D-006

# ラベル指向情報検索における分類ラベル統合方式の検討

## A Label Unification Method for Label-Based IR-System using Informatively Named Entities

向井 景洋 †　戸田 浩之 †　片岡 良治 †
Kagehiro Mukai　Hiroyuki Toda　Ryoji Kataoka

## 1．はじめに

Ｗｅｂサーチエンジンを始めとする全文検索システムにおいて，ユーザが検索結果から必要な情報を探し出すコストが大きいという問題がある．近年，この問題への１つの対処として，検索結果を分類し提示する手法が研究されている．その一手法として我々は検索結果から動的に「特徴的な固有表現」を抽出し，これを検索結果に対するインデクス（分類ラベル）として提示することにより，検索結果内容の概観性を向上させる手法を提案している．本手法では，検索結果から重要な固有表現をラベルとして動的に選出しているが，個々のラベルを独立に評価する為，提示するラベル集合内に意味的に重複するラベルを含んでいる．本稿では，ラベルの表層情報，共起情報を複合的に利用し，冗長なラベル同士を動的に統合し，更に概観性を向上させる動的分類ラベル統合方式を提案する．

## 2．問題点の抽出

### 2.1　固有表現を用いたラベル指向情報検索

膨大な情報からユーザに適確な情報を提供するナビゲーション手段として検索システムが普及しているが，一般的に広く利用されているキーワード入力型の検索システムでは，以下の問題点が知られている．

- 検索結果リストからユーザが必要な情報を選別するコストが大きいこと
- 検索結果集合全体を把握し難いこと（概観性の悪さ）

我々のグループでは，この２つの問題点に対し，検索結果集合から抽出した固有表現[1]をラベルとして用い，検索結果を動的に分類することで概観性及びユーザの選別コストを改善するラベル指向情報検索手法を提案し，新聞記事を対象とした検索において，従来手法と比較し精度の良いナビゲーションが可能であるとの結果を得ている．[2]

### 2.2　ラベル指向情報検索の問題点

ラベル指向情報検索では，ユーザの与えた検索キーワードにより選出された文書集合から，「人物」，「場所」，「組織」といったカテゴリ毎に，ユーザナビゲーションに効果的な固有表現をラベルとして選出し，そのラベルに基づき階層的に検索結果を提示する．本手法では，検索時に個々のラベルの独立な評価に基づきラベル選出を行っている為，現状の新聞記事検索において生成されたラベルには以下に示す冗長性を含む．

１つ目は，同義ラベルの冗長性である．同義ラベルの種別は，表１に例示しているようないくつかの原因に基づき生じた，完全に（グローバルに）同義な関係にあるラベルに基づく冗長性と，“田中”と“田中真紀子”及び“田中耕一”あるいは，“東京都”と“東京都千代田区”の関係

表 1. グローバルに同義なラベル

| ラベルの関係 | ラベル例 |
|---|---|
| 和名／英略名 | “国際原子力機関：IAEA”, “欧州連合：EU” |
| 正式名称／略名 | “日本長期信用銀行：長銀” |
| ひらがな／カタカナ／漢字 | “米国：アメリカ”, “金正日：キムジョンイル” |
| 表記ゆれ | “横須賀市：神奈川県横須賀市”, “ゼネラル・モーターズ：ゼネラル・モーターズ社” |

といった意味的に包含関係を持ち，状況に応じて（ローカルに）同義な関係にあるラベルに基づく冗長性の 2 種類に大きく分類される．

ユーザナビゲーション（概観性）の観点から考えると，グローバルに同義な関係にあるラベルについては常に統合されることが期待され，ローカルに同義な関係にあるラベルは状況に応じた統合が期待される．つまり，前述の例をもとに説明すると，与えられた検索キーワードが“政治”の場合と“研究”の場合で，“田中”を“田中真紀子”と“田中耕一”のどちらと統合するか，あるいは検索キーワードが“日本”の場合であればどちらとも統合しない，といった状況に応じた判断に基づくラベル統合が必要となる．

２つ目は，文書集合に着目したラベルの冗長性である．いま説明を行った冗長性は，ラベル自体の冗長性であるが，ラベル指向情報検索では非排他的なクラスタリング手法を使用しており，一つの文書が複数のラベルに含まれることを許容している為，利用者が検索を進め，話題が絞り込まれた場合に，異なる複数のラベルが含む文書がほぼ同じ文書集合となる場合があるという問題点がある．ユーザの本システムにおける検索ログの分析を行った結果，現状のラベル指向情報検索では，ユーザが正解集合を閲覧した回数の約半分が既閲覧文書であることが判った．

## 3. 提案手法

### 3.1 アプローチ方針

前章で説明したローカルに同義であるラベルの統合を行う為には，各ラベルに含まれる文書情報を考慮した検索毎の判断に基づく手法を考える必要がある．グローバルに同義であるラベルについては，予め辞書を用意し，静的に定義することも可能であるが，現状利用を想定している新聞記事等のニュースコンテンツを対象とした検索の場合，新語が次々に出現することを考慮すると，動的に自動で行う手法が望ましい．

各ラベルに含まれる文書情報を考慮した手法としては，ラベル内の文書集合の内容（文書ベクトル等）の分析に基づく手法と，ラベル内の文書集合自体に着目する手法が考えられる．今回は，

- 和名／英略名や正式名称／略名などは言い換えの形で同一文書に出現する可能性が高い

† 日本電信電話株式会社　NTT サイバーソリューション研究所

・検索毎にラベル統合の判断が必要である為，処理的に軽い手法である必要がある
・固有表現をキーとした分類を行っている為，文書ベクトルなどの内容による類似判定は適さない

という事項から，ラベルに含まれる文書集合自体に着目し，文書の重複度合いに基づく手法をベースに検討を行う．また，重複度合いに基づく手法とすることにより，前述の文書集合に着目したラベルの冗長性も併せて低減することが期待できる．

### 3.2 文書重複率に基づくラベル統合手法

文書重複率に基づくラベル統合手法の基本的な考え方としては，検索時に生成されたラベルAとラベルBに含まれる文書数をそれぞれ $Ra$, $Rb$, 両ラベルに含まれる文書数を $Rab$ とし以下の条件を満たすラベル同士を統合対象とする．

ラベル統合条件：$\frac{R_a}{R_{ab}} > T$ ,かつ $\frac{R_b}{R_{ab}} > T$　$T$：閾値

上記条件を用いることにより，文書集合に着目したラベルの冗長性は排除可能となる．次に本手法による同義ラベルの統合の精度検証を行った．実験は毎日新聞 94, 95 年の記事を対象とし，記事に対して IREX で作成されたトピック 30 件に対し，複数のユーザに検索してもらう．その際に出力されたラベル集合に関し，表1に提示したような同じ実体を表すラベル同士を同義ラベルと考え，正解集合とした．また，ローカルに同義と考えられるラベルは含まれる文書内容を閲覧し，80%以上の文書が実体として同じものを示している場合のみを同義ラベルとした．正解件数は全部で 313 件であった．本手法による同義ラベルの統合に関する適合率及び再現率を表2に示す．

表2．文書重複率に基づくラベル統合結果

| 単位（%） | 閾値 T=0.8 | 閾値 T=0.6 |
|---|---|---|
| 適合率 | 37 | 27 |
| 再現率 | 6 | 11 |

本手法では，文書数に大きく差があるラベル同士の場合，統合対象となりにくい手法である為，再現率が低くなっている．この問題を解決し再現率を上げる為には，閾値を緩和する必要があり，それを行った場合，適合率が低下する．

### 3.3 改良手法

前節で示した問題を解決する為に，上記手法をベースとし，何らかの確度の高い外部情報を補助情報として利用し，集合数が異なる場合にラベル統合を行うべきか否かの判断を補助し，さらに確度に併せて閾値を緩和する手法を提案する．以下にその手法について説明する．

#### 3.3.1 表層情報を利用した手法

本手法では，ラベル自体の情報を補助情報として利用し，ラベルの表層情報が近ければ近いほどラベル同士が同義である可能性が高いと考え，表層情報によりラベル間の類似度を判定し，その類似度を確度と捉え，3.2 節で提示した手法の閾値 $T$ を確度により可変とする．類似度判定の尺度としては，編集距離を利用する方法もあるが，今回は簡易の為，前方一致及び後方一致に基づく一致度を考える．

ラベルA，Bにおいて，ラベルAがラベルBの文字列を含む場合，一致したラベル B の文字長 $L_b$ を確度とする。また，閾値Tは，

$T = 1.0 - \alpha * L_b$　($\alpha = 0.30, 4 > L_b > 1$)

とし，片方のラベルが統合条件を満たせば，ラベル統合の対象とする．また，$L_b > 3$ の場合は，表層情報のみの判断により統合を行う．式中の係数 $\alpha$ は予備実験により決定した．

#### 3.3.2 文書集合全体の共起情報を利用した手法

文書集合全体において相互に関係の深いラベル同士は，ローカルな状況においてもある程度の関連性をもつ可能性が高いと考え，集合数が異なる場合にも統合対象として考える．ラベルAとラベルBが文書集合全体において，相互に高い共起傾向を持つ場合，閾値 $T$ を片方が満たせば，統合対象とする．高い共起傾向の判断については，各ラベルが文書集合全体において高い共起傾向を示すラベル情報を数個保持し，その情報をもとに判断を行う．今回の実験では，予備実験により精度の高かった T=0.8 で検証を行った。

#### 3.3.3 ハイブリッドな手法

表層情報及び文書集合全体の共起情報の双方を利用したハイブリッドな手法を考える．各手法は前節までの説明に従う．

## 4. 結果

前章で説明したそれぞれの手法に関し 3.2 節で説明した実験による比較を行った．その結果を図1に示す．

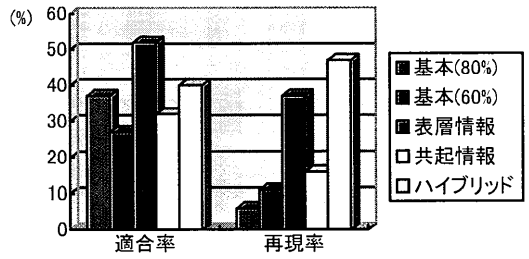


図1．適合率と再現率の比較

図1に示されるように，基本手法と比較し，改良手法が再現率の面において大幅に改善されていることがわかる．特にハイブリッドな手法では，約半数近くの同義ラベルが統合されている．また，提案した3つの改良手法の再現率を比較すると，表層情報に基づく手法と共起情報に基づく手法では，抽出される同義ラベルが大部分において排他的な関係にあり，ハイブリッド手法では双方の手法が効率よく利用出来ていることが判る．これは，表層情報に基づく手法が主に表記ゆれの関係にあるラベル同士の統合に有効なのに対し，共起情報に基づく手法は主に和名／英略名や正式名称／略名に関するラベルの統合に有効である為と考えられる．

## 5. まとめ

本稿では，ラベル指向情報検索において検索毎に生成されたラベルの冗長性を整理し，その冗長なラベルを統合する手法として，文書重複率をベースとし，表層情報及び文書集合全体の共起情報を利用したラベル統合手法を提案した．また，94, 95 年の新聞記事を利用し評価を行った結果，再現率において大幅な改善が実現され，本手法が有効であることを示した．今後は，被験者による情報ナビゲーションの観点からの効果の検証を行う予定である．

### 参考文献

[1] 磯崎秀樹，賀沢秀人: “固有表現抽出のための SVM の高速化” 情報処理学会論文誌 Vol.44, No.3, 2003.
[2] 戸田浩之，長浜光俊，片岡良治: “特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案” 情報処理学会研究報告 2004-FI-75, pp. 99-106, (2004)