情報処理学会論文誌 Vol. 50 No. 4 1399–1409 (Apr. 2009)

キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰

馬場康夫[†1,*1] 新里圭司[†1]
柴田知秀[†1] 黒橋禎夫[†1]

本論文では，数千件のウェブ検索結果を対象としてクラスタリングを行い，その際に抽出されるラベルをキーワード蒸留によって精錬する手法を提案する．検索エンジン基盤 TSUBAKI と連携し，事前に行った言語処理結果を利用し，さらに並列計算機環境を利用することで，千件の検索結果の処理を約 1 分で行うことができる．また，表記揺れ，音訳，同義関係，包含関係，多義性解消などを，さまざまな言語リソースや検索結果中の文書頻度を利用することによって処理するキーワード蒸留という手法により，クエリに関連する良質なラベルを抽出することを可能とする．評価実験を行ったところ，提案システムは表記揺れの関係にあるラベルを正しく集約でき，また，検索結果の下位に埋もれた話題を発見するうえで有効であることが分かった．

Web Information Observation Using Keyword Distillation Based Clustering

Yasuo Bamba,[†1,*1] Keiji Shinzato,[†1]
Tomohide Shibata[†1] and Sadao Kurohashi[†1]

In this paper, we propose a system that conducts search result clustering for several thousands of web pages, and elaborates cluster labels extracted from the web pages through keyword distillation. The proposed system provides a user a clustering result from 1,000 web pages in approximately one minute by using a search engine infrastructure and grid computing environment. Keyword distillation is a method that properly handles spelling variations, transliterations, synonymous relations, inclusion relations and word ambiguity using linguistic resources and document frequencies in the search result. Experimental results show that the system correctly merges synonymous keywords and is useful for finding topics hidden in the lower-ranked web pages of a search result.

1. はじめに

現在，ウェブには膨大な量の情報が蓄積されており，その利活用が人々の生活基盤となりつつある．ウェブ情報を対象とする検索は誘導型と調査型に大別することができる．誘導型検索とは，たとえば「トヨタ自動車」や「福田康夫」のように会社や人物に関するトップページの検索であり，その場合には，ページリストをランク付けして返す既存のリスト型検索エンジンが有効に機能する．一方，たとえば「ゆとり教育」や「捕鯨問題」のように，ある話題についての情報を広く調査することを目的とする調査型の検索の場合には，リスト型検索エンジンで満足な結果が得られることは少ない．多くの場合，まずその話題についての概要の把握，ウェブ情報の全体像の把握が求められるが，ランク付けされたページリストからそれを知ることは難しい．

このような問題の解決策として，クラスタリングなどによって検索結果を集約する方法が考えられる．1 つのアプローチは，通常の文書クラスタリングと同じく文書ベクトルに基づく方法である[1)–4)]．また，クラスタリングにおける視点の多様性に注目し，ファセットの半自動獲得に着目したアプローチも存在する[5),6)]．

しかし，検索結果集約では，クラスタに対してより良いラベルを与えることが重要であることから，まずラベルを抽出し，それを含む文書をクラスタとするラベルベースの手法が主流となっている[7)–9)]．また，Clusty[⋆1]，Grokker[⋆2]，Mooter[⋆3]などの商用システムも存在している．

これらの研究・システムでは，まず，いくつかのリスト型検索エンジンから検索結果を収集し，そのスニペット（各ページの 2，3 文の要約）の集合から適切なキーワード（ラベル）を選択し，各ラベルを含むスニペットをまとめてクラスタとする．ラベルは，クエリ（話題）に対する重要関連語となっているはずであり，ユーザはラベル集合によって話題の概要を知ることができ，またラベルを通してページへアクセスすることができる．

しかし，このような既存のクラスタリングシステムには次のような問題点がある．

†1 京都大学
Kyoto University
∗1 現在，キヤノン株式会社
Presently with Canon Inc.
⋆1 http://clusty.com/, http://clusty.jp/（日本版）
⋆2 http://www.grokker.com/
⋆3 http://www.mooter.com/, http://www.mooter.co.jp/（日本版）

図 1　Clusty での「ゆとり教育」の検索結果
Fig. 1　An example of clustering search result in “Clusty” (query: “relaxed education”).

- 既存のリスト型検索エンジンをベースとしているためクラスタリングの対象となっているページ数が数百件程度に制限される．また，元のページを取得して利用することは行わず，検索エンジンが返すスニペット集合（数文）だけをラベル抽出の対象としている．このため，話題に対して十分な関連語を抽出することができていない．先にあげた既存システム Clusty でのクエリ「ゆとり教育」のクラスタリング結果を**図 1** に示す．この例では 204 件のウェブ検索結果のスニペットがクラスタリング対象となっており，左側のラベル一覧に，「ゆとり教育」の重要な概念である「詰め込み教育」「生きる力」といったラベルが含まれていない．
- ラベル抽出の対象が大きくないため（数百ページのスニペット）顕在化していないが，ラベル抽出において同義表現の処理などはほとんど行われていない．図 1 の例では「中央教育審議会」と「転換，中教審」はマージされるべきである．

本論文では，このような問題を解決し，大規模ウェブページ情報の俯瞰システムを提案する．**図 2** にクエリ「ゆとり教育」でのシステムの実行画面を示す．既存システムでは数百

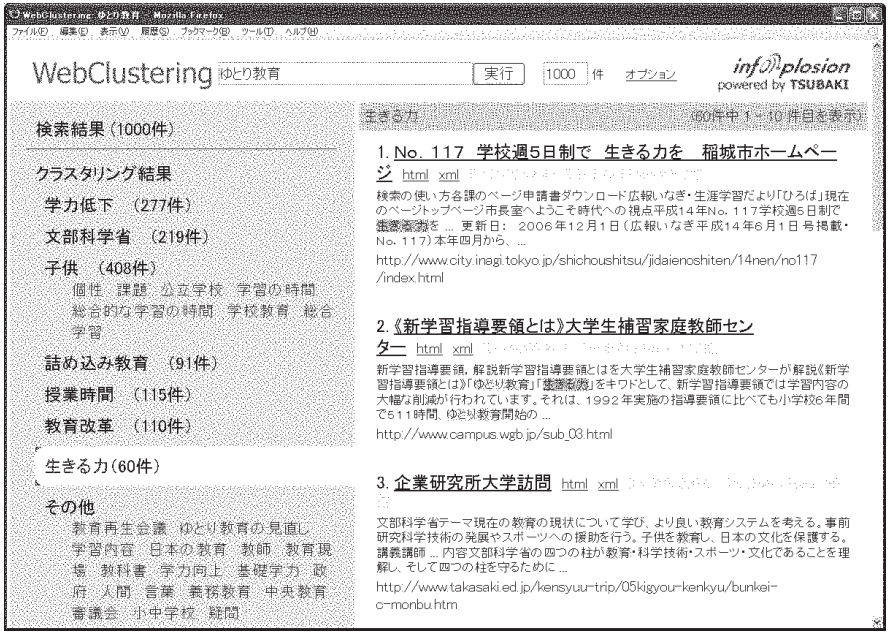

図 2　クラスタリングシステムの実行例（クエリ：ゆとり教育，件数：1,000 件）
Fig. 2　An example of clustering result (query: “relaxed education”, 1,000 pages).

件のウェブページの数百文がラベル抽出の対象であったのに対して，本システムでは検索エンジン基盤と連携し，さらに並列計算環境を用いることにより，数千件のウェブページの数万文を対象とするラベル抽出を可能とした．さらに，表記揺れや同義表現，部分全体関係のキーワードを徐々に集約していくキーワード蒸留により，話題に対する重要関連語を高純度で抽出することを可能とした．

2. システムの概要

2.1　開放型検索エンジン基盤 TSUBAKI

本システムは，その基盤として開放型検索エンジン基盤 TSUBAKI を利用する[10]．TSUBAKI は，次世代サーチ研究のための基盤を提供することを目的に構築された検索エンジンであり，2007 年 5 月〜7 月にクローリングした日本語ウェブページ約 1 億件を保

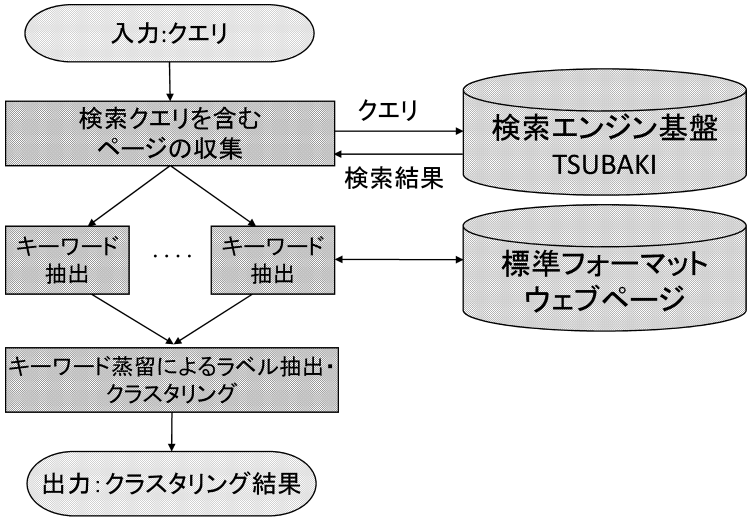

図 3　クラスタリングシステムの構成
Fig. 3　An overview of the proposed clustering system.

持している．

TSUBAKI の特徴の 1 つは，ウェブページに対する高度言語処理用の標準フォーマット（XML 形式）を定義し，その形にウェブページを変換して保存していることである．そこでは，HTML ファイルから文区切りを検出して，文単位で管理し，さらに文の形態素解析，構文解析結果も管理している．

商用サーチエンジンでは，ランキング尺度が公開されておらず，取得文書数・API 利用回数に制限があり，検索結果は随時更新されるため再現性がない．これに対して，TSUBAKI では，ランキング尺度を公開し，また API などによって無制限に検索結果，オリジナルウェブページ，標準フォーマットページを取得することができる．これによって，再現性のある形で次世代サーチの研究を支援している．

2.2　システム構成

本システムの構成を**図 3** に示す．本システムでは，まず TSUBAKI を用いて文書 ID を取得し，次に複数のキーワードサーバで各文書から並列にキーワードを取得し，これをマスタサーバで集約して結果を表示する．このような構成とすることにより，与えられたクエリに対して千件の文書から重要関連語を抽出し，クラスタリングすることが 1 分程度で可能となっている．以下，各処理の概要を示す．

2.2.1　クエリに対するページ集合の取得

クエリ Q およびクラスタリングしたいページ数 N が与えられると，TSUBAKI を用いてトップ N 件の検索結果（文書 ID 集合 D_Q）を得る．なお，TSUBAKI では OKAPI [11] を用いてページをランキングしている．

2.2.2　各ページからのキーワードの抽出

次に，検索結果の各ページからキーワード候補を抽出する．この処理は 26 並列のキーワードサーバで，詳細な言語解析結果を保持した標準フォーマットページを用いて行うことにより高速化されている．

ウェブページは複数の話題について述べていたり，広告などを含むことも多く，ページ全体を対象とするとクエリとは無関係なキーワードが多く抽出されてしまう．そこで，各ページからクエリと関連の強い重要文を抽出し，重要文のみからキーワードの抽出を行う．

重要文の抽出は独自に考案した以下のスコア計算をもとに行う．ここでのスコアは文の重要度を表す．まずクエリ Q に対する文 s の単独のスコアを以下の式により計算する．

$$w(s,Q) = l_Q \cdot m_Q \cdot \log(length(s)) \tag{1}$$

ここで，l_Q は文 s に含まれるクエリ Q 中の内容語の異なり数，m_Q は文 s に含まれる Q の内容語の出現回数の総和，$length(s)$ は文 s に含まれる単語数とする．クエリ Q 中の内容語を多く含みかつ，長い文ほど高いスコアを得ることになる．

さらに，文脈としてクエリと強く関連するものを選ぶために，前後それぞれ 2 文のスコアを伝播させ，次の値を文 s_0 のスコアとする．

$$score(s_0,Q) = \sum_{i=-2}^{2} \frac{w(s_i,Q)}{2^{|i|}} \tag{2}$$

ここで s の添字は文の相対位置を示すものとする．このようにして，各ページからスコアの上位 15 文を重要文として抽出する．

次に各重要文から，キーワードとして，名詞句（修飾節のないもの）およびかぎ括弧（「」，『』）で囲まれた表現を抽出する．

名詞句からはその部分単語列もキーワードとして抽出する．たとえば，「教育基本法」からは「教育」「基本」「法」「教育基本」「基本法」なども抽出する．この中から不適切な「教育基本」などを排除することは後述のキーワード蒸留で行う．

括弧表現は「分数ができない大学生」のような修飾節を含む重要なキーワードを取り出す

ために抽出する．括弧表現に限定せずに修飾節を含む名詞句をすべて抽出すると膨大な数となってしまうが，経験的には重要なキーワードはいずれかの文書で括弧表現として出現していることが多い．なお，各ページからの括弧表現抽出の段階では括弧のない表現が他の文書で括弧表現として出現しているかどうかは分からないので，すべてのページからの括弧表現抽出後にもう 1 度，重要文集合全体を調べ，括弧のない出現があればそのページからも括弧表現が抽出されたと見なす．

キーワード抽出では，「サイトマップ」や「プライバシーポリシー」のような一般的なキーワードは対象外とする．ここでは，TSUBAKI が持つ日本語ウェブページ 1 億件での文書頻度の上位 500 語（名詞句）を一般的なキーワードとしている．

2.2.3　キーワード蒸留

キーワードサーバで並列に抽出されたキーワードをマスタサーバに戻し，キーワードの集約を行う．この過程をキーワード蒸留と呼ぶことにする．キーワード蒸留については次節で詳しく説明する．この結果，最終的に約 30 個のラベルを得る．

2.2.4　結果の表示

先にあげた図 2 のシステムの実行画面で説明する．まず，キーワード蒸留の結果得られたラベルを画面左側に表示している．この一覧がクエリ（調べたい話題）の関連項目の集約となっており，これによって検索結果を鳥瞰図的に把握することができる．

そして，各ラベルは，そのラベルを含む文書集合のクラスタに対応付けられており，ラベルを選択（クリック）すると，画面右側にそのラベルを含むページのリスト（タイトルとスニペッツ）が表示される．このときのページ順位は TSUBAKI のランキングを基本とするが，ラベルがタイトルに含まれるページや，本文中に多数出現するページは順位を上げて表示する．

3.　キーワード蒸留

従来のラベルベースクラスタリング研究では，キーワード抽出の対象が数百文程度と少なかったために，単純に頻度上位のいくつかをラベルとして表示するということが行われていた．しかし，本システムのように数万文を対象として網羅的にキーワード抽出を行った場合には，それでは閲覧性の高い情報提示とはならない．それは，「つめこみ教育」「詰込み教育」「詰め込み型教育」，「カリキュラム」「教育課程」，「IWC」「IWC 総会」「IWC 総会開催」など，同じ内容を表す表現や包含関係にある表現がキーワード中に多数存在し，それらが別々に表示されると，人間の全体把握のプロセスを大きく阻害するからである．

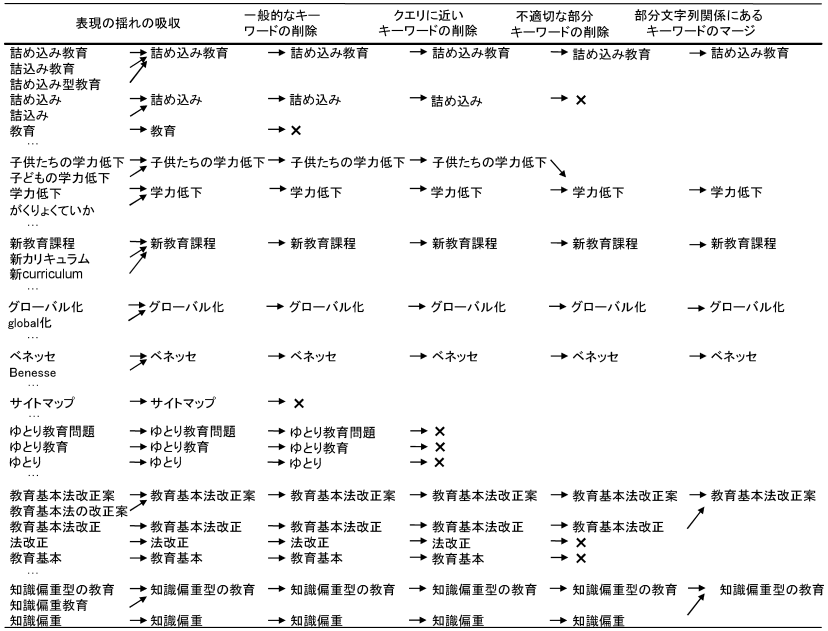

図 4　キーワードが蒸留される様子（クエリ：ゆとり教育）
Fig. 4　A process of keyword distillation (query: “relaxed education”).

そこで本システムでは，このような関係にあるキーワードを段階的に集約していくキーワード蒸留という処理を行う．例として，**図 4** にクエリ「ゆとり教育」に対するキーワード蒸留の過程を示す．

キーワード蒸留は，各キーワードサーバで抽出されたキーワードを，マスタサーバに集めて行う．各キーワードサーバは，キーワード一覧とそれぞれの文書頻度（自分の受け持つ検索文書中での文書頻度）をマスタサーバに返す．マスタサーバではまずこれを集計し，各キーワード l について検索文書全体 D_Q での文書頻度 $ldf(l)$ を得る．さらに，TSUBAKI が検索対象とするページ集合全体（1 億件）における l の文書頻度 $gdf(l)$ を得て，これをもとに l のスコアを次のように計算する．

$$score_{rel}(l) = ldf(l) \cdot \log \frac{N_T}{gdf(l)} \tag{3}$$

ここで，N_T は TSUBAKI が検索対象とするページ数（1 億）を表す．これは，検索結果中

の多くのページに現れ，かつ一般的ではないものほど高い値を持つスコアとなっている．

キーワード蒸留は，上記のスコアの上位 1 万個を対象としてスタートする．蒸留の過程でキーワードがマージされると，*ldf* および *gdf* の値が更新され，新たなスコアが与えられ，段階的にスコア上位のキーワードだけを残して蒸留が進められる．以下，キーワード蒸留の詳細を説明する．

3.1　同義異表記のキーワードのマージ

キーワード蒸留の最大の目的は，同じ意味を表し，表記の異なるキーワードをマージすることである．日本語文書におけるこのような現象を調査したところ，以下の 4 種類に分類できることが分かった．

A：日本語の字種，送り仮名などの表記揺れ
- 「詰め込み 教育」と「詰込み 教育」
- 「癌」と「ガン」と「がん」

B：音訳
- 「グローバル 化」と「global 化」
- 「ベネッセ」と「Benesse」

C：同義表現（略称などを含む）
- 「カリキュラム」と「教育課程」
- 「中央教育審議会」と「中教審」
- 「経済協力開発機構」と「OECD」

D：軽微な形態素の有無
- 「学力低下」と「学力 の 低下」
- 「小中学生」と「小 ・ 中学生」
- 「詰め込み教育」と「詰め込み 型 教育」

以下に上記 4 種類の問題への対処方法を述べる．

3.1.1　日本語の字種，送り仮名などの表記揺れ

日本語には平仮名，片仮名，漢字の 3 種類の文字種があり，これらの組合せで多数の表記がありうる．ある特定の表記が一般的（dominant）である語もあるが，「詰め込み」「詰込み」「つめこみ」，「癌」「がん」「ガン」など，よく用いられる複数の表記が存在する語も多い．

このような表記揺れについては，形態素解析システム JUMAN [⋆1]が基本語彙 3 万語について，代表表記（共通の ID）を与えているので，キーワードを代表表記に変換することでマージを行う．このとき，代表表記は必ずしも最も一般的な表記ではないので（そもそも，一般的な表記はドメインごとで変化することも多い），もとの表記で *ldf* の最も高い表記を表示のための表記として記憶しておく．これは，他のマージ処理においても共通に行う．

このような処理は，もとの表記（語）に曖昧性がある場合にはそのまま適用することはできない．たとえば，「がん」という表記は曖昧性があり，「癌」「雁」「岩」の 3 つの代表表記候補が得られる．このような場合，クエリに関連するコンテキスト（文書集合）においては，曖昧性は 1 つのものに解消され，それは文書集合において最も頻度の高い曖昧性のない表記であると仮定してマージを行う．すなわち，たとえば「成人病」というクエリの場合には，「がん」という曖昧なキーワードに対して，検索文書集合中に「癌」は頻出しているが，「雁」や「岩」はほとんど出現しておらず，このことから「がん」は代表表記「癌」に変換され，「癌」とマージされる[⋆2]．

3.1.2　音　　訳

日本語では，外来語を音訳して片仮名表記で用いることが多い．たとえば「global」は「世界的」のような日本語訳よりも音訳の「グローバル」の方が一般的に用いられる．さらに，このような語は，日本語文章の中で英単語そのままで用いられることも少なくない．このようなことから，「グローバル 化」と「global 化」のような音訳の関係にあるキーワードが抽出されることがある．

このうちの一部については，英和辞書において，「global」の訳として「グローバル」が与えられているような場合があり，これをもとにキーワードのマージを行うことができる．このとき，曖昧性の問題が発生することもあるが，前項の日本語の表記揺れの場合と同様に，クエリによって与えられるコンテキストを利用して解消する．

一方，音訳の関係は英和辞書だけを用いて高いカバレッジで扱うことはできない．たとえば，「Benesse」と「ベネッセ」は辞書にはのっていない．そこで音訳関係の動的な検出を行う．これは，片仮名のキーワードをいったん英語アルファベットに音訳し，それと英語キーワードとの編集距離を求め，ある閾値以下のものは同義異表記関係であると判断し，マージ

⋆1 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

⋆2 より本格的な解決策として，各出現について意味的曖昧性解消を行い代表表記を決定することが考えられる．しかし，あらゆる語について曖昧性解消の学習器を作り，それを大規模文書に適応することは現時点では難しい．一方，日本語の単語の場合，漢字表記の場合にはほとんど意味的曖昧性はなく，問題となるのはこの例のように平仮名の場合の曖昧性である．その場合には，この方法のように，他の文書における漢字表記の出現が有効利用できる．

する．

3.1.3 同 義 表 現

「教育課程」と「カリキュラム」のような同義表現については，大規模ウェブページから括弧表現に着目して収集した同義表現辞書を用いて対処する[12]．これはウェブページ 1 億件から，「... 教育課程（カリキュラム）...」と「... カリキュラム（教育課程）...」のように，括弧による注釈関係が対称に存在する場合に同義関係を抽出するもので，5,225 組の同義関係が 98%の精度で抽出されている．

これを用いて同義表現であるキーワードをマージする．曖昧性がある場合の解消はこれまでの処理と同様に行う．

3.1.4 軽微な形態素の有無

名詞間の修飾関係を示す助詞「の」，並列関係を示す中点「・」，さらに「詰め込み（型）教育」「子供（たち）の体力」などのように意味をほとんど変化させない接辞の有無によって，同義異表記のキーワードが抽出されていることがある．

これらは特定の形態素であるので辞書的に列挙し，それらを削除したものをキーワードの標準形と考え，標準形が一致するものはマージを行う．この際，3.1.1 項で述べたように，その中で頻出する表現を記憶しておくことにより，最終的にラベルとして表示する際には自然な表記が用いられるようにする．

3.2 不適切な部分キーワードの削除

2.2.2 項で述べたように，最初のキーワード抽出の段階では，「教育基本法」中の「教育」や「教育基本」のように，名詞句の中の部分単語列もキーワードとして抽出する．これは，部分単語列としての出現も考慮することで，そのキーワードとしての適切さをより正確に測るための処理である．

しかし，こうして抽出された部分単語列の中には，名詞句内の語の結び付きを考慮していないために，「教育基本」のような不適切なキーワードも含まれている．そこで，キーワード蒸留の段階でそのような不適切なものを排除する．

ここでは，適切な区切りのキーワードは単独で（最大の名詞句として）もある程度の頻度で出現していると仮定し，単独での出現回数が部分単語列としての出現回数よりも少ないものは不適切なものとして排除する．

3.3 包含関係にあるキーワードのマージ

最後に，「知識偏重」と「知識偏重型教育」のような包含関係にあるキーワードについて，マージを行うかどうかを判断する．

本手法では，部分単語列についても頻度を数えているので，「知識偏重」の文書頻度が「知識偏重教育」の文書頻度よりも必ず大きい．そのうえで，たとえばこの 2 つのキーワードの頻度に大きな差がなければ，与えられたクエリに関連するドメインでは，「知識偏重」はほぼ「知識偏重教育」に関連して述べられていると考えることができ，これらはマージすることが適当である．また，その際には情報量の多い「知識偏重教育」でこれらを代表させる．

一方，「基礎学力」「基礎学力低下」「基礎学力調査」などのキーワードがあり，「基礎学力」の頻度が相対的に大きい場合には，「基礎学力」が重要な概念であり，これを残しておくことが適切であると考えられる．

そこで，包含関係にある短いキーワードと長いキーワードの頻度をみて，前者の頻度が後者の頻度の 2 倍以下であれば長いキーワードにマージし，そうでなければマージを行わないこととした．なお，この 2 倍というパラメータは経験的に定めた．

ここまでの処理で残ったキーワードについて，スコアの上位 30 個のキーワードを最終的に採用する．最終的に採用されたキーワードは**ラベル**と呼ぶことにする．

3.4 ラベル間の関連性の解析

クエリに関連する 30 個のラベルを提示する際，その中で関連するものはまとめて表示することが望ましい．各ラベルは，それが出現するページ集合に対応付けられているので，ページ集合の重なりが大きいラベルは関連するラベルであると考えることにする．

各ラベルのページ集合の重なりは Simpson 係数によって計算する．2 つのラベル l_1，l_2 について，それぞれに含まれるページ集合を P_1，P_2 とするとき，Simpson 係数は次の式で求められる．

$$\mathrm{Simpson}(P_1, P_2) = \frac{|P_1 \cap P_2|}{\min(|P_1|, |P_2|)} \tag{4}$$

この値が閾値（0.7）を上回る場合，*ldf* が大きい方のラベルを親，小さな方を子と定義する．1 つのラベルに対して親になりうるラベルが複数存在する場合は，その中から最も Simpson 係数の高いラベルを親とする．

このように抽出した親子関係を集約し，親を持たないラベルを根ラベルとし，根ラベルとその子孫ラベルを関連するラベルとしてまとめて扱うこととする．

本システムの解析結果画面（図 2）では，左側画面に，スコア順に 7 つの根ラベルとその子孫ラベルを表示する．この 7 つの集合に含まれない他のラベルについては，まとめて「その他」とする．図 2 の例では「学力低下」「文部科学省」「詰め込み教育」などが根ラベルである．

4. 評価実験

本システムの有効性を調べるために，まずシステムの処理スピードを測定し，次に，ラベルの評価を行った．全体を通じて，評価は以下の 15 個の調査型のクエリを対象に行った．

BSE 問題，携帯電話の電磁波，年金制度，NHK の受信料，子供の体力低下，パレスチナ問題，アンチエイジング，少子化問題，捕鯨問題，癌の予防，ダイエット食品，マイナスイオン，京都観光，地球温暖化，ゆとり教育

本クラスタリングシステムは 1 台のマスタサーバと 26 台のキーワード抽出サーバの上に構築した．マスタサーバ，キーワード抽出サーバともにスペックは CPU 3.60 GHz（2 コア），メモリ 6 GB である．

4.1 実行時間

15 個のクエリを対象に，1,000 件の検索結果の取得およびクラスタリングの実行に要する時間を計測した．クエリごとの時間ならびにその平均値を**表 1** に示す．キーワード抽出の並列化および言語処理を事前に行い標準フォーマットとしておくことにより，検索結果 1,000 件に対する処理を平均約 1 分で行うことができている．また，それほど顕著な傾向ではないが，クエリに含まれる自立語数が多いほど実行時間が長くなっていることが分かる．

表 1　1,000 件のクラスタリングに要する時間の平均値（秒）
Table 1　Average time required to clustering 1,000 pages.

クエリ	検索結果の取得	キーワード抽出	キーワード蒸留	合計
BSE 問題	17	33	16	66
携帯電話の電磁波	24	38	17	79
年金制度	13	28	16	57
NHK の受信料	15	43	13	71
子供の体力低下	11	33	17	61
パレスチナ問題	11	35	17	63
アンチエイジング	7	28	12	47
少子化問題	10	37	17	64
捕鯨問題	9	45	22	76
癌の予防	14	28	16	58
ダイエット食品	14	26	11	51
マイナスイオン	11	29	12	52
京都観光	21	22	10	53
地球温暖化	11	40	14	65
ゆとり教育	8	41	20	69
平均実行時間	13.1	33.7	15.3	62.1

4.2 ラベルの評価

システムが出力するラベルをいくつかの観点から評価した．ラベルを評価する方法は Zeng ら[8] や Ferragina ら[9] の研究においても採用されており，そこでは，システムが出力した上位のラベルについて人手でクエリと関連があるかどうかを評価している．

本研究では，まず，キーワード蒸留においてマージが適切に行われているかどうかを調べ，次に，クエリに対してラベルが検索結果を集約するうえで適切であるかどうかを総合的に判断した．さらに，リスト型検索エンジンとの比較によって，本システムの話題検出の有効性を調査した．

4.2.1 ラベルのマージの適切さ

キーワード蒸留においてマージが適切に行われているか，あるいは過不足があるか，すなわち，マージされるべきでないキーワードとマージされていたり，マージされるべきキーワードが残っていたりするかどうかを調査した．具体的には，15 個のクエリそれぞれについて，キーワード蒸留の結果得られた約 30 個のラベル，計 450 個について適切かどうかを評価した．その結果，96.4%のラベルは適切にマージされたものであった．また，クエリ別の精度を**表 2** の左側に示す．ほぼ全クエリにわたって同じような傾向がみられた．

マージが不十分であったラベルの例としては，「マイナスイオン発生 器」と「マイナスイオン発生 機」（マイナスイオン），「親」と「保護者」（子供の体力低下），「電化製品」と「家電製品」（携帯電話の電磁波），などがあった．これらは一般的な同義表現というわけではないので，与えられたクエリの文脈における類似性によって，柔軟にまとめることを判断する必要である．

4.2.2 ラベルの検索結果集約としての適切さ

ラベルおよびそれに対応するクラスタが，クエリの検索結果を集約するうえで適切であるかどうかを評価した．ここでは，15 個のクエリについて，上位 7 個（7 個存在しない場合はすべて）の根ラベル，計 103 個のラベルを評価対象とした．それぞれのラベルについて，上位 10 ページを見て総合的に判定した．一見ラベルがクエリと関連があるようにみえても，実際に集まっているページ集合とラベルが深く関わるものでなかったり，クラスタが多くのページを含む漠然としたものであれば適切でないとした．結果は，適切なものが 84.4%，ほぼ適切であるものが 12.6%，適切でないものは 2.9%であり，本手法のラベル抽出は総合的に有効なものであった．また，クエリ別の精度を表 2 の右側に示す．こちらの評価に関してもほぼ全クエリにわたって同じような傾向がみられた．

表 2　クエリ別のラベルのマージの適切さと検索結果集約としての適切さ
Table 2　The appropriateness of mergeing labels and organizing the search result in each query.

クエリ	ラベルのマージ	検索結果集約		
		適切	ほぼ適切	適切でない
BSE 問題	30 / 30	5 / 7	2 / 7	0 / 7
携帯電話の電磁波	29 / 30	7 / 7	0 / 7	0 / 7
年金制度	30 / 30	5 / 7	2 / 7	0 / 7
NHK の受信料	30 / 30	6 / 7	1 / 7	0 / 7
子供の体力低下	29 / 30	7 / 7	0 / 7	0 / 7
パレスチナ問題	30 / 30	5 / 7	2 / 7	0 / 7
アンチエイジング	28 / 30	5 / 7	2 / 7	0 / 7
少子化問題	30 / 30	6 / 7	1 / 7	0 / 7
捕鯨問題	28 / 30	3 / 5	0 / 5	2 / 5
癌の予防	29 / 30	7 / 7	0 / 7	0 / 7
ダイエット食品	29 / 30	7 / 7	0 / 7	0 / 7
マイナスイオン	29 / 30	7 / 7	0 / 7	0 / 7
京都観光	26 / 30	6 / 7	0 / 7	1 / 7
地球温暖化	28 / 30	4 / 7	3 / 7	0 / 7
ゆとり教育	29 / 30	7 / 7	0 / 7	0 / 7
合計	434 / 450 (96.4%)	87 / 103 (84.4%)	13 / 103 (12.6%)	3 / 103 (2.9%)

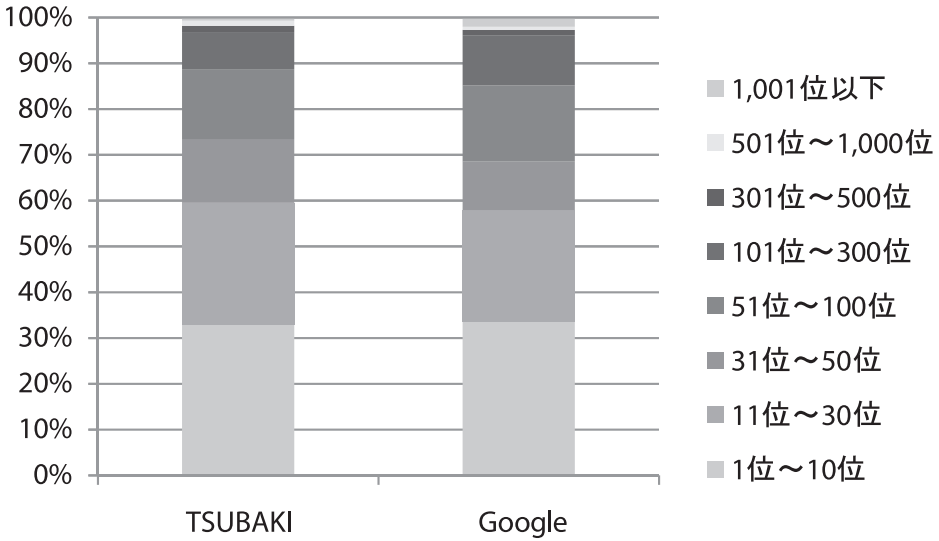

図 5　リスト型検索エンジンにおけるクラスタのラベルの初出の順位の分布
Fig. 5　Distribution of the first-appearance rank of labels in list-style search engines.

適切でないと判定された例として，「京都駅」（京都観光）がある．このクラスタでは，京都駅そのものの紹介をしているページは 1 ページにとどまり，残りのページはすべて，他の観光名所への経路を説明している部分に「京都駅」という語が出現するページであった．

また，クエリ「少子化問題」に対するラベル「日本の少子化」，クエリ「地球温暖化」に対するラベル「地球温暖化対策」「地球温暖化防止」などでは，1,000 件の検索結果中 4〜5 割のページを含むものであり，検索結果を集約するうえでは有効なラベルとは考えられないものであった．

4.2.3　リスト型検索エンジンとの比較

Jansen ら[13] が Excite [⋆1] の検索ログを用いて行った調査によれば，利用者が閲覧する検索結果中のページの数は平均 2.35 ページにすぎない．このように，従来のリスト型検索エンジンの利用者は，検索結果の上位しか閲覧しないことが知られており，検索の下位に埋もれた話題を見逃してしまう．

⋆1 http://www.excite.com/

たとえば，クエリ「ゆとり教育」について TSUBAKI で検索したとき，TSUBAKI が出力する結果表示画面（スニペットおよびタイトル）中に「生きる力」というキーワードが最初に現れるのは 32 位である．つまり TSUBAKI の利用者は，最低 32 個の検索結果を閲覧しないと「生きる力」という話題の存在に気づくことができない．これに対して，クラスタリングシステムを用いる場合には，「生きる力」はラベルとして抽出されており，ユーザはその存在に即座に気づくことができる．

このように，リスト型検索エンジンでは下位に埋もれて発見が難しい話題が，本システムを用いることでどの程度検出できるかを 15 クエリ，450 個のラベルで定量的に評価した．リスト型検索エンジンとして TSUBAKI と Google を用いた．この結果を**図 5** に示す．図より，TSUBAKI，Google ともに検索結果の 10 位以内に存在するラベルは 3 割程度であり，残り 7 割は 11 位以下の検索結果を見ないと出現しないものであることが分かる．さらに，TSUBAKI ではラベルの約 1 割，Google ではラベルの約 1.5 割が 100 位以下まで検索結果をたどらないと出現しないことも分かる．

本システムは処理に 1 分ほどかかるのに対して，リスト型検索エンジンでは瞬時に結果

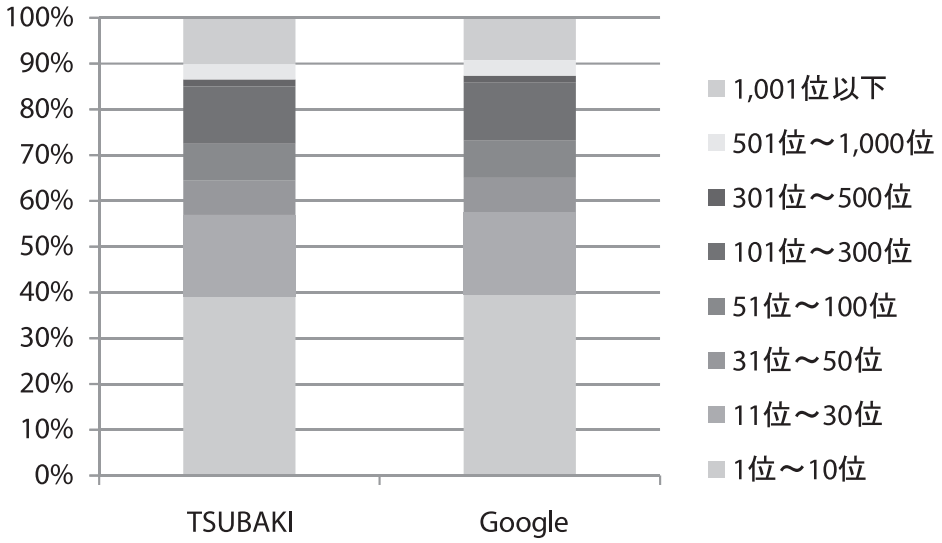

図 6　リスト型検索エンジンにおけるクラスタのラベルの初出の順位の分布（Clusty）
Fig. 6　Distribution of the first-appearance rank of labels in list-style search engines (Clusty).

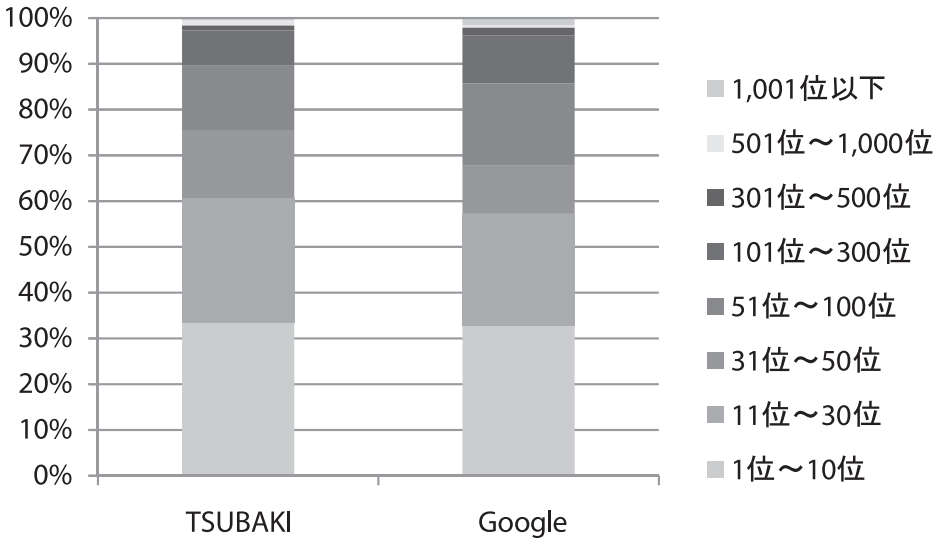

図 7　リスト型検索エンジンにおけるクラスタのラベルの初出の順位の分布（キーワード蒸留を行わない場合）
Fig. 7　Distribution of the first-appearance rank of labels in list-style search engines (w/o keyword distillation).

が得られて検索結果の閲覧が始められるという差があるが，それでも 100 件あるいはそれ以上の検索結果を調べることは 1 分では難しい．これらの結果から，リスト型検索エンジンと比較して，提案するシステムは検索結果の下位に埋もれた話題を発見するうえで有効であるといえる．

4.3　既存システムとの比較

既存システムの 1 つである Clusty と本システムの比較を行った．まず実行時間については Clusty では 2，3 秒で検索結果が返ってきており，高速に動作する．

次に，4.2.2 項で述べたラベルの検索結果集約としての適切さを評価した．Clusty についても，15 個のクエリにおいて，1 階層目のラベル，計 146 個のラベルに対して評価を行ったところ，適切なものが 58.2%，ほぼ適切であるものが 18.3%，適切でないものは 23.9%となり，本システムの方がかなり良い結果が得られていることが分かった．適切でないものの例としては，クエリを含むページに出現するがクエリと関係のないもの（クエリ「ダイエット食品」に対する「化粧品」，クエリ「ゆとり教育」に対する「入試」）や，ラベルとクラスタの内容と関連の薄いもの（クエリ「少子化問題」に対する「時代」，クエリ「マイナスイオン」に対する「専門」）などがあった．

続いて，4.2.3 項で述べたリスト型検索エンジンとの比較を行った．15 個のクエリに対して 2 階層目までのラベル，計 534 個のラベルの TSUBAKI，Google での初出の順位を調べた．結果を**図 6** に示す．図 5 と比べると，1,000 位までに見つからなかったラベルの割合が多いことが分かる．たとえば，クエリ「ゆとり教育」に対する「読売新聞社の速報」，クエリ「NHK の受信料」に対する「12 年度」などといったクエリと関連が薄いものがあり，これらは Clusty の対象文数が少ないために文書頻度が少しでも多いとラベルとして抽出されるためだと思われる．また，検索結果の 10 位以内に存在するラベルは 4 割程度となっており，本システムの 3 割程度に対して多くなっている．したがって，既存のシステム Clusty ではリスト型検索エンジンで比較的容易に見つかるような話題を本システムよりも多く含んでいることが分かる．

以上より，本システムは既存のシステムよりも動作に時間はかかるものの，ラベルの適切さ，リスト型検索エンジンで下位に埋もれる話題の発見において，既存のシステムを上回る

ことが分かった．

4.4 キーワード蒸留の効果

本研究で提案したキーワード蒸留がどの程度有効であるかを調べた．キーワード蒸留を行わないシステムに対して，まず，4.2.2 項で述べたラベルの検索結果集約としての適切さを評価した．4.2.2 項と同様の評価を行ったところ，適切なものが 69.5%，ほぼ適切であるものが 18.1%，適切でないものは 12.4%であり，キーワード蒸留を行わないと精度が低下することが分かる．低下した例としては，クエリと同義関係にあるラベルが出現してしまっているものが多く，クエリ「少子化問題」に対する「少子化の問題」，クエリ「BSE 問題」における「狂牛病」，「牛海綿状脳症」などがあった．

同様に，キーワード蒸留を行わないシステムに対して 4.2.3 項で述べたリスト型検索エンジンとの比較を行った結果を**図 7** に示す．図 5 と比べてほとんど差がないことが分かる．これは，ラベルの初出の順位が高いラベルがマージされないと検索結果の順位が高いものの割合が増えるが，ほぼ同様の割合で初出の順位の低いラベルがマージされなかったことを示している．

5. おわりに

本論文では，数千件のウェブ検索結果を対象としてクラスタリングを行い，その際に抽出されるラベルをキーワード蒸留によって精錬する手法を提案した．検索エンジン基盤 TSUBAKI と連携し，事前に行った言語処理結果を利用し，さらに並列計算機環境を利用することで，千件の検索結果の処理を約 1 分で行うことが可能となった．また，表記揺れ，音訳，同義関係，包含関係，多義性解消などを，さまざまな言語リソースや検索結果中の文書頻度を利用することによって処理するキーワード蒸留という手法により，クエリに関連する良質なラベルを抽出することを可能とした．

今後の最も重要な課題としては，抽出されたラベルに対するより高度な関連付け，組織化がある．現在は，ラベルを含むページの重複だけを手がかりとしているが，固有表現解析結果の利用や，シソーラス・オントロジなどの利用によって，より高度な組織化が可能であると考えている．

参考文献

1) Cutting, D.R., Pedersen, J.O., Karger, D. and Tukey, J.W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *Proc. 15th Annual Int'l ACM SIGIR Conference on R&D in IR*, pp.318–329 (1992).
2) Hearst, M.A. and Pedersen, J.O.: Reexamining the cluster hypothesis: Scatter/gather on retrieval results, *Proc. SIGIR-96, 19th ACM Int'l Conference on R&D in IR*, Zürich, CH, pp.76–84 (1996).
3) Osinski, S., Stefanowski, J. and Weiss, D.: Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition, *Intelligent Information Processing And Web Mining: Proc. International IIS: IIPWM'04 Conference*, Zakopane, Poland, May 17–20 (2004).
4) Geraci, F., Pellegrini, M., Pisati, P. and Sebastiani, F.: A scalable algorithm for high-quality clustering of web snippets, *Proc. 2006 ACM symposium on Applied computing*, pp.1058–1062 (2006).
5) Hearst, M.A.: Clustering versus faceted categories for information exploration, *Comm. ACM*, Vol.49, No.4, pp.59–61 (2006).
6) Stoica, E., Hearst, M. and Richardson, M.: Automating Creation of Hierarchical Faceted Metadata Structures, *Proc. NAACL HLT*, pp.244–251 (2007).
7) Zamir, O. and Etzioni, O.: Grouper: A Dynamic Clustering Interface to Web Search Results, *Proc. 8th International World Wide Web Conference*, Vol.31, No.11-16, pp.1361–1374 (1999).
8) Zeng, H., He, Q., Chen, Z., Ma, W. and Ma, J.: Learning to cluster web search results, *Proc. 27th Annual International Conference on Research and Development in Information Retrieval*, pp.210–217 (2004).
9) Ferragina, P. and Gulli, A.: A personalized search engine based on Web-snippet hierarchical clustering, *Software: Practice and Experience*, pp.189–225 (2007).
10) Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S.: TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology, *Proc. IJCNLP2008*, pp.189–196 (2008).
11) Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M.: Okapi at TREC-3, *The 3rd Text REtrieval Conference* (*TREC-3*) (1994).
12) Sasano, R., Kawahara, D. and Kurohashi, S.: Improving coreference resolution using bridging reference resolution and automatically acquired synonyms, *Proc. DAARC2007*, pp.125–136 (2007).
13) Jansen, B.J., Spink, A. and Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web, *Information Processing and Management*, Vol.36, No.2, pp.207–227 (2000).

(平成 20 年 8 月 25 日受付)
(平成 21 年 1 月 7 日採録)

馬場　康夫

2006 年京都大学工学部電気電子工学科卒業．2008 年京都大学大学院情報学研究科修士課程修了．現在，キヤノン株式会社勤務．

新里　圭司

2002 年東京電機大学工学部情報通信工学科卒業．2006 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（情報科学）．現在，京都大学大学院情報学研究科特定研究員．自然言語処理の研究に従事．

柴田　知秀

2002 年東京大学工学部電子情報工学科卒業．2007 年東京大学大学院情報理工学系研究科博士課程修了．博士（情報理工学）．現在，京都大学大学院情報学研究科助教．自然言語処理の研究に従事．

黒橋　禎夫（正会員）

1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了．博士（工学）．2006 年 4 月より京都大学大学院情報学研究科教授．自然言語処理，知識情報処理の研究に従事．