conference on information and knowledge management, pp. 1043–1052 (2008).
- [3] Elsas, J.L., Arguello, J., Callan, J. and Carbonell, J.G.: Retrieval and feedback models for blog feed search, Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, pp.347–354 (2008).
- [4] Lee, Y., Na, S.-H., Kim, J., Nam, S.-H., Jung, H.-Y. and Lee, J.-H.: KLE at TREC 2008 blog track: blog post and feed retrieval, TREC 2008 (2008).
- [5] Lee, J. H.: Analyses of multiple evidence combination, Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval, pp.267–276 (1997).
- [6] Aslam, J.A. and Montague, M.: Models for Metasearch, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp.276–284 (2001).
- [7] Farah, M. and Vanderpooten, D.: An outranking approach for rank aggregation for optimizing average precision, Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pp.591–598 (2007).
- [8] Y.Yue, T.F., Radlinski, F. and Joachims, T.: A support vector method for optimizing average precision, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp.271–278 (2007).
- [9] Polikar, R.: Ensemble based systems in decision making, IEEE circuits and systems magazine, Vol.6, No.3, pp.21–45 (2006).
- [10] Jacobs, R., Jordan, M. and Hinton, G.: Adaptive mixtures of local experts, Neural Computation, Vol.3, pp.79–87 (1991).
- [11] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web., Technical report, Stanford InfoLab (1999).