E-063

Web 検索結果のラベリングにおける閾値の利用について

Labeling Method with Threshold in Web Search Results

吉田 俊広 † 松原 雅文 ‡ Goutam Chakraborty‡ 馬淵 浩司 ‡
Toshihiro Yoshida Masafumi Matsuhara Goutam Chakraborty Hiroshi Mabuchi

1. はじめに

近年，Web 上の情報が増加している．その情報を検索するにあたって，多くのユーザーはロボット型検索エンジンを使用する．しかし，的確なキーワードを思いつくのが難しいことの他に，検索結果をリスト形式で出力するため，ユーザーが必要としない情報が混在しているという問題点がある．そのため，ユーザーの負担が大きい．それらを解決する手段の一つとして，検索結果をクラスタリングして表示する手法がある [1]．Web 検索結果のクラスタリングとは，検索キーワードにより取得した検索結果中に存在する，類似した Web ページのクラスタを生成することである．この生成されたクラスタに所属する Web ページの内容を示したラベルをクラスタに付与する．ユーザーはこのラベルを見ることで，どのような Web ページが含まれているか把握することができる．しかし，クラスタの内容を示していないラベリングがなされる場合もある．これにより，目的のページがクラスタ内に埋もれて発見できない場合や，ユーザーが必要な情報を見つけるのに時間がかかるという問題点がある．

ある検索キーワードで取得した検索結果をクラスタリングする．それにより，生成されたクラスタに付与するラベルが適切であるならば，このラベルと元の検索キーワードで AND 検索した場合，当該クラスタに含まれる Web ページはその AND 検索結果中に多数存在するものと考えられる．この一致した度合いを一致率として用いてラベリングを行うことにより，適切なラベリングが行えると考えられる．本研究では，一致率を用いてラベリングを行う手法を提案している [2]．

本手法では形態素を次元としてクラスタリングを行っており，さらに，この形態素をラベル候補として利用している．そのため，次元数が多いと，生成されたクラスタのラベル候補が増加し，ラベリングの精度に悪影響を及ぼすと考えられる．そこで，クラスタリングの際，閾値を設けて Web ページの次元数を削減してからクラスタリングをする．これにより，ラベリングにどのような影響があるか調査する．

本稿では，提案手法の概要を示し，行った評価実験の結果から，本手法の有効性を示す．

2. 提案手法

2.1 概要

提案手法の流れを図 1 に示す．まず，検索結果データを取得し，タイトルとスニペットのみを抽出する．この抽出したタイトルとスニペットを形態素解析し，名詞のみを抽出する．次に，これらに重み付けを行い，正規化する．この正規化された重みに閾値を設けて，閾値以下を排除し，次元数削減を行う．続いて，削減された次元数を使用し，クラスタリングを行う．このクラスタリングされた結果をもとに一致率を算出し，この一致率を利用してラベリングを行う．

2.2 形態素解析

形態素解析には，今回茶筌*を用いた．形態素解析結果から，名詞のみを抽出して，半角記号を排除したものを使用する．

†岩手県立大学大学院ソフトウェア情報学研究科
‡岩手県立大学ソフトウェア情報学部
*茶筌, http://chasen-legacy.sourceforce.jp

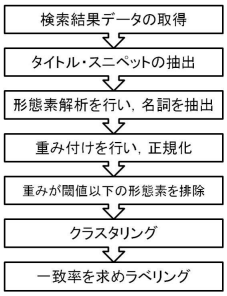

図 1: 提案手法の流れ

2.3 重み付け

Web ページ内の各形態素に重み付けをするために TF-IDF[3] を使用する．この重み $tfidf(t,d)$ を式 (1) に示す．$tf(t,d)$ は Web ページ d に存在する形態素 t の頻度であり，$idf(t)$ は $df(t)$ と N の比の対数である．ここで $df(t)$ は形態素 t が存在する Web ページ数で，N は全 Web ページ数である．これにより，Web ページをよく特徴づける形態素の重要度を高くすることが可能となる．

$$tfidf(t,d) = tf(t,d) \cdot idf(t) = tf(t,d) \cdot \log \frac{N}{df(t)} + 1 \quad (1)$$

続いて，重み $tfidf(t,d)$ を正規化する．正規化した重みを $w(t,d)$ と定義し，式 (2) に示す．$w(t,d)$ は $0 \leq w(t,d) \leq 1$ の値を取る．

次に，閾値を設定し，正規化された形態素の重み $w(t,d)$ が閾値以下となるものを排除する．これにより，Web ページの特徴をあまり表していない形態素を排除することができ，次元数の削減が可能となる．

$$w(t,d) = \frac{tfidf(t,d) - \min(t,d)}{\max(t,d) - \min(t,d)} \quad (2)$$

2.4 クラスタリング

クラスタリングには k-means[4] を使用する．なお，クラスタリングに使用する値は，Web ページに存在する形態素 t の正規化された重み $w(t,d)$ であり，次元数はその形態素数である．

2.5 一致率

一致率 CR を式 (3) に示す．ここで，$C(l)$ は正解ラベルが l となるクラスタ，$R(t,q)$ は検索キーワードで取得した検索結果集合内で，形態素 t を含んでいる Web ページの集合である．$C(l)$ と $R(t,q)$ のどちらにも共通する Web ページの集合が $C(l) \cap R(t,q)$ である．

Num はそれらの集合に含まれる Web ページ数を表している．これにより，クラスタ $C(l)$ に所属する形態素 t の一致率 CR を求める．

$$CR(t, C(l)) = \frac{Num(C(l) \cap R(t,q))}{Num(C(l))} \times 100[\%] \quad (3)$$

2.6 ラベリング

提案手法でのラベリングは，TF-IDF に加え一致率 CR を考慮して行う．ラベリングに使用する形態素 t の重み

表 1: 各閾値でのラベル候補数と正解ラベル平均値

閾値	0.0	0.1	0.2	0.3	0.4
ラベル候補数	52.0	44.9	34.0	23.7	13.7
1 位	53.4%	57.3%	55.2%	56.2%	54.1%
2 位以下	14.4%	7.2%	10.8%	2.3%	3.7%
圏外	32.2%	36.1%	34.0%	41.5%	42.2%

0.5	0.6	0.7	0.8	0.9	1.0
10.7	7.5	5.7	4.8	3.7	2.1
44.6%	39.0%	30.8%	26.3%	19.3%	12.2%
6.0%	3.0%	2.8%	0.9%	0.0%	0.0%
49.3%	58.0%	66.4%	72.8%	80.7%	87.8%

表 2: AKB48 のクラスタ数と正解ラベルの割合

AKB48	0.0	0.1	0.2	0.3	0.4
クラスタ数	43	41	44	37	43
1 位	39.5%	48.8%	52.3%	54.1%	48.8%
2 位以下	18.6%	9.8%	13.6%	2.7%	4.7%
圏外	41.9%	43.9%	34.1%	43.2%	46.5%

0.5	0.6	0.7	0.8	0.9	1
39	40	41	42	42	40
30.8%	35.0%	19.5%	21.4%	11.9%	7.5%
7.7%	7.5%	4.9%	2.4%	0.0%	0.0%
61.5%	57.5%	75.6%	76.2%	88.1%	92.5%

W を式 (4) に示す．クラスタ $C(l)$ 内の Web ページに存在する $w(t,d)$ の総和が，形態素 t の重み $W(t,C(l))$ となる．クラスタ内に同じ形態素が存在した場合，これは，重要な形態素であると考えられる．そのため，同じ形態素の重み $w(t,d)$ を足し合わせることにより，その形態素の重要度を高くすることができる．

提案手法での重みを CRW とし，式 (5) に示す．式 (4) によって求められた形態素 t の重み $W(t,C(l))$ に，式 (3) によって求められた形態素 t の一致率 $CR(t,C(l))$ を掛け合わせて求める．これにより，一致率を考慮した重み付けをすることができる．

$$W(t,C(l)) = \sum_{C(l)} w(t,d) \tag{4}$$

$$CRW(t,C(l)) = CR(t,C(l)) \cdot W(t,C(l)) \tag{5}$$

3. 評価実験

3.1 実験方法

今回の実験では検索キーワードには「AKB48」,「アマゾン」,「地震」,「楽天」,「価格」の 5 つを使用した．これらの検索キーワードは Yahoo!ランキングの上位 20 件以内のものである．検索結果取得数はそれぞれ 100 件である．閾値は 0.0，0.1，0.2，$\cdots$1.0 と変化させる．これにより，ラベルにどのような影響があるかを調査する．

ラベルの評価には作成した正解ラベルを用いる．この正解ラベルの順位で評価を行う．はじめに，生成されたクラスタに所属する Web ページを見てもらう．次に，重み CRW の大きい順に並んだクラスタのラベル候補集合から適切であると考えられるものを選んでもらう．これを正解ラベルとした．正解ラベルがラベル候補集合の一番上にあるならば 1 位とし，それ以下は 2 位以下とした．また，ラベル候補集合に正解ラベルが存在していなかった場合は圏外とした．なお，正解ラベルの作成は本稿の第一著者が行った．

3.2 実験結果と考察

5 つの検索キーワードでのラベル候補数，正解ラベルが 1 位のもの，2 位以下のもの，圏外それぞれの割合を平均した値を表 1 に示す．正解ラベルが 1 位の割合を閾値ごとに見ると，閾値を高く設定した場合，正解ラベルが 1 位の割合が小さくなっている．続いて，正解ラベルが圏外の割合を閾値ごとに見ると，閾値が上がるごとに割合が大きくなっている．これは，Web ページがもつ形態素が，クラスタリングする前に閾値により排除されるためであると考えられる．このことから，閾値を高く設定すると，Web ページの特徴を強く表している形態素が排除され，クラスタに所属する Web ページの内容を表す正解ラベルを付与することができないということが分かる．これに対して，閾値を低く設定した場合，正解ラベルが一位の割合は同程度の割合である．このことから，閾値を低く設定することで，提案手法が有効に作用する可能性が示された．

次に，ラベル候補数の平均を見ると，閾値が上がるごとに，ラベル候補数が減少している．これは，次元数が減ったためである．これにより，クラスタリングにかかる時間を減らすことができると考えられる．

次に，検索キーワード 1 つの結果を見る．検索キーワード「AKB48」のクラスタ数と正解ラベルが 1 位，2 位以下，圏外の割合を表 2 に示す．Web 検索結果のクラスタリングでは，クラスタ数が多すぎても意味がない．そのため，少ないクラスタ数が好ましいと考えられる．クラスタ数を見ると，閾値が 0.3 のときが最も少ないクラスタ数である．さらに，正解ラベルが 1 位である割合が 54.1%と最も高い．この結果から，閾値を 0.3 と設定することで，クラスタ数を減少させつつ，提案手法の精度が向上する可能性が示された．

以上のことから，閾値を 0.3 と低く設定することで，計算にかかる時間やクラスタ数を減少させつつ，提案手法の精度を維持できる可能性が示された．

4. まとめ

本稿では，Web 検索結果のクラスタリングにおけるラベリング手法に閾値を用いることで，どのような影響があるかを調査した．実験では閾値を変化させ，正解ラベルの順位を調査した．実験の結果，閾値を 0.3 と低く設定することで，計算にかかる時間やクラスタ数を減少させつつ，提案手法の精度を維持できる可能性が示された．

今後は，評価の人数を増やして実験を行う．現在は評価を行った人数が 1 人のため，別の人から見た場合，正解ラベルが異なる可能性がある．それにより，より適切な閾値が変動する可能性があるため，人数を増やして正解ラベルとその順位の評価を行う予定である．また，正解ラベルが未知語名詞の場合，クラスタリングは適切に行われていても，形態素解析により正解ラベルが分解され，異なる意味となってしまい，圏外になってしまうものが多かった．そのため，名詞が連続した場合，複合名詞として扱い，実験を行う予定である．

参考文献

[1] 成田宏和, 太田学, 片山薫, 石川博, “Web 文書検索のための非排他的クラスタリング手法の提案, ” DEWS2003 2-P-01

[2] 吉田俊広, 松原雅文, Chakraborty Goutam, 馬淵浩司, “Web 検索結果における一致率を利用したラベリング手法の提案, ” 電気関係学会東北支部連合大会, 2C02, pp.87, 2010

[3] 徳永健伸, “情報検索と言語処理, ” 東京大学出版会, 1999/11/25

[4] 神嶌敏弘, “データマイニング分野のクラスタリング手法 (1) −クラスタリングを使ってみよう !− , ” 人工知能学会誌 18 巻 1 号, 2003 年 1 月