E-003

大規模ウェブ情報クラスタリングにおけるラベルの組織化

Label Organization in Large-scale Web Information Clustering

原島 純　柴田 知秀　新里 圭司　黒橋 禎夫
Jun Harashima　Tomohide Shibata　Keiji Shinzato　Sadao Kurohashi

1　はじめに

現在ウェブには膨大な情報が蓄積されており，その量は日々増えつづけている．この中から有用な情報を効率よく取得するためには検索エンジンの利用が必要不可欠である．しかし既存の検索エンジンは検索結果をリスト形式で提示するのみに止まっている．このようなリスト形式の提示結果では，検索結果全体に含まれる内容を俯瞰するのが難しく，またリストの下位に埋もれた有用な情報を見逃す危険性がある．そのため近年，検索結果を内容ごとにクラスタリングするシステムの研究・開発が盛んである[1] [2]．

検索結果をクラスタリングするシステムでは，ユーザは各クラスタに付与されたラベルを見て求める情報を探索するため，ラベルの質がシステムの利便性を大きく左右する．そのため，まず検索結果から可読性の高いラベルを抽出し，ラベルが出現した文書集合をクラスタとするラベルベースの手法が主流である．しかし抽出したラベルを単純に並べて提示するだけでは，クラスタリング結果は単なるクエリの関連語のリスト程度の価値しか持たず，また提示している意味内容が理解しにくいことが多い．

以上を踏まえ，本研究ではラベルベースの手法によって抽出されたラベルを組織化し，ユーザに有用かつ理解しやすい情報を提示することを目指す．具体的には，(i) トピックによるラベルのグルーピング，(ii) トピック中における重要なラベルのグルーピング，という 2 つの軸のグルーピングを行うことで，クエリに関する情報を組織化して提示するシステムを提案する．

2　提案システムの概要

本節では我々が開発中のクラスタリングシステムの概要を述べる．提案システムではラベルベースの手法を用いて検索結果をクラスタリングしている [1]．以下に提案システムの主な特徴を挙げる．

京都大学大学院 情報学研究科 知能情報学専攻

[1] Clusty:http://clusty.jp/

[2] Mooter:http://www.mooter.co.jp/

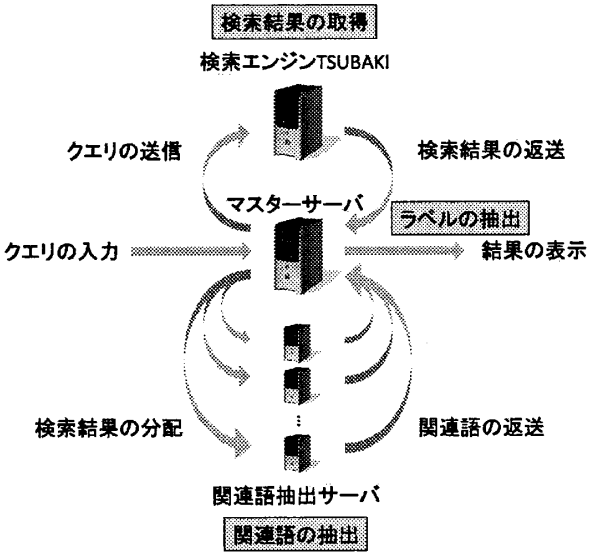

図 1: 提案システムの概要

開放型検索エンジン基盤 TSUBAKI の利用　既存のシステムの多くは商用の検索エンジンを用いて検索結果を取得している．そのためラベルの抽出対象として，数百件程度の文書の 2 ～ 3 文のスニペット，計数百文しか利用できない．これに対して提案システムでは，開放型検索エンジン基盤 TSUBAKI[2] と連携することで，数千件規模の文書のそれぞれから抽出したクエリと関連が深い数十文，計数万文をラベルの抽出対象とすることが可能である．

高度な言語処理の活用　既存のシステムは単純に高頻度の表現をラベルとして採用している．これに対して提案システムでは，ラベルを抽出する過程において，表記揺れの解消，同義表現のマージなどの高度な言語処理を行っている．

このように既存のシステムは少量のテキストに対して単純な処理を行っているため，ラベルの抽出精度が低い．これに対して提案システムでは上記で述べたように，大量のテキストに対して高度な言語処理を行うことで，十分かつ適切なラベルの抽出を実現している．

以下に提案システムが行う処理の流れを示す (図 1).

Step 1. 検索結果の取得 TSUBAKI を用いてクエリに対する検索結果を取得する.

Step 2. 関連語の抽出 検索して得られた各文書を複数の関連語抽出サーバに分配する. 各文書からクエリと関連が深い文を抽出し, 文中の名詞句や鍵括弧で囲まれた文字列をクエリの関連語として抽出する.

Step 3. ラベルの抽出 各サーバで抽出された関連語をマスターサーバで集約し, 表記揺れの解消 (「子供」[3]と「子ども」), 同義表現のマージ (「赤ちゃん」と「赤ん坊」) などの高度な言語処理を行う. 以上の処理で精錬された各関連語 w に対して次式を用いてスコア付けを行い, 上位 n 語をラベルとして採用する. 現在は $n = 300$ としている.

$$score(w) = ldf(w) * log\frac{N}{gdf(w)} \qquad (1)$$

ただし N は TSUBAKI が検索対象とする文書数 (= 1 億), $ldf(w)$ は検索結果中における w の文書頻度, $gdf(w)$ は TSUBAKI が検索対象とする 1 億文書中における w の文書頻度とする.

3 ラベルの組織化

本研究では前節 Step 3 で述べたラベルの抽出に続く処理として、ラベルの組織化を行う. この処理では, (i) トピックによるラベルのグルーピング, (ii) トピック中における重要なラベルのグルーピング、という 2 つの軸のグルーピングを行うことで, ラベルを組織化する.

3.1 トピックによるラベルのグルーピング

一般に, 検索結果中には複数の異なるトピックに関する文書が混在している. 特にクエリに曖昧性がある場合, このような現象は顕著である. 例えば TSUBAKI を用いて『ガンバ』を検索すると, ガンバ大阪 (サッカーチーム), ヴィオラ・ダ・ガンバ (楽器), ガンバの冒険 (アニメ), 家庭教師のガンバ (人材派遣業者) などの複数のトピックに関する文書が一緒に検索される. また『ゆとり教育』などの曖昧性がないクエリにおいても,「ゆとり」のみ, もしくは「教育」のみが出現するゆとり教育とは関係ない文書が検索される可能性がある. このような検索結果から抽出したラベルをそのまま提示すると, 異なるトピックに関するラベルが混在し, ユーザに混乱を与える結果を提示してしまう恐れがある. そこで, まず抽出したラベルをトピックごとにグルーピングする.

同じトピックに関するラベルは同じ共起語を持ちやすいと考えられる. 例えばガンバ大阪に関するラベル「宮本恒靖」「加地」は共に「浦和レッズ」「優勝争い」などのラベルと共起している. そこでラベル間の共起関係を用いることで各ラベルをトピックごとにグルーピングする. この処理は次の 2 つのステップからなる.

Step 1. 各ラベルを表すベクトルの作成 各ラベルを表すベクトルを作成する. ベクトルの要素には他のラベルとの共起頻度を用いる. 例えばラベル l_k がラベル l_1, l_2, l_3, l_4, … とそれぞれ 2 回, 0 回, 1 回, 2 回, … 共起していたとき, l_k を表すベクトル $\boldsymbol{v_k}$ は次のようになる.

$$\boldsymbol{v_k} = (\ 2,\ 0,\ 1,\ 2, \cdots\ ) \qquad (2)$$

Step 2. ラベルのクラスタリング ラベル集合全体を 1 つのクラスタとみなした状態から始めて, 再帰的にクラスタを分割する. 具体的には次のサブステップを繰り返す.

Step 2-1. クラスタ c に含まれるラベルの中から, 式 (1) のスコアがもっとも高いラベル l_a を取得する.

Step 2-2. l_a と c 内の他のラベルの *cosine* 類似度を計算し, l_a との類似度がもっとも低いラベル l_b を取得する. l_a と l_b の類似度が閾値以上であれば c を一定以上の類似度を持つラベルの集合 (トピック) とみなし、クラスタの分割を終了する. 閾値未満であれば Step 2-3 に移る. 閾値は 0.2 とした.

Step 2-3. c 内に含まれる l_a, l_b 以外のすべてのラベルに対して, l_a, l_b との類似度を計算し, 類似度の高い方のラベルにグルーピングする. 生成された 2 つのグループをクラスタとみなし, 両クラスタに対して Step2-1 から Step 2-3 を行う.

以上の処理を行うことで各ラベルをトピックごとにグルーピングする.

3.2 重要なラベルのグルーピング

前節の処理により各ラベルをトピックごとにグルーピングした後, 各トピックにおいて重要なラベルをグルーピングする. グルーピングは次の 4 つの観点から行う.

[3]以降本稿では,「」で囲んだ文字列はラベル,『』で囲んだ文字列はクエリ,【】で囲んだ文字列はグループ (3.2 節で述べる) を表すものとする.

表 1: ラベルをエンティティとみなす条件

エンティティ	固有表現	カテゴリ	分類
【人・主体】	PERSON	人	人名
【組織・団体】	ORGANIZATION	組織・団体	-
【場所】	LOCATION	場所-施設 場所-自然	地名

クエリを含むラベルのグルーピング クエリを含むラベルは重要であるとみなし，このようなラベルが m 個以上存在したとき，これらをグルーピングする．現在は $m = 3$ としている．

エンティティによるグルーピング 【人・主体】【組織・団体】【地名】などのエンティティを表すラベルは，各トピックにおいて重要な役割を持っていると考えられる．そこで各ラベルに対して形態素解析[4] 及び固有表現解析 [3] を行い，各エンティティを表すと判定されたラベルをグルーピングする．ラベルがエンティティを表すとみなされる条件は表 1 のとおりである．例えばラベルが【人・主体】を表すとみなされるのは，(i) 固有表現解析の結果 PERSON と判定される，(ii) JUMAN の辞書にカテゴリ体系が人であると付与されている，(iii) JUMAN の分類が人名である，のいずれかの条件を満たした場合である．各ラベルを解析した結果，各エンティティを表すラベルが m 個以上存在した場合，これらをグルーピングする．

共有形態素列によるグルーピング 末尾の形態素列，もしくは先頭の形態素列が同じラベルが m 個以上存在していたとき，この形態素列はクエリに対して重要な役割を持つとみなし，これらのラベルをグルーピングする．

スコアによるグルーピング 式 (1) のスコアが高いラベルについては，上記のいずれのグループにも属さなかったとしても，重要であるとみなす．具体的には，上記のグルーピングを行った後，スコア上位 10 ラベルの中で上記のいずれのグループにも属さなかったラベルを【重要キーワード】にグルーピングする．また上位 11 ～ 30 ラベルの中で，上記のいずれのグループに属さなかったラベルを【その他】にグルーピングする．

以上の処理を行うことで各トピックにおいて重要なラベルをグルーピングする．

4 実験

まずトピックによるラベルのグルーピング精度について提案手法の有効性を調査した．具体的には，検索結果中に複数のトピックについての文書が混在するクエリ『ガンバ』『コナン』『イースター』『ACL』『マック』に対して，提案手法を用いて各ラベルをグルーピングし，同じトピックのラベルをグルーピングできたか，その正否を人手で判定した．判定には各クエリに対してランダムに選択した 100 ラベルを用いた．ただし，複数のトピック $t_1, \cdots, t_n$ にグルーピングされうる曖昧なラベル l(ex.「写真」「記事」「プロフィール」などの様々なトピックに出現しうるラベル) については，l が出現する文書集合において，半分を越える文書が t_k に関する文書である場合は正解を t_k とし，そのような t_k が存在しない場合は l を評価の対象外とした．実験結果を表 2 に示す．

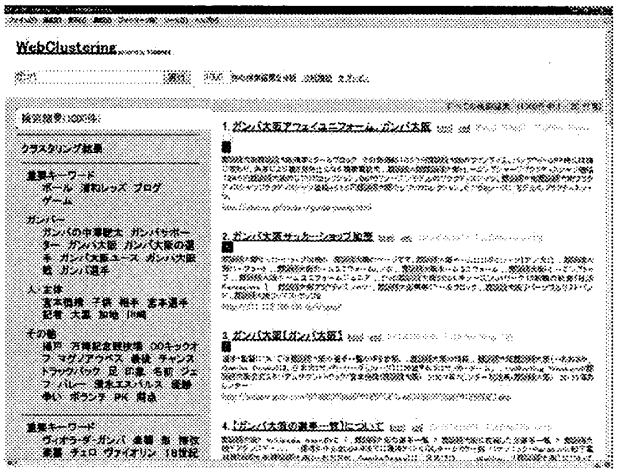

図 2: クエリ『ガンバ』をクラスタリングした結果

表 2 から『ガンバ』『コナン』『イースター』については 80% を越える高い精度で各ラベルをグルーピングできた．一方で『ACL』『マック』については 50.6%, 33.3% とそれぞれ低い精度が示された．原因を調査した結果『ガンバ』『コナン』『イースター』に対する検索結果中に含まれていたトピック数が 4～7 個であった[5]のに対して,『ACL』は 13 個,『マック』では 41 個もの異なるトピックが含まれていたことがわかった．抽出するラベル数は固定 (2 節 Step3) しているため，トピック数が多くなるにつれ，各トピックに所属すべきラベルは少なくなる．各トピックに所属すべきラベルが少なくなると，これらのラベル間における共起語情報が乏しくなり，結果としてグルーピングの精度が落ちたものと思われる．このように，著しく曖昧性のあるクエリについては検索結果から得られる情報だけでは正確なグルーピングが難しい．それゆえ，外部から獲得される知識を利用して，グルーピングをサポートする必要があると思われる．具体的には Wikipedia

[4] JUMAN:http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[5] TSUBAKI を用いて取得した 1,000 件の検索結果中で著者が観測できた範囲でのトピック数.『ACL』『マック』についても同様.

表 2: トピックによるグルーピング精度

クエリ	分類精度 (%)	検索結果中に含まれるトピックの例	トピック数
『ガンバ』	99.0 (90/91)	ガンバ大阪, ヴィオラ・ダ・ガンバ, ガンバの冒険, …	7
『コナン』	99.0 (97/98)	名探偵コナン, 未来少年コナン, アーサー・コナン・ドイル, …	4
『イースター』	84.2 (80/95)	イースター島, イースター祭, イースター株式会社, …	6
『ACL』	50.6 (45/89)	アクセス制御リスト, アジアチャンピオンズリーグ, …	13
『マック』	33.3 (24/72)	マッキントッシュ, マクドナルド, 株式会社マック, …	41

表 3: 各トピック及び 3.2 節の各観点によるグルーピングの結果

トピック	各トピックに分類されたラベル				
	【クエリ】	エンティティ【人・主体】	【共有形態素】	スコア【重要キーワード】	【その他】
ガンバ大阪	ガンバ大阪, ガンバサポーター	宮本恒靖, 加地, 大黒	-	浦和レッズ, ボール, ゲーム	優勝争い, ボランチ
ヴィオラ・ダ・ガンバ	-	平尾雅子, 小澤絵里子	-	擦弦楽器, ヴァイオリン	調弦法, 歴史, 魅力
ガンバの冒険	-	斎藤惇夫, 出崎統	カワウソの冒険, グリックの冒険	ノロイ, アニメ, 十五匹の仲間	テーマ, マンプク

の曖昧さ回避のページなどの利用を検討中である.

重要なラベルをグルーピングする有用性を調査するため『ガンバ』についてクラスタリングした結果を図 2 に示す. また参考のため,『ガンバ』の各トピック及び 3.2 節の各観点に基づいてグルーピングされたラベルの一例を表 3 に示す. 図 2 から, 例えば「浦和レッズ」がガンバ大阪に関する【重要キーワード】であること,「宮本恒靖」や「加地」がガンバ大阪に関連する【人・主体】であることがすぐにわかる. ラベルを単純に並べて提示するだけではこのようなことは難しく, 提案手法の有用性が示されていると言える.

5 関連研究

文書集合から抽出した単語を組織化する研究として, Stoica らによるファセットを自動で構築する研究が挙げられる [4]. Stoica らは WordNet において同じ上位語を持つ単語をグルーピングすることで, 単語を抽出した文書集合の内容を反映したファセットを自動で構築している. これらのファセットは Flamenco というクラスタリングシステムで実際に利用されている [5]. Flamenco ではファセットを利用することでクエリに関する情報を様々な観点から検索することができる.

鳥澤らは鳥式というシステムを開発している [6]. 鳥式では抽出したラベルを, 各ラベルのマイナー度とラベル間の類似度を用いてグラフ状に可視化している. また藤井らは時事問題に関するクエリを対象とし, 抽出したラベルを賛否の軸で分類・可視化する OpinionReader というシステムを開発している [7]. 具体的には各ラベルが賛成意見中に出現する確率と反対意見中に出現する確率を比較することで, ラベルを賛否に分類し, 2 次元グラフ上に可視化している.

6 おわりに

本稿では検索結果をクラスタリングするシステムにおいて, 抽出されたラベルを組織化する必要性を述べ, その手法を提案した. 提案手法では, ラベル間の共起関係を用いてラベルをトピックごとにグルーピングした後, 各トピックにおける重要なラベルをグルーピングする. トピックによるラベルのグルーピング精度を測った結果, 著しく多義のクエリを除いて, 高い精度でグルーピングできることが確認された. また, 各トピックにおける重要なラベルのグルーピングについてもその有用性が確認できた. 今後の課題として, ラベルのグルーピングに止まることなく, 情報をより正確に伝達可能な表現形態を模索する必要がある. 具体的には, ラベル間の関係の同定, 検索結果に対する複数文書要約などについて検討中である.

参考文献

[1] 馬場康夫, 新里圭司, 黒橋禎夫. 検索エンジン基盤 TSUBAKI を用いた大規模ウェブ情報クラスタリングシステムの構築. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2008, No. 4, pp. 67–74, 2008.

[2] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. *In Proceedings of IJCNLP*, pp. 189–196, 2008.

[3] Ryohei Sasano and Sadao Kurohashi. Japanese named entity recognition using structural natural language processing. *In Proceedings of IJCNLP*, pp. 607–612, 2008.

[4] E. Stoica., M. Hearst, and M. Richardson. Automating creation of hierarchical faceted metadata structures. *In Proceedings of NAACL/HLT*, pp. 244–251, 2007.

[5] M. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, Vol. 49, No. 4, pp. 59–61, 2006.

[6] 鳥澤健太郎, 隅田飛鳥, 野口大輔, 風間淳一. 自動生成された検索ディレクトリ「鳥式」の現状. 言語処理学会 第 14 回年次大会 発表論文集, pp. 729–732, 2008.

[7] 藤井敦. Opinionreader–意思決定支援を目的とした主観情報の集約・可視化システム. 電子情報通信学会論文誌. D, 情報・システム, Vol. 91, No. 2, pp. 459–470, 2008.