

能動学習による効率的な情報フィルタリング

Graham Neubig †‡

森 信介 †

† 京都大学 情報学研究科

‡ 日本学術振興会 特別研究員

1 はじめに

インターネットが豊かな情報源であり、その中からユーザーにとって有用な情報を整理・提供する研究は様々な分野でなされてきた [1, 2]。その多くは自動分類などの機械学習技術を利用し、高い情報抽出精度を実現している。

しかし、災害時などにおいて人命に関わる情報を整理・提供する際、誤った情報の提供が1件でも許されない。現在までの災害時における情報抽出の試みは、大勢の作業者を集めて全作業を人手で行う手動情報抽出 [5] と、自動情報フィルタリングで大量の情報から有用そうな候補を取り出してから人手でチェックする半自動情報抽出 [7] がある。また、災害時における情報抽出は信頼性以外にも、提供するスピードが求められる。しかし、人手による作業が必要である以上、情報提供できるスピードは作業効率によって大きく左右される。

本研究は、能動学習を用いて、この両方の要件を満たす効率的な情報フィルタリング法を提案する。具体的には、提供する情報に作業者のチェックが必要であるという前提のもと、チェック候補を特定する分類器を学習する。さらに、チェックされた事例で定期的に分類器を再学習し、候補提示の精度を向上させる。従来の能動学習が正例・負例の境界面に近い事例を提示するのに対して、本手法は正例の可能性が最も高い候補を提示するため、作業者がチェックする負例の数を最小限に抑え、作業効率を向上させる。

評価実験では、震災関連のツイートコーパスから有用なツイートを分類するタスクで提案手法を用いた情報フィルタリングは30分で平均204件の有用情報をチェックできた。これは従来の能動学習法の63件を大幅に上回り、約3倍の作業効率を実現している。

2 文書分類と能動学習

2.1 文書分類

文書分類はある文書 d_j がクラス c_i に属するか属さないかを判定するタスクとして定義される。本研究で扱う分類タスクは「有用な情報あり (正例)」 c_y か「有用な

情報なし (負例)」 c_n かという2値分類の問題である。この設定の文書分類は「情報フィルタリング」とも言われる [3]。

近年の文書分類は機械学習で構築された分類器で行うことが多い [10]。2値分類の場合、文書 d_j から素性の値を計算する K 個の素性関数 $\phi_k(d_j)$ を定義し、各素性に対する重みを与える λ_k を用いてスコア s_j を計算する。

$$s(d_j) = \sum_{k=1}^K \lambda_k \phi_k(d_j)$$

$s(d_j) \geq 0$ の場合、 d_j を正例と判定し、 $s(d_j) < 0$ の場合、 d_j を負例と判定する。

重み $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_K)$ を学習するために、文書 \mathcal{D} とその正解ラベル \mathcal{C} からなる学習コーパスを用意し、サポート・ベクター・マシン (SVM) やナイーブ・ベイズなどの学習アルゴリズムを利用する [6]。

2.2 能動学習

分類対象の文書 \mathcal{D} は大量に収集可能であるとし、正解ラベル \mathcal{C} は人手により作成するとする。 \mathcal{C} が多ければ多いほど分類器の精度が向上すると期待できるが、ラベル付け作業に利用できる時間やコストは限られている。限られたラベル付け時間で最大の精度向上を実現する枠組みとして能動学習が広く使われている。

広義では、能動学習は「分類器の信頼度に基づいて、 \mathcal{D} の中からどの事例にラベル付けするかを選択する」と定義される。その中で、 \mathcal{D} の中からラベル付けする事例を選択する基準はいくつか提案されており、代表的な例として分類平面からの距離が最も小さい事例を選択する手法がある [12]。

$$\hat{d}_j = \operatorname{argmin}_{d_j \in \mathcal{D}} |s(d_j)| \quad (1)$$

これは分類平面に近い事例は誤分類の可能性が高く、ラベル付けすれば確実に正しい分類ができるようになり、この事例から学習された素性をさらに他の事例に適用できるという考え方に基づいている。

2.3 素性のラベル付け

従来の分類器学習は事例に対するラベル付けを行うが、素性に対するラベル付けを行う研究も提案されている [4]。例えば、災害後において Twitter などから避難所に関する情報のツイートを抽出しようとする際、「避難所」という単語 2-gram が存在すれば正例の可能性が高く、「避難所」を正例の素性としてラベル付けする。特に素性と事例のラベル付けを組み合わせる手法は先行研究で効果的であるとされており [11]、本研究でもこの枠組みを採用する。

有用な素性を発見する方法として主に 2 つの手順がある。1 つは、[11] で提案されているような自動素性提示法である。学習器が情報量の多い素性を自動的に計算して提示し、作業者がそれを正例（負例）らしい素性としてラベル付けする。もう 1 つの方法として、ラベル付けする事例の中から発見した単語の中から正例か負例を指す単語を選び出し、ラベル付けする手法がある。

3 半自動情報抽出

3.1 情報抽出タスクの特徴

本研究では、Web からの情報抽出に着目し、Twitter からの災害関係情報抽出をその一例として扱う。このタスクには 3 つの大きな特徴がある。

1. 絶対的な信頼性が求められる：災害関係の情報を誤って提供すれば人命に関わる可能性もある。このため、情報を提供する前に人手でチェックする事はどうしても必要となる。医療や法律などの場面でも同等の条件が設けられると考えられる。
2. スピードが重要：例えば、被災者に避難所や救援物資の情報を届ける場合など、災害時の対応は一刻を争う。先行研究 [7, 5] では大勢のボランティアで 1 つのタスクを集中的に行うことである程度のスピードを実現しているが、ボランティア間の基準のゆれが生じるなど、問題も指摘されている。
3. 有用でない情報が大多数：4.1 項で詳しく説明する東日本大震災直後に集められたツイートコーパスでは、震災関係の検索クエリで絞っても、生存情報を提供しているツイートはわずか 2.7% しかない。クエリを利用しない場合はこの割合は更に低下する。

次節では、このような条件を全て満たす文書分類器構築法を提案する。

3.2 正例候補に着目した能動学習

本研究の基本的な分類器構築法は 2.2 項で紹介された能動学習法と類似している。異なるところは、式 (1) の

ラベル付け選択基準を以下の基準に変更する点のみである。

$$\hat{d}_j = \operatorname{argmax}_{d_j \in \mathcal{D}} s(d_j) \quad (2)$$

つまり、現在の分類器にとって最も判断が難しい事例ではなく、正例の可能性が最も高い事例を選択する。

このような事例の選択基準を利用することで、作業者に提示される事例のほとんどが正例であることが期待される。特に、前節で述べたような、負例が非常に多く含まれている場合にはこの選択基準が有用である。逆に、偏ったデータで従来手法の分類平面からの距離を基準にすると、作業者に提示される事例の分布が実際のデータの分布に従い、結果的にほとんどが負例となることが考えられる。

正例を多く選択することで 3 つの効果が期待される。

1. ラベル付けする正例はチェック済み情報ともなり、半自動情報抽出の目的である信頼性の高い情報の提供が実現できる。
2. 正例を多くラベル付けするため、正例に対する安定した重みが学習可能となり、正例の抽出精度向上が期待される。これに対して、分類平面に近い負例をラベル付けしても、学習器のスコアが大きくなる負例を除外することができず、スコアの高い負例は残ったままとなる。
3. ラベル付け作業の精神的な負担を減らす。分類平面からの距離を利用すると、外国語を含む事例や非常に短い事例など、実際に抽出したい情報と縁のない事例の提示が多く、実際にラベル付けしたい事例がデータに存在するかどうかさえも怪しまれる。

1 つ断っておくべきことは、本手法は必ずしも精度の高い分類器の構築に向いているという訳ではない。逆に実際のデータが負例を多く含むにも関わらず、正例を多く含む学習データを作成しているため、分類器が実際より多くの事例を正例と判定することになる。しかし、ここで重要なのは、分類器が正しく正例かどうかと判定できるのではなく、実際に作業者に提示する、スコアの高い未ラベル付け事例が正例であるかどうかである。

このモチベーションは情報検索等で広く用いられる関連フィードバックに類似している [13]。しかし、情報検索の場合では主に成熟した情報検索システムをユーザーの趣向に適應する手法として用いられ、ゼロから新しい応用のためにシステムを作るために利用されていない。

全ツイート	29,919
安否情報要請	3,974 (13.2%)
安否情報提供	813 (2.7%)

表 1: コーパスのツイート数。安否情報関連のツイートは外部リンクを中心とするツイートを除く。

種類	正例	負例
安否情報要請	取れません、とれません	http
安否情報提供	無事、無事だった	http
避難所・物資	避難所、水、充電、炊き出し	安否

表 2: 分類器学習の初期化に利用した素性ラベル。

4 実験評価

4.1 実験設定

提案手法の評価には Twitter からの震災関連情報抽出をタスクとする。ANPLI NLP プロジェクト [7] の一部として収集された東日本大震災の直後のツイートコーパスをデータとして利用する。具体的には、このコーパスは「#anpi」「#hinan」「#j.j_helpme」「#save_[地名]」などのハッシュタグを含んだツイートが中心となっており、予め震災関係のツイートに絞ったものである。また、全てのツイートは内容に応じて人手でラベル付けされている。

コーパスの諸元を表 1 に示す。この中で、震災関連のハッシュタグで絞ったツイートであるにも関わらず、「安否情報要請」や「安否情報提供」のような有用な情報が非常に少ないということが分かる。

比較対象は、従来の分類器に基づく能動学習と、提案手法の正例候補に基づく能動学習の 2 通りとする。効率よく作業を行うために、事例と素性のラベル付けが両方可能な能動学習インターフェース DUALIST[11] を利用した¹。文書分類の素性として、単語 1-gram と単語 2-gram を利用し、効率の良い学習を行うために学習アルゴリズムとしてナイーブ・ベイズを利用している。その前処理として、コーパスの形態素解析を KyTea[8] で行い、「語尾」と「助動詞」を直前の単語と一緒にまとめた。実験の評価基準は、1 節の背景を踏まえて「1 人の作業者が 30 分の作業で何件のチェック済有用情報が抽出できるか」とする。この 30 分の中で、15 分を能動学習による分類器構築に利用し、15 分を分類器が出力したスコアを上から見てラベル付けを行った。

情報抽出の対象は「安否情報要請」、「安否情報提供」、

¹<http://code.google.com/p/dualist/> にてオープンソースで入手可能。

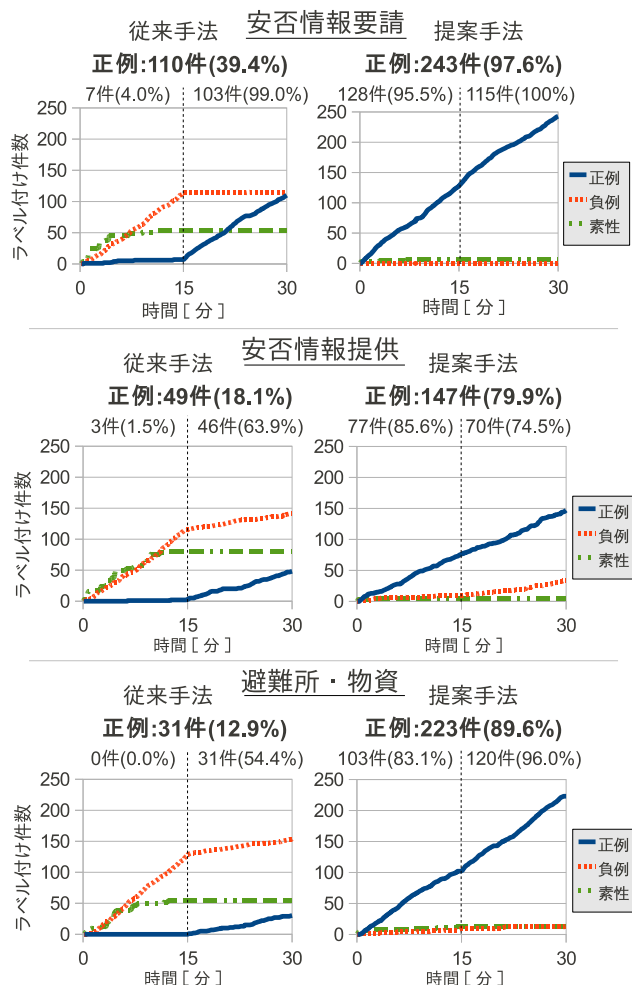


図 1: 3 種類の情報の抽出効率。件数はラベル付けされた正例の件数であり、割合は全ラベル中の正例が占める割合。前半 15 分（能動学習）と後半 15 分（信頼度の高い順のラベル付け）も分けて表示してある。

「避難所・物資情報提供」の 3 通りとした。学習の最初に、従来手法、提案法ともに表 2 の通りの素性ラベルで学習を初期化した。

4.2 実験結果

主な実験結果を図 1 に示す。提案手法を用いて、作業者は 30 分で平均 204 件の有用なツイートを抽出してチェックすることができた。これは、従来の能動学習法を用いた場合の平均 63 件の 3 倍を超える。

特に、分類器の学習に用いられる最初の 15 分に注目すると、従来手法でラベル付けされる事例のほとんどが負例、または素性であったのに対して、提案手法では学習が始まってからすぐに正例のラベル付けがほとんどとなった。このように能動学習と正例のチェック作業を組み合わせていることが従来手法との抽出件数の差の大き

従来手法	提案手法
給水、物資、電気、 学校、水道、ガス、 ガソリン、食料、営業	給水、営業、銀行 風呂、ガソリン

表 3: 「避難所・物資情報」の正例を指すとして追加された素性。下線を引いた素性は事例から発見したものであり、下線のない素性は自動提示によるものである。

な要因となっている。

また、分類器のスコアに基づいてスコアの高い順に処理を行っていく後半の 15 分では、従来手法の正例件数が大きく向上する。しかし、全ての場合において、後半でも提案手法によって学習された分類器の方が高い割合で正例を提示していることが分かる。

この差が特に著しかった「避難所・物資情報提供」でその原因を探てみると、2つの要因があることが分かった。まず、従来手法によって学習された分類器は、能動学習終了後でも物資情報を提供している外部リンクを含むツイートに高いスコアを与えていた。これは今回のラベル付けでは対象外としていたものの、これらのツイートは表 2 に含まれる正例素性を含む場合があり、学習の最初から高いスコアとなっていたため、スコアが分類平面に近い事例を選択する従来の能動学習法では 1 回も選択されなかった。

もう 1 つの要因は、提案手法による能動学習の過程で、提示された正例の中からより正確な素性ラベル付けを行うことができたことである。この過程で得られたラベル素性は表 3 の通りである。2.3 項で述べたように、自動提示と事例中の発見という 2 つのラベル発見法がある。しかし、文脈なしに自動提示された素性をラベル付けすることが危険であることは評価実験を通して分かった。例えば、自動提示された「水道」「ガス」「電気」などは、「電気はまだ復旧せず」のような否定的な表現で利用されることが多く、決して正例を指す素性ではない。逆に、「銀行」や「風呂」のような正例から発見された素性は比較的頻度が低かったものの、正例との強い結びつきを持ち、分類精度を下げずに適切に抽出する情報の規模を拡張できた。

5 おわりに

本研究は正例に着目することで、効率的かつ信頼度の高い情報フィルタリングを行う手法を提案した。正例の数が比較的少ないコーパスの中から、30 分で平均 204 件の有用な事例を特定することができ、従来の能動学習法より約 3 倍の効率を実現することができた。

これからの課題としては、多クラス分類への適応が考

えられる。本研究では 3 種類の有用情報に対して、3 回のラベル付けで 2 値分類器を学習したが、この 3 種類の情報を同時に学習できればさらなる効率の向上につながる。また、正例のみから学習可能な分類アルゴリズム [9] も研究されており、これらを適用すればさらなる精度向上も見込まれる。有用な情報を含む文書やツイートを特定すると同時に、その中の更に細かい単位の固有表現を抽出する枠組みもこれからの有望な研究課題である。最後に、文書が正例かどうかを判定する時に、作業者はある特定の表現（「確認できました」「給水情報」等）を探して判断をするが、この判定に有用な情報を自動的に特定し、色付けや太文字で見やすくすることで作業効率を更に向上させられる可能性もある。

参考文献

- [1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proc. EMNLP*, 2011.
- [2] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction for the web. In *Proc. of IJCAI2007*, 2007.
- [3] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35(12), 1992.
- [4] G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. In *EMNLP*, 2009.
- [5] Google Japan. 共有された被災者名簿のパーソナライズ登録についてご協力をお願い。 http://googlejapan.blogspot.com/2011/03/blog-post_17.html, 2011.
- [6] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *ECML-98*, 1998.
- [7] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami. Safety information mining - what can NLP do in a disaster -. In *Proc. IJCNLP*, pp. 965–973, Chiang Mai, Thailand, November 2011.
- [8] G. Neubig, Y. Nakata, and S. Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. ACL*, pp. 529–533, Portland, USA, June 2011.
- [9] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- [10] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1), 2002.
- [11] B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proc. EMNLP*, 2011.
- [12] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *JMLR*, 2, 2002.
- [13] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 8(6), 2003.