D-008

トピック判定における n-gram の高速組み合わせ手法の検討

Applying n-gram Assortment Methods for Topic Determination

柳原 正† 松本 一則† 小野 智弘† 滝嶋 康弘†

Tadashi Yanagihara Kazunori Matsumoto Chihiro Ono Yasuhiro Takishima

概要

トピックごとに事前に定めておいたキーワードを含むテキストを該当トピックのテキストであると判定するトピック判定手法が存在する。同手法では、判定に用いるキーワード集合の良し悪しがトピック抽出の精度に大きな影響を与える。また、単一のキーワードの代わりに単語ペア（word bigram）を使うことでトピック判定の精度を向上できると思われるが、最適な n-gram の組み合わせを発見するための処理量が多いという欠点を持つ。本稿では、unigram でのスコアに基づいて、n-gram の候補を限定することで処理量を軽減する手法を提案する。

1. はじめに

特定話題のテキストを抽出するためにトピック判定を行う手法が数多く提案されている。例えば Naïve Bayes Classifier などの識別器を使った判定方式[1]がある。この方式の場合、識別器の入力となる特徴量は、通常、tf・idf のように対象テキストを形態素処理して得ている。

一方、事前に定めておいた語を含むテキストを該当トピックのテキストであると判定するトピック判定手法（文字列マッチングによるトピック判定手法）も存在する。同時発生する大量のテキストに対してトピック判定を行う場合、形態素処理を行って特徴量を生成する判定手法より、文字列マッチングによる判定手法の方が、処理量の点で有利である。

ただし、文字列マッチングを用いる手法の場合、各トピック毎に、トピックへの関連性が高い単語を選定しておく必要がある。この仕組みを実装したシステムとして、基本的な構造は図 1 のようになる。

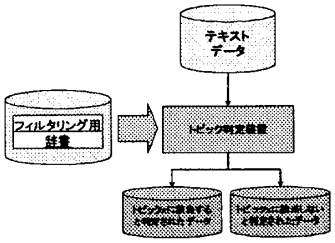

図 1. パターンマッチング方式に基づく判定システムの図

† 株式会社 KDDI 研究所

文字列マッチング方式における判定処理のナイーブな実装では、キーワード集合を Suffix Array 等の適切なデータ構造に変換して、文字列マッチングを高速に行うことになる。しかし、単一の語（unigram）のマッチングを条件にトピック判定を行う限り、トピックに該当すべきではないのに該当すると判定してしまうという誤検出が多くなり、判定精度の点で限界がある。

この問題を解決するためには単純な unigram ではなく、bigram 等の n-gram の利用が考えられる。一般的に、n-gram は誤検出を減少させることができ、適合率(precision)が向上するが、再現率(recall)は低下することが多い。再現率の低下を避けるためには、マッチング時に利用する n-gram の種類数を増やすことで対処できるが、大量の n-gram の中から判定に適したものを抽出するための計算量が膨大になることが問題である。

この問題を解決するために、本稿では、unigram でのスコアに基づいて、n-gram の候補を限定することで処理量を軽減する手法を提案する。

2. 提案方式

本提案では、トピック判定を行う上で効果が高い unigram を推定し、これらの unigram から効果が高い bigram を形成するための計算方式を提案する。これにより、判定する上で有益となる単語の組み合わせが効率的に形成すると同時に、計算処理量を大幅に抑えることができる。

はじめに、特定のトピックに該当する学習用に用いるテキストデータであるD_p、特定のトピックに該当しない学習データであるD_nを用意する。次に、これらのデータから単語候補となる単語集 W を抽出する。単語集 W に含まれるそれぞれの単語 w に対し、(a) 単語 w がD_pに出現する文書数、(b) 単語 w がD_pに出現しない文書数、(c) 単語 w がD_nに出現する文書数、(d) 単語 w がD_nに出現しない文書数として求める。ここで求まる値をそれぞれ(a) N_{11}, (b) N_{12}, (c) N_{21}, (d) N_{22}とする。これらの値を使い、各単語 w に対する 2x2 分割表 T を形成する。

	単語 w が含まれる	単語 w が含まれない
D_p	N_{11}	N_{12}
D_n	N_{21}	N_{22}

表 1. 単語 w に対する 2x2 分割表の生成例

次に、それぞれの単語に対するトピックへの関連度を表すスコアを計算する。これには様々な方法があるが、本提案では情報量基準に基づくモデル検定方式[2]を利用する。本提案で求める値は、単語の unigram に対するスコアとし

てはシングルスタティックスコア(SSS)、単語の組み合わせである bigram に対するスコアとしてはマルチスタティックスコア(MSS)とマルチダイナミックスコア(MDS)の計 3 種類である。

2.1. unigram のスタティックスコアの計算方法

各単語の unigram に対し、2x2 分割表を作成し、モデル検定を行う。ここでモデル検定を行うためには、情報量基準に基づいた独立モデル・従属モデルの値を求めなければならないが、これは以下の通りで求められる。

1. *単語 w が与えられたとき、2x2 分割表を作成する際に利用する変数を以下のように定義する：*
   - $a = N_{11(w)}$：*トピックに該当する文書D_pのうち、単語wが含まれる文書の数*
   - $b = N_{12(w)}$：*トピックに該当しない文書D_nのうち、単語wが含まれる文書の数*
   - $c = N_{21(w)}$：*トピックに該当する文書D_pのうち、単語wが含まれる文書の数*
   - $d = N_{22(w)}$：*トピックに該当する文書D_nのうち、単語wが含まれない文書の数*

   このとき、以下の式が成り立つ：
   - $q = N_{11(w)} + N_{12(w)}$：*単語wを含む文書数*
   - $r = N_{21(w)} + N_{22(w)}$：*全文書の数*
   - $N_p = N_{11(w)} + N_{21(w)}$：*特定のトピックに該当する文書の数*
   - $N_n = N_{12(w)} + N_{22(w)}$：*特定のトピックに該当しない文書の数*
   - $z = N_{11(w)} + N_{12(w)} + N_{21(w)} + N_{22(w)}$
2. *独立モデルの AIC 値である AIC(IM)を以下の式に基づいて計算する。(但し、0log0=0 とする)*

   $MLL = a\log a + b\log b + c\log c + d\log d - z\log z$

   $AIC(IM) = -2 \times MLL + 2 \times 3$
3. *従属モデルの AIC 値である AIC(DM)を以下の式に基づいて計算する。(但し、0log0=0 とする)*

   $MLL = q\log q + r\log r + (z - q)\log(z - q) + (z - r)\log(z - r) - 2z\log z$

   $AIC(DM) = -2 \times MLL + 2 \times 2$

これによって、AIC(IM)及び AIC(DM)が求まるが、これらの値の大小関係からは独立モデルか従属モデルとして仮定した方が良いかという情報しか得られない。これは、特定のトピックに該当するかを判別する上で役に立つ単語 w を調べたい場合では情報として不十分である。

そこで、[3]や[4]で述べられているように、モデルの適合度の大小関係を組み合わせることで、単語の重要度を表す値である E(w)を以下のように求める。

4. *単語重要度 E(w)を以下の式に基づいて計算する。*
   - *(ア)* $N_{11(w)} / (N_{11(w)} + N_{12(w)}) > N_{21(w)} / (N_{21(w)} + N_{22(w)})$ *のとき、* $E(w) = AIC(IM) - AIC(DM)$
   - *(イ)* $N_{11(w)} / (N_{11(w)} + N_{12(w)}) < N_{21(w)} / (N_{21(w)} + N_{22(w)})$ *のとき、* $E(w) = AIC(DM) - AIC(IM)$

これにより、「トピックに関連する文書」を判定する上で役に立つ単語に対して正の値として表し、「トピックに関連しない文書」を判定する上で役に立つ単語に対して負の値として表されるスコアである E(w)が得られる。これらの単一の単語ごとのスコアをシングルスタティックスコア(SSS)として定義する。

2.2. bigram のスタティックスコアの計算方法

各単語の unigram に対する SSS から、トピック判定を行う上で有益そうな単語の組み合わせである bigram を形成し、有益度に対するスコアを付加する。これは、SSS を求める際と同様に 2x2 分割表を形成するが、文書集合 Dp 及び Dn 内の含有数を単語 w によって決定するのではなく、単語 w_i と w_j の単語組み合わせ ω を使って集計を行う。

	組み合わせ ω が含まれる	単語の組み合わせ ω が含まれない
D_p	N_{11}	N_{12}
D_n	N_{21}	N_{22}

表 2. 単語の組み合わせ ω に対する 2x2 分割表の生成例

単語 wi 及び単語 wj と成りえる単語を全て選び、単語の組み合わせを形成することになるが、組み合わせを全て形成し集計を行うための処理量を考慮すると実用的に問題があるため、以下のように効果がありそうな単語 w_i と単語 w_j を選択し、単語の組み合わせ ω を形成する。

1. *初期化*
   - *(ア) 採用済み組み合わせの集合を G とする。G = {}*
   - *(イ) 採用済み組み合わせの数を u とする。 u = 0*
2. *繰り返し*
   - *(ア) シングルスタティックスコアが最も大きい単語 w を選択する。*
   - *(イ) 単語w以外にシングルスタティックスコアが大きい l 個の単語 w_1～w_l を選ぶ。(l は事前に決定した閾値)*
   - *(ウ) 単語w と単語 w_j (1≦j≦l)の組み合わせ ω_j に対する 2x2 分割表を求める。(但し、すでに ω_j に対する 2x2 分割表が計算済みである場合は再計算しない。)*
   - *(エ) $\omega_1, \omega_2, \ldots \omega_l$ において、それぞれの 2x2 分割表を使って、スタティックスコアを計算し、最もスタティックスコアが高い ω を求める。*
   - *(オ) ω を G に採用する。$G = G + \{\omega\}$、$u = u + 1$。以下、u が規定値に達するまで(ア)から繰り返す。*

これにより、スタティックスコアの大きい順に並べた単語組み合わせのリストが求まる。これらのスコアの集合を総じてマルチスタティックスコア(MSS)と定義する。

2.3. bigram のダイナミックスコアの計算方法

SSS を以上のように求めたが、この方式の欠点としては、組み合わせ ω 間に独立性があることが保証されていない。例えば、リストの第一候補となる組み合わせ ω_1 と第二候補となる組み合せ ω_2 において、ω_2 が判定対象とする文書は全て ω_1 に含有されている可能性がある。

この問題を解決するために、本提案の中に、単語の組み合わせ ω を選定する際に、学習対象とする文書の量を動的に調整しながら再集計を行うことで、リスト内の組み合わせ ω 間の間に独立性を保つ方式を提案する。

1. 初期化
   (ア) 順序未定組み合わせの集合Cに対して、$C = \{\omega_1, \omega_2 \ldots \omega_{NALL}\}$とする。
   (イ) 順序決定済み組み合わせの集合 L に対して、$L = \{\}$とする。(空集合)
   (ウ) C の各組み合わせωに対して、仮スコア $ts(\omega) = E(\omega)$を求める。
2. 繰り返し
   (ア) Cの中から、$ts(\omega_i)$が最大となるω_iを求める。
   (イ) $C = C - \{\omega_i\}$、$L = L + \{\omega_i\}$、$SDS(\omega_i) = TS(\omega_i)$
   (ウ) Cの中の各組み合わせω_jに対して、以下のように 2x2 分割表を更新する。
   - $N_{11}(\omega_j) = N_{11}(\omega_j) - n_{11}(ij)$
   - $N_{12}(\omega_j) = N_{12}(\omega_j) - n_{12}(ij)$
   - $N_{21}(\omega_j) = N_{21}(\omega_j) - n_{21}(ij)$
   - $N_{22}(\omega_j) = N_{22}(\omega_j) - n_{22}(ij)$

   但し
   $N_{11}(ij)$ = トピックに該当する文書のうち、ω_iとω_jが含まれる文書の数
   $N_{12}(ij)$ = トピックに該当しない文書のうち、ω_iとω_jが含まれる文書の数
   $N_{21}(ij)$ = トピックに該当する文書のうち、ω_iとω_jが同時には含まれない文書の数
   $N_{22}(ij)$ = トピックに該当しない文書のうち、ω_iとω_jが同時には含まれない文書の数
3. Cの中の各組み合わせω_jに対して、$E(\omega_j)$を更新する。以降、集合Cが空集合になるまで(ア)から繰り返す。

Cの中の各組み合わせωに対して、仮スコア$ts(\omega_j) = E(\omega_j)$を求める。これによって、ダイナミックスコアの大きい順に並べた組み合わせのリストが求まる。これらの組み合わせの集合を総じてマルチダイナミックスコア(MDS)と定義する。

3. 評価方針

以下に本方式の有効性を検証するための評価方針を想定している。トピック判定を行う対象としては、有害な情報を含むブログの検出を取り上げる。このときの「有害」とは、主に未成年者に対し適切ではない公序良俗に反する情報のことを指す。これらのデータを基に、「有害なトピックであるか」についてのトピック判定を行う予定である。

データセットとしては、ブログサイトから収集したブログ記事を使う。収集したブログ記事のうち、有害な表現を含むものを有害というラベリング付けを行い、それ以外のものを無害というラベリング付けを行う。次に、このデータからランダムに半分を取り出し、取り出したデータを学習用データに、残りのデータを評価用データとして利用する。これらのデータを基に、SSS・MSS・MDS の精度を計測する。さらに、網羅的に求めた bigram との比較も行う。

方式の評価尺度としては、テキスト検索で用いられる適合率及び再現率で精度を計測する。適合率及び再現率の二つの値をもとに、トピック判定に対する精度の評価を行う。

4. おわりに

本論文では、文章集合から特定のトピックに関する文章のみを自動で取り出すために利用するパターンマッチングベースのシステムについて述べた。特に単語を 2 つ以上組み合わせた bigram の集合のうち、効率的な組み合わせを発見するため組み合わせ手法を提案した。今後はこれらの方式の有効性を 3 章に述べた方法で評価し、精度向上を検証する。

参考文献

[1] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, Tom Mitchell, "Text Classification from Labeled and Unlabled Documents using EM", Machine Learning, Vol. 39, pp. 103-134, 2000.

[2] "情報量基準による統計解析入門", 鈴木一郎著 1995 年 講談社サイエンティフィク刊

[3] Kazunori Matsumoto, Kazuo Hashimoto, "Schema Design for Causal Law Mining from Incomplete Database", Discovery Science, Second International Conference, DS '99, Tokyo, Japan, December, 1999, Proceedings. Lecture Notes in Computer Science 1721 Springer, pp.92-102, 1999.

[4] 内山 将夫, 中篠 清美, 山本 英子, 井佐原 均, "英語教育のための分野特徴単語の選定尺度の比較", 自然言語処理, Vol 11, No. 3, pp. 165-198, 2004.