# URLを用いた検索結果の分類手法

山下真理子† 鈴木優‡ 川越恭二‡

†立命館大学大学院 理工学研究科 ‡立命館大学 情報理工学部

## 1 はじめに

現在，利用者に Web 検索エンジンの検索結果をわかりやすく提示するために，類似した検索結果の Web 文書を分類して提示する研究 [1] が行われている．既存研究では，Web 文書のタイトルや文書内容から特徴ベクトルを作成し，特徴ベクトル間の類似度にもとづいてクラスタリングを行う手法が用いられている．しかし，この手法では文書内容の解析を行うため分類精度は向上するが，Web 文書数が増加するほど処理時間が膨大になるという問題がある．

そこで本研究では，Web 文書の内容と URL の相関関係に着目する．なぜならば，URL のドメイン名やファイル名には文書内容を表す文字列が含まれている可能性が高いと考えるためである．そこで，トライグラム [2] によって Web 文書の URL から抽出した文字列を用いて特徴ベクトルを作成し，特徴ベクトル間の距離にもとづいてクラスタリングを行う手法について提案を行う．その結果，Web 文書の内容を用いたクラスタリングに対して，分類精度は同程度のまま処理時間を削減することが可能であると考えられる．

## 2 URL を用いた検索結果の分類

本研究では，検索結果の Web 文書を分類する際に URL 間の類似度を用いる．なぜならば，URL と Web 文書には相関関係があると考えるためである．例えば，立命館大学トップページの URL である “http://www.ritsumei.jp/index_j.html”に含まれる，“ritsumei”は立命館に関連し，“index”はトップページに関連している Web 文書であることが推測できる．このように，URL には Web 文書の内容を表す文字列が含まれていると考えられる．そこで，トライグラムを用いて URL から抽出した文字列を特徴語として特徴ベクトルを作成する．作成した特徴ベクトルを用いて URL 間の距離を算出し，その距離にもとづいて Web 文書のクラスタリングを行う．

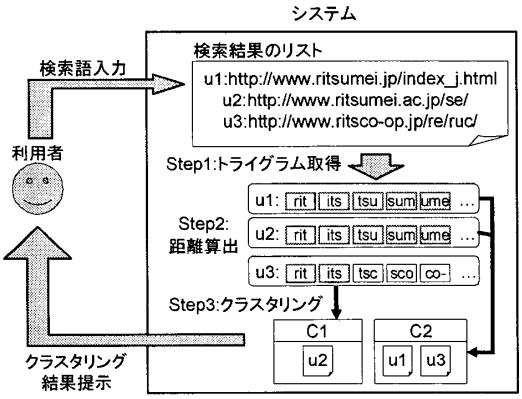


図 1: システム概要

### 2.1 システムの概要

提案手法を用いたシステムの概要を図 1 に示す．まず，利用者が入力した検索語の検索結果を取得し，トライグラムにより各 URL から文字列を抽出し，特徴ベクトルを作成する．次に，特徴ベクトルから URL 間の距離を算出し，URL 間の距離にもとづいて各 Web 文書をクラスタリングする．そして，クラスタリング結果を利用者に提示する．

### 2.2 トライグラムを用いた URL の特徴語抽出

URL に含まれる “http://”以降のドメイン名とファイル名を対象として，先頭から一文字ずつずらしながら三文字ずつの文字列を取得する．取得した文字列を特徴語とし，各特徴語の重みを要素とする URL の特徴ベクトルを作成する．

$N$ 件の検索結果に含まれる Web 文書の各 URL $u_n(n = 1, 2, \cdots, N)$ に出現する $M$ 個の特徴語 $t_m(m = 1, 2, \cdots, M)$ を，重複が無いように取得する．また，特徴語 $t_m$ が URL $u_n$ に出現する回数を重み $w(t_m, u_n)$ とし，この重み $w(t_m, u_n)$ を要素とする特徴ベクトル $\vec{d_n} = (w(t_1, u_n), w(t_2, u_n), \cdots, w(t_M, u_n))$ を作成する．

**A Classification Method of Search Results based on URL**

Mariko YAMASHITA†,Yu SUZUKI‡and Kyoji KAWAGOE‡

†Graduate School of Science and Engineering, Ritsumeikan University

‡College of Information Science and Engineering, Ritsumeikan University

Nojihigashi 1-1-1, Kusatsu, Shiga, 525-8577 Japan

†yamashita@coms.ics.ritsumei.ac.jp

‡{yusuzuki, kawagoe}@is.ritsumei.ac.jp

### 2.3 URL 間の距離算出とクラスタリング

本研究では，クラスタリングを行う際に，階層型クラスタリングの手法である Ward 法 [3] を用いる．Ward 法では，データのばらつきの大きさを表す指標である平方和を用いて，クラスタ間の距離を算出する．

クラスタを $l(l=1,2,\cdots,N-1)$ 回統合したときのクラスタを $C_i^l(i=1,2,\cdots,N-l)$ とする．Web 文書の各 URL をそれぞれ一つずつ含む初期クラスタは $C_i^0(i=1,2,\cdots,N)$ となる．また，クラスタ $C_i^l$ に含まれる $K$ 件の URL を $u_k(C_i^l)(k=1,2,\cdots,K)$ とする．クラスタ $C_i^l$ に含まれる URL の特徴ベクトル $\vec{d}_n$ を用いて平均特徴ベクトル $\vec{v}_n=(\bar{w}(t_1,u_n),\bar{w}(t_2,u_n),\cdots,\bar{w}(t_M,u_n))$ を作成する．これらの値により，(1) 式を用いてクラスタ $C_i^l$ に含まれる特徴ベクトルの平方和 $E(C_i^l)$ を算出し，(2) 式を用いてクラスタ $C_a^l$，$C_b^l(a,b=1,2,\cdots,N-l;a\neq b)$ 間の距離 $S(C_a^l,C_b^l)$ を算出する．

$$E(C_i^l)=\sum_{k=1}^{K}\sum_{m=1}^{M}(w(t_m,u_k(C_i^l))-\bar{w}(t_m,C_i^l))^2 \quad (1)$$

$$S(C_a^l,C_b^l)=E(C_a^l\cup C_b^l)-E(C_a^l)-E(C_b^l) \quad (2)$$

クラスタ間の距離 $S(C_a^l,C_b^l)$ が近いほどクラスタ同士の類似度が高いとし，クラスタ間の距離が最小のクラスタ同士を統合する．そして，統合されたクラスタと他のクラスタとの距離を算出し直す．この処理を繰り返し，クラスタ数がある一定個数以下になった場合にクラスタリングを終了する．

## 3 評価実験

### 3.1 実験方法

本実験では正解集合を人手で作成するため，分類する際の指標をわかりやすくするために，検索語として一つの単語で複数の異なる対象に関して記述されている Web 文書が検索結果となる単語 9 個を検索語として用いる．例えば，『マック』で検索した場合，Mac やマクドナルドなどの Web 文書が検索結果に含まれる．これらの文書は，計算機と食べ物という全く異なる対象に関して記述されているため，異なる分類となる．何を対象にして記述されているかを人手で判断し，正解集合を作成した．

表 1: 実験結果

| | 処理時間（ミリ秒） | 編集回数（回） |
|---|---|---|
| 提案手法 | 1059.11 | 49.44 |
| 既存手法 | 43402.22 | 60.44 |

また，提案手法を実装したシステムと，Web 文書の内容から算出した類似度にもとづいてクラスタリングを行うシステムを作成する．Yahoo! JAPAN の検索結果の中で文書内容を取得することができた Web 文書上位 100 件に対して，これらのシステムを用いて正解集合のクラスタ数と同じクラスタ数になるようにクラスタリングを行う．そして，平均の処理時間と分類精度の点から二つのシステムを比較する．処理時間とは，利用者が検索語を入力してから検索結果を分類するまでにかかった時間とする．分類精度は，あらかじめ作成した正解集合と分類結果との比較を行い，正解集合と同じクラスタを作成するまでに編集した回数とする．

### 3.2 実験結果と評価

実験結果を表 1 に示す．この結果から，既存手法に比べて提案手法の方が大幅に処理時間が短いことがわかる．また，編集回数も既存手法よりも少ない回数となった．例えば，『ドライバー』という検索語で検索した場合を示す．ソフトウェアのドライバに関する Web 文書の URL には download という単語が多く含まれていた．また，自動車の運転に関する Web 文書の URL には paper などの単語が含まれていた．このように，URL には Web 文書の内容を表す文字列が含まれているため，文書中にノイズも多く含まれる既存手法に比べて分類精度が向上したと考えられる．

## 4 おわりに

本研究では，トライグラムにより URL から抽出した文字列を用いた検索結果の分類手法の提案と，提案手法を用いて実装したシステムの評価結果について述べた．今後の課題として，本提案手法の試作システムから得た Web 文書の分類結果を用いた，利用者とシステムとの対話的なクラスタ編集に適応することを考えている．

## 参考文献


[1] 丸山謙志, 王冠超, 徳山豪：“Web 検索結果におけるクラスタリングアルゴリズムの研究”, 情報処理学会研究報告, **2005-AL-100(3)**, pp. 17–24 (2005).

[2] 徳永健伸：“言語と計算 5 情報検索と言語処理”, 東京大学出版会 (1999).

[3] 新納浩幸：“R で学ぶクラスタ解析”, オーム社 (2007).