3P-10

一致率を利用した検索結果クラスタへのラベル付け手法の性能評価

吉田 俊広† 松原 雅文‡ Goutam Chakraborty‡ 馬淵 浩司‡
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‡

1 はじめに

現在，インターネットの普及により Web 上の情報が増加してきている．それらの情報を検索する際，多くのユーザはロボット型検索エンジンを利用する．しかし，それらの検索エンジンが返す検索結果は大量である．また，検索結果はリスト形式で返される．そのため，必要な情報であるかどうか，検索結果すべてを評価するにはユーザに多くの負担がかかる．この問題を解決するために，検索結果をクラスタリングして，ユーザに提示するシステムがある [1].

このシステムでは，生成されたクラスタにそれらの内容を示したラベルが付与されている．ユーザはこのラベルを閲覧することで，検索結果にどのような情報が含まれているかを容易に把握することができる．

しかし，クラスタに付与されるラベルがそのクラスタの内容を示していない不適切なものだった場合，ユーザはそのクラスタにどのような内容のドキュメントが含まれているか把握することができない．これではユーザの負担を増やしてしまうことになる．そこで，本研究では一致率を用いた検索結果クラスタへのラベル付け手法を提案している [2][3]．これにより，より適切なラベル付けを目指す．

2 提案手法

2.1 概要

提案手法における処理の流れを図 1 に示す．ロボット型検索エンジンに検索キーワードを入力し，検索結果を取得する．その検索結果から，タイトル・サマリを抽出し，茶筌¶で形態素解析を行う．この形態素解析結果から，品詞が名詞，未知語である形態素のみを抽出する．また，形態素解析結果を利用して TermExtract‖ で複合名詞も生成する．これら形態素，複合名詞をドキュメントの素性として利用する．

次に，素性に $tf \cdot idf$ で重み付けを行い，その重みを Web ドキュメントごとに正規化する．これら正規化された素性を利用して k-means でクラスタリングを行う．最後に，一致率を計算し，これを用いて検索結果クラスタの内容を適切に示したラベル付けを行う．

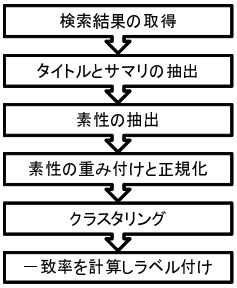

図 1: 提案手法の流れ

2.2 一致率

ある検索キーワードを利用して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検索結果クラスタに付与するラベルが適切であるならば，はじめの検索キーワードとそのラベルを利用して AND 検索した場合，そのラベルが付与される検索結果クラスタに含まれる Web ドキュメントは，その AND 検索結果中に多数含まれるものと考えられる．提案手法では，この一致した度合いを一致率として用いる．

一致率 CR を式 (1) に示す．

$$CR(t, C(l)) = \frac{|\, C(l) \cap R(q,t) \,|}{|\, C(l) \,|} \cdot 100[\%] \tag{1}$$

ここで，$C(l)$ は正解ラベルが l となるクラスタ，$R(q,t)$ は素性 t と検索キーワード q の AND 検索結果を表している．また，クラスタ $C(l)$ と AND 検索結果 $R(q,t)$ に共通する Web ページの集合が $C(l) \cap R(q,t)$ である．これにより，クラスタ $C(l)$ に所属する素性 t の一致率 $CR(t, C(l))$ を求める．

こうすることで，クラスタに所属する Web ドキュメントを多く含む AND 検索結果を取得できる素性 t の重みを大きくすることができる．

2.3 ラベル付け

提案手法でのラベル付けは $tf \cdot idf$ に一致率 CR を掛け合わせた値を用いて行う．ラベル付けに使用する素性 t の重み W を式 (2) に示す．

$$W(t, C(l)) = \sum_{C(l)} w(t,d) \tag{2}$$

クラスタ $C(l)$ 内の Web ドキュメントに存在する素性の正規化された重み $w(t,d)$ の総和が，素性 t の重み $W(t, C(l))$ となる．クラスタ内に同じ素性が多く存在した場合，これは，重要な素性であると考えられる．そ

Performance Evaluation of Labeling Method Using Concordance Ratio for Clusters of Search Results
†Toshihiro YOSHIDA: Graduate School of Software and Information Science, Iwate Prefectural University Graduate School
‡Masafumi MATSUHARA, Goutam CHAKRABORTY, Hiroshi MABUCHI: Faculty of Software and Information Science, Iwate Prefectural University

¶ChaSen – 形態素解析器, http://chasen-legacy.sourceforge.jp/
‖TermExtract, http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

のため，同じ素性の重み $w(t,d)$ を足し合わせることにより，その素性の重要度を高くすることができる．

提案手法の重みを CRW とし，式 (3) に示す．

$$CRW(t,C(l)) = CR(t,C(l)) \cdot W(t,C(l)) \tag{3}$$

式 (2) によって求められた素性 t の重み $W(t,C(l))$ に，式 (1) によって求められた素性 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を付けることができる．

この重み CRW を利用して，クラスタにラベル付けを行う．

3 評価実験

3.1 実験データ

実験には Open Directory Project*（以下 ODP）のディレクトリ「レクリエーション」以下の Web ドキュメントを利用する．このディレクトリ以下の Web ドキュメントのうち，タイトル・サマリにディレクトリ名を含んでいる Web ドキュメントのみを実験データとして利用する．

正解ラベルには，Web ドキュメントが所属しているディレクトリのディレクトリ名を利用する．

3.2 評価方法

評価では ODP のディレクトリ名を正解ラベルとして使用する．正解ラベルとして用いるディレクトリ名は，クラスタに所属する Web ドキュメントの 80%以上がもつものとする．クラスタの Web ドキュメントに共通するディレクトリ名がない場合は，そのクラスタを評価しない．

評価は再現率と精度によって行う．再現率は式 (4) によって求められる．精度は式 (5) によって求められる．ここで，R は指定した順位までの正解ラベルの個数である．N は全正解ラベルの個数である．C は指定した順位までのラベル候補の個数である．

$$recall = \frac{R}{N} \tag{4}$$

$$precision = \frac{R}{C} \tag{5}$$

評価における比較手法として，David らの Wikipedia の知識ベースを活用してクラスタに付与されるラベルを強化する手法で利用されている，$ctf \cdot cdf \cdot idf$ を利用する [4].

3.3 実験結果と考察

評価可能なクラスタ数は 573 となり，全正解ラベルの個数は 725 となった．

再現率と精度の結果を表 1 に示す．再現率の結果において，第 4 位までは $ctf \cdot cdf \cdot idf$ よりも提案手法のほうが高い．しかし，第 5 位以降を含めた場合は，提案手法の再現率は $ctf \cdot cdf \cdot idf$ よりも低い．一致率は，その値が低いラベル候補を下位に下げる性質がある．つまり，ラベル候補の下位に存在した正解ラベルの一致率が低く，さらに下位に下がってしまったため，再現率が低下したと考えられる．

表 1: 再現率と精度

順位	再現率		精度	
	提案手法	$ctf \cdot cdf \cdot idf$	提案手法	$ctf \cdot cdf \cdot idf$
1	48.70%	45.80%	50.90%	45.00%
2	11.90%	10.80%	21.90%	18.80%
3	7.70%	10.60%	13.60%	15.50%
4	5.10%	5.90%	9.50%	9.10%
5	4.80%	5.50%	9.60%	7.10%
6	3.30%	4.70%	7.80%	6.30%
7	1.80%	2.60%	5.30%	5.30%
8	1.40%	1.90%	7.50%	3.40%
9	1.50%	1.90%	4.70%	4.20%
10	0.40%	1.20%	8.30%	3.50%
11 ~	13.40%	9.00%	4.10%	2.60%

精度の結果では，第 3 位，第 7 位以外で提案手法の精度が $ctf \cdot cdf \cdot idf$ よりも高い．これは，一致率によって，正解ラベル以外のラベル候補を下位に下げたためと考えられる．このことから，一致率の重要度を大きくすれば，上位の精度がさらに向上すると考えられる．

再現率と精度両方の結果から，提案手法は少ないラベル候補数で正解ラベルを上位に残すことが可能なことを示唆している．以上のことから，提案手法は検索結果クラスタへのラベル付けに有効である．

4 まとめ

本稿では，一致率を用いた検索結果クラスタへのラベル付け手法を提案し，再現率と精度でその性能評価を行った．評価実験の結果，検索結果クラスタへのラベル付けに一致率を用いるのは有効であることが示された．

今後の予定として，一致率の重要度を大きくし，さらなる再現率と精度の向上を目指す．また，クラスタにいくつのラベルを付与するのが適切か調査を行う予定である．

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”, 第 14 回データ工学ワークショップ (DEWS2003) 論文集, 2-P-01

[2] 吉田俊広, 松原雅文, Chakraborty Goutam, 馬淵浩司: “Web 検索結果のラベリングにおける閾値の利用について” FIT2011 第 10 回情報科学技術フォーラム講演論文集, E-063, pp.365-366, September 2011.

[3] Toshihiro Yoshida, Masafumi Matsuhara, Goutam Chakraborty and Hiroshi Mabuchi: “A Novel Ranking Method of Web Search Result Using Clustering and Concordance Count” Proc. of WCCI 2012 IEEE World Congress on Computational Intelligence, pp.902-907, Brisbane, Australia, June, 10-15, 2012.

[4] David Carmel, Haggai Roitman and Naama Zwerdling: “Enhancing Cluster Labeling Using Wikipedia”, *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp.139-146, 2009.

*Open Directory Project, http://www.dmoz.org/