社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (10)
2008－ＮＬ－183 (10)
2008／1／22

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

馬場 康夫　新里 圭司　黒橋 禎夫
京都大学大学院情報学研究科知能情報学専攻
〒606-8501 京都市左京区吉田本町
{banba,shinzato,kuro}@nlp.kuee.kyoto-u.ac.jp

あらまし

本稿では，検索エンジン基盤TSUBAKIを使って検索されたウェブページを，ページ中の複合名詞に注目して自動的にクラスタリングするシステムについて述べる．検索エンジン基盤TSUBAKIとは，日本語ウェブページ1億件を対象にした研究用途に主眼をおいた検索エンジンである．本クラスタリングシステムは，このTSUBAKIと連係することで，数千から数万件のウェブページを分類することが可能であり，さらに，豊富な言語情報を利用した高精度な複合名詞抽出を行うことが可能である．簡単な評価実験の結果，本システムを用いることでTSUBAKIの検索結果中で下位に埋もれているウェブページに対し効率よくアクセスできること，さらには，抽出した複合名詞が有用な情報へアクセスする際に有効であることがわかった．

キーワード　大規模ウェブページクラスタリング，検索エンジン基盤TSUBAKI，複合名詞抽出，異表記の吸収

Development of a Large-scale Web Page Clustering System using an Open Search Engine Infrastructure TSUBAKI

Yasuo BAMBA　Keiji SHINZATO　Sadao KUROHASHI
Department of Intelligence Science and Technology,
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{banba,shinzato,kuro}@nlp.kuee.kyoto-u.ac.jp

Abstract

This paper describes a system that organizes a large number of web pages retrieved from the search engine TSUBAKI into clusters according to compound nouns extracted from the pages. TSUBAKI is a search engine infrastructure that can retrieve pages from 100 million web pages. Our clustering system deeply cooperates with TSUBAKI. This enables the system to generate clusters from several thousand web pages, and to give elaborate labels to the clusters. Experimental results showed that our system users can efficiently access low-ranked web pages in a search result obtained from TSUBAKI, and that generated labels navigate the users to information that they want.

Key Words　large-scale Web page clustering, search engine infrastructure TSUBAKI, compound noun extraction, expressive divergence assimilation

1 はじめに

現在，ウェブ上には膨大な量の情報が氾濫しており，その中から求める情報を得るためには検索エンジンの利用は不可欠である．検索クエリには，調査型，誘導型，取引型の3種類あることが知られており [1]，既存の検索エンジンを用いて，調査型の検索クエリに適合するウェブページを瞬時に発見することは難しい．調査型の検索クエリに適合するウェブページの発見を難しくしている理由として，既存の検索エンジンの抱える以下の問題点が考えられる．

- 検索結果として数万を超えるページが得られることも多く，結果の全体像を瞬時に俯瞰することが難しい
- クエリと関連するウェブページをランキングしたリストを検索結果として提示するだけであるため，利用者は自分の欲しい情報を結果の中から探し出さなければならない

上記の問題点を解決し，利用者を，自身の必要とする情報へ素早く導くための新しい検索エンジンの研究・開発は非常に重要である．

本稿では我々が研究・開発を進めているウェブクラスタリングシステムについて述べる．ウェブクラスタリングシステムとは，検索エンジンより得られる検索結果を，自動的に分類するシステムであり，上述した既存の検索エンジンの抱える問題点を解消することが可能である．図1はシステムの実行画面である．画面左にはクラスタリングによって生成されたクラスタのラベル一覧が，画面右には選択されたクラスタに属すページのタイトルやスニペッツがリスト形式で表示される．利用者は画面左に列挙されたラベル全体を眺めることで，検索結果の鳥瞰図的把握を得ることが可能である．さらにラベルを手がかりにすることで，自身の必要とする情報へ効率的にアクセスすることも可能である．

検索結果をクラスタリングするシステムは現在までに数多く提案されている [2–4]．また商用システムとしても，Clusty[1]をはじめ，SRC[2], Mooter[3]など多くのシステムが存在する．これら既存のシステムでは，利用者からクエリが与えられると，複数の検索エンジンから検索結果を収集し，そこに含まれる高々200件程度のウェブページを自動分類するに留まっている．さらに，即時性を重視するため，クラスタに付与されるラベルは，検索結果に含まれるスニペッツやページのタイトルだけから獲得されることが多い．

その一方で，我々のシステムでは検索エンジン基盤 TSUBAKI [5] を利用することにより，数千件から数万件規模の検索結果を分類することが可能である．これにより，既存の検索エンジンを用いた場合は検索結果の奥底に埋もれてしまうページであっても，検索結果中のページを多角的に分類することで，ある観点からは上位のページとして利用者へと提供できる可能性がある．さらに我々のシステムでは，TSUBAKI より提供されるページの解析済みデータを利用することで，豊富な言語情報を手がかりに可読性が高く，意味的に冗長性の少ないラベルの集合を獲得することも可能である．

本論文の構成は以下の通りである．まず，2節でシステムの全体像について述べ，3節で我々のシステムにおいて重要なラベル抽出処理について述べる．続いて4節でシステムの構成について述べ，5節でシステムの簡単な評価実験について述べる．

[1] http://clusty.jp/
[2] http://rwsm.directtaps.net/
[3] http://www.mooter.co.jp/

2 システムの概要

本稿で提案するクラスタリングシステムの概要を図2に示す．本節では，各ステップについて述べる．

Step 1: クラスタリング対象となる文書の取得

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，検索エンジン基盤 TSUBAKI [5] を使ってクエリ Q の検索結果を N 件取得する．一般に，検索結果中には商品販売ページや，ほぼ写真だけからなるページなど，クラスタリングの対象として好ましくないページが含まれている．そこで，本システムでは，商品販売ページ，写真ページ，リンク集ページの3種類のページタイプを設け，これらのタイプに該当するページを簡単な統計量およびヒューリスティック・ルールを使って自動的に判定し，検索結果から除く．以下，検索結果中から上記のページタイプに属すページを除いてできるページの集合を D_Q とする．

D_Q に含まれる各ページについて，標準フォーマット [6] に変換されたデータを取得する．標準フォーマット化されたデータには，あらかじめ抽出された日本語文や，日本語文を Juman [7]・KNP [8] で解析した結果が埋め込まれている．

Step 2: ラベルの抽出

Step 2 では D_Q から複合名詞を抽出し，ラベルとする．まず，入力されたクエリと関連の強い文を重要文として D_Q 中の各文書から抽出する．そして，抽出された重要文から品詞の並びや，かぎ括弧（「」）に注目し複合名詞を抽出する．その後，抽出

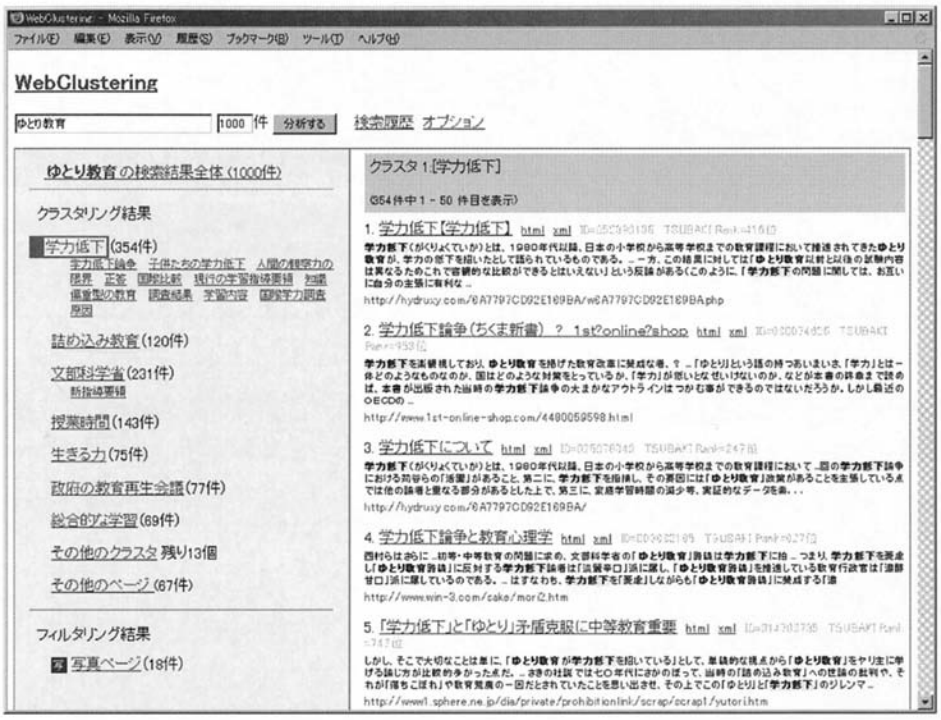

図 1: クラスタリングシステムの実行例（クエリ:ゆとり教育，件数:1,000 件）

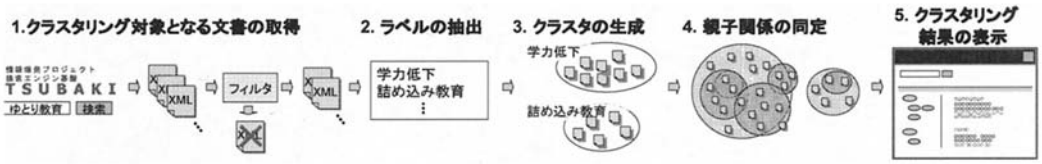

図 2: クラスタリングシステムの概要

された複合名詞の表現の揺れを吸収し，文書頻度などの統計量を使って最終的にラベルとする複合名詞の選別を行う.（詳細については 3 節で述べる）

Step 3: クラスタの生成

Step 3 では，Step 2 で抽出されたラベルをもとに，同一のラベルを含むページ同士をまとめることでクラスタを生成する．この時，複数のラベルを含むページは，各クラスタの要素とする．さらに，Step 2 で抽出されたどのラベルも含まないページについては,「その他のページ」という特別なクラスタに含める.

Step 4: 親子関係の同定

Step 4 では，各クラスタに含まれるページ集合の包含関係から親子関係を定義し，クラスタ間に階層構造を構築する．クラスタ間の包含関係は Simpson 係数により計算する．任意のクラスタ C_1, C_2 に含まれるページの集合を P_1, P_2 としたとき，Simpson 係数は

$$\mathrm{Simpson}(P_1, P_2) = \frac{|P_1 \cap P_2|}{\min(|P_1|, |P_2|)} \qquad (1)$$

で求められる．この値が閾値（0.7）を上回るクラスタ同士について，サイズの大きな方のクラスタを親，小さな方のクラスタを子と定義する．ただし，1 つのクラスタに対して親になりうるクラスタが複数存在する場合は，その中からもっとも Simpson 係数の高いクラスタを選び，それを親とする.

Step 5: クラスタリング結果の表示

Step 5では，図1のようなクラスタリング結果を表示する．画面左には，Step 4までで生成されたクラスタのラベルの一覧および，Step 1で自動判定されたページタイプに属すページの集合が表示される．また，画面右には選択されたクラスタに属するページのタイトルおよびスニペッツなどがリスト形式で表示される．

ラベルを表示する際，まず，3.6節で定義されるスコアの高い上位20件のラベルを取得する．そして，スコア上位7件のラベルについては，そのまま画面に表示し，残りの13件については，画面が煩雑にならないようにデフォルトでは表示しない．この13件のラベルについては，「その他のクラスタ」をクリックすることで画面に表示される．また，あるクラスタの子要素になっているクラスタのラベルは，親クラスの下に字下げして表示される．

その一方で，画面右に表示されるクラスタ中の各ページは，以下の条件を満たすページから順に表示する．これは，ラベルと関連が深いページが上位に表示された方が閲覧性が向上すると考えたためである．

- ページのタイトルにクラスタのラベルが含まれている
- ページの本文中におけるラベルの出現頻度が高い
- 検索エンジンTSUBAKIの検索スコアが高い

3 ラベルの抽出

本システムのラベル抽出処理は以下の手順からなる．

1. 標準フォーマットからの重要文抽出
2. 重要文中に含まれる複合名詞の抽出
3. 抽出された複合名詞の表現の揺れの吸収
4. 不適切な複合名詞の削除
5. 部分文字列関係にある複合名詞のマージ
6. クエリと関連性の高い複合名詞の選択

ラベルベース手法に基づくクラスタリングシステムにおいて，ラベル抽出処理は，クラスタリング結果の閲覧性を左右する重要なコンポーネントである．本システムでは，D_Q 中に出現する複合名詞をラベルとして抽出するが，単に抽出しただけでは，閲覧性の高いクラスタリング結果を生成することは難しい．それは，「詰め込み型教育」と「詰込み教育」などの表現の揺れの問題が存在するためである．ラベルとして抽出された複合名詞の表現が揺れていると，表現は異なるが意味的に同じ複合名詞がラベルとして表示されることとなり，クラスタリング結果の閲覧性低下の原因となる．そこで本システムでは，複合名詞の文書頻度や同義表現データベースなどの外部リソースを用いることで，ラベルとして抽出される複合名詞の表現の揺れの問題を解決し，意味的に冗長でないラベルの集合の抽出を目指す．図3に，クエリ「ゆとり教育」で抽出されるラベルの例を示す．図より，以下で述べる各処理を経ることで，ラベルが洗練されていく様子がわかる。

以下，ラベル抽出に伴う各処理について述べる．

3.1 重要文の抽出

ラベル抽出に先立ち，D_Q 中の各ページについて，クエリと関連の強い重要文を選択する．そのため，Step 1で取得された標準フォーマット化データに含まれている文とその解析結果を利用する．

標準フォーマットに埋め込まれている文 s について，クエリ Q に対する重要度を以下の式により計算する．

$$w(s,Q) = l_Q \cdot m_Q \cdot \log(length(s)) \qquad (2)$$

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 に含まれる単語数をそれぞれ意味しており，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．

クエリを含む文の周辺からも複合名詞を抽出したいため，前後2文の w の値を考慮した以下の式により，最終的な文 s のスコアを決定する．

$$score(s,Q) = w(s,Q) + \sum_{i=1}^{2} \frac{w(s_{-i},Q) + w(s_{+i},Q)}{2^i} \qquad (3)$$

ここで s_{-i} および s_{+i} は，文 s の前後 i 番目に位置する文をそれぞれ表す．

以上のスコアの上位 M 文を選び出し，これらにページのタイトルを追加した $M+1$ 文を重要文として抽出する．本システムでは，$M = 15$ としている．

3.2 複合名詞の抽出

前節で抽出された重要文から，名詞（形式名詞，副詞的名詞は除く）および名詞に相当する語（例：ナ形容詞「重要だ」の語幹である「重要」），助詞「の」が1語以上連続する形態素列を複合名詞とし

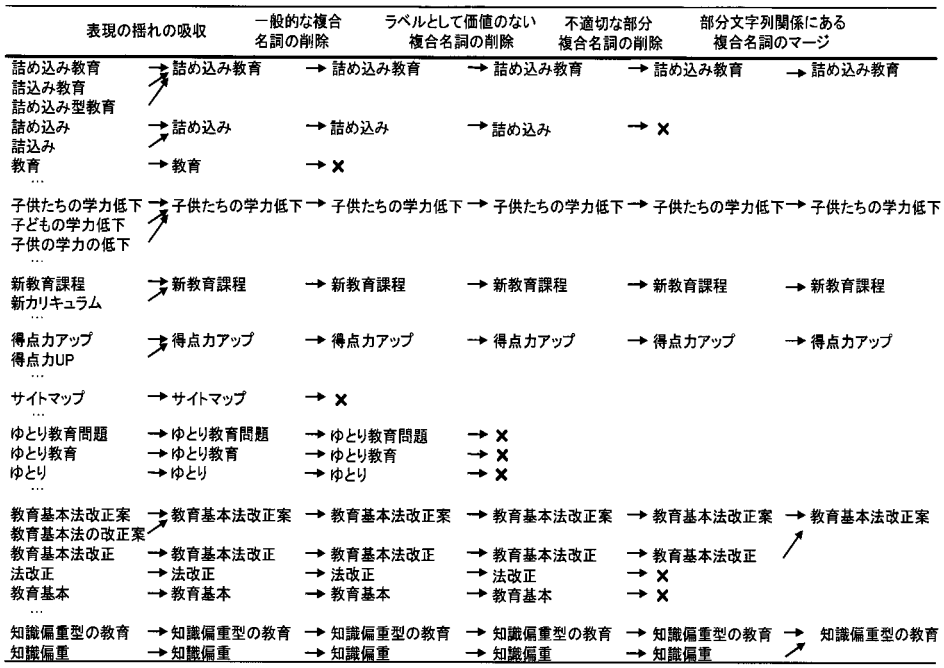

図 3: 処理を経ることで洗練されるラベルの様子（クエリ：ゆとり教育）

て抽出する[4]．このとき，複合名詞として抽出された形態素列の部分列も複合名詞として抽出する．例えば，「こどもの学力低下」からは，「こどもの学力低下」自身に加え，「こども」，「こどもの学力」，「学力低下」といった表現も複合名詞として抽出される．

以上より，クラスタのラベルの候補となる複合名詞が抽出されるが，このままでは，“分数ができない大学生”のような表現をラベルとして持つクラスタを生成することができない．そこで，“「分数ができない大学生」”のようにかぎ括弧（「」，『』）で囲まれている表現もラベル候補として抽出する．

以下では，抽出された複合名詞の集合を**ラベル候補集合**と呼び L で表す．

3.3 表現の揺れの吸収

ラベル候補集合 L 中には，表現の揺れた複合名詞が多数含まれている．そこで，表現の揺れを吸収し，閲覧性の高いラベルを出力するようにするため，表現の揺れの原因を調査した．その結果，以下の 2 種類の原因が顕著であることがわかった．

A: 形態素の表記が揺れている，または異表記である（「詰込み 教育」と「つめこみ 教育」，「新 カリキュラム」と「新 教育課程」，「得点力 UP」と「得点力 アップ」など）

B: 形態素が挿入されている（「詰め込み教育」と「詰め込み 型 教育」など）

本節では，上記の 2 種類の表現の揺れの問題の対処方法について述べる．

3.3.1 形態素の表記揺れ，異表記の吸収

本システムでは，以下の 3 つのリソースを用いることにより，形態素の表記揺れ，異表記の問題に対処する．

- Juman を用いて形態素解析した結果得られる代表表記
- Shibata らがウェブページおよび新聞記事から構築した同義表現データベース [9]
- Nakazawa ら [10] の手法により獲得される訳語関係にある英語とカタカナ表記の対

具体的には，上記のリソースを使って，複合名詞中で表記が揺れている表現，異表記になっている表現を一旦単一の表記に置換する．そして，同じ表記に置換された複合名詞のうち，もっとも文書頻度の高い複合名詞を，代表的な表記として採用する．

[4]名詞 1 語からなる語も便宜上複合名詞と呼ぶこととする．

例えば,「つめこみ教育」,「詰込み教育」,「詰め込み教育」を考えた場合，上述の手順を経ることで「詰め込み教育」が代表的な表記として獲得される．

3.3.2 形態素の挿入による表現の揺れの吸収

本システムでは，複合名詞中に現れる中黒「・」や助詞「の」，助数辞（「年」や「個」など）以外の接尾辞の有無は，複合名詞の表す意味には影響を及ぼさないと仮定し，複合名詞中から該当する形態素を一旦削除する．そして，3.3.1 節で述べた手順に従い，この種の複合名詞についても表現の揺れの吸収を行う．

3.4 不適切な複合名詞の削除

ここでは，ラベルとして不適切な複合名詞をラベル候補集合 L から削除する．以下では，この処理で削除される複合名詞について述べる．

3.4.1 一般的な複合名詞

ラベル候補集合 L 中には,「サイトマップ」や「プライバシーポリシー」のような一般的な複合名詞も含まれる．しかしながら，このような複合名詞は，クラスタのラベルとして不適切である．そこで，一般的な複合名詞をあらかじめ求めておき，これらを L から除くことでラベルの洗練をはかる．

本システムでは，TSUBAKI が検索対象とする日本語ウェブページ 1 億件から抽出した複合名詞のうち，文書頻度の高い上位 500 語を一般的な複合名詞とし，それらを L 中から削除する．

さらに，上述した 500 個の一般的な複合名詞以外に，人手で整備した 101 語の複合名詞（「方向性」や「私たち」など）についても L から削除する．

3.4.2 ラベルとして価値のない複合名詞

多くの場合,「(クエリ) +情報」や「(クエリ) +問題」のような複合名詞はラベルとしての価値がないものと考えられる．そこで，このような複合名詞を L から削除する．さらに，クエリの部分文字列になっている複合名詞についても同様の理由で L から除く．

3.4.3 不適切な部分複合名詞

3.2 節で述べたように，本システムでは部分的な複合名詞をすべて抽出しているため，例えば「教育基本法改正案」からは,「教育基本法改正」のようなラベルとして適切な複合名詞だけでなく,「教育基本」や「基本法改正」といったラベルとして不適切な複合名詞も多く抽出される．そこで，抽出された複合名詞 $l \in L$ について，それ自身がどの程度単独で文書中に出現しやすいかを計算し，ラベルとして不適切な複合名詞を L 中から削除する．本システムでは，複合名詞が適切なラベルかどうかの判定に以下のスコアを用いる．

$$\frac{ldf_{\text{longest}}(l)}{ldf(l)} \tag{4}$$

ここで，$ldf(l)$ は D_Q 中での l を含む文書数，$ldf_{\text{longest}}(l)$ は D_Q において l が複合名詞として単独で出現した文書数を表す．全ての部分複合名詞について上記のスコアを計算した後，その値が α 以上のものをラベルとして適切な複合名詞と見なす．本システムでは α の値を経験的に 0.5 としている．

3.5 部分文字列関係にある複合名詞のマージ

3.4 節までの結果には,「知識偏重」と「知識偏重型の教育」のような，包含関係にある複合名詞の組が存在する．このような複合名詞は，最終的なクラスタリング結果において，ラベル一覧の閲覧性低下の原因となるためマージする．

3.6 クエリと関連性の高い複合名詞の選択

2 節で述べたように，複合名詞抽出処理で抽出された複合名詞を元にクラスタの生成は行われるため，クエリと関係のあるクラスタが生成されるかどうかは，抽出される複合名詞次第である．本システムでは，クエリと特に共起しやすい複合名詞はクエリとの関連性が高いと仮定し，複合名詞 $l \in L$ について，クエリ Q との関連度を表すスコア $score_{rel}$ を計算する．

$$score_{rel}(l) = ldf(l) \cdot \log \frac{1 \times 10^8}{gdf(l)} \tag{5}$$

ここで，$ldf(l)$ は D_Q 中における l の文書頻度，$gdf(l)$ は 1 億件のウェブページから獲得した l の文書頻度を表す．L 中の全複合名詞について，上記のスコアを計算した後，スコアの上位 100 個の複合名詞をラベルとして見なし，残りの複合名詞は破棄する．

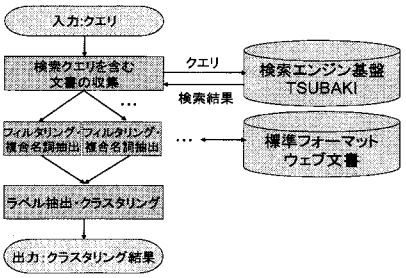

図 4: クラスタリングシステムの構成

4 システム構成

本システムの構成を図 4 に示す．ページタイプの判定処理および複合名詞抽出処理は，16CPU コアで並列に行うことで処理全体の高速化を図っている．実行時間は，任意のクエリについて 1,000 件の検索結果をクラスタリングする場合，およそ 3 分である．

5 評価

この節では，クラスタリングシステムのトピック検出能力についての簡単な評価実験結果および抽出されたラベルの考察について述べる．そして最後に，今後の課題についても述べる．

5.1 トピック検出能力の評価

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 [11]，検索の下位に埋もれたトピックを見逃してしまう．例えば，クエリ「捕鯨問題」について TSUBAKI で検索した結果の上位 1,000 ページ中には，特定の原住民に対して与えられる捕鯨の権利である「原住民生存捕鯨」に関するページが 37 位以下にしか存在せず，発見することが難しい．しかし，クラスタリングシステムを用いることで「原住民生存捕鯨」というクラスタが生成され，このトピックに気づくことができる．この節では，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トピックがクラスタリングシステムを用いることでどの程度検出できるかを，以下の方法により定量的に評価した．

1. 15 個のクエリ[5]について，1,000 件の検索結果に対するクラスタリングを実行する．
2. 生成された計 1,500 個のクラスタ（画面に表示されないクラスタも含む）それぞれに対し，そのクラスタ内のページ中で，検索結果の順位がもっとも上位に位置するページの順位を調査する．

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

表 1: トピック検出能力の評価

順位レンジ	クラスタの個数	割合
1 位〜10 位	410	27.3%
11 位〜30 位	472	31.5%
31 位〜50 位	219	14.6%
51 位〜100 位	244	16.3%
101 位〜300 位	138	9.2%
301 位〜500 位	8	0.5%
501 位〜1,000 位	9	0.6%
合計	1500	100%

実験の結果を表 1 に示す．表より，リスト型の検索エンジンでは下位に埋もれてしまうトピックがクラスタリングによって発見される可能性があることがわかる．

5.2 ラベルの考察

クラスタリングシステムの利用者は，各クラスタのラベルを見て，それぞれのクラスタの内容を推測し，閲覧するクラスタを決定すると考えられる．よって，ラベルの良さはシステムにとって非常に重要な要素である．そこでこの節では，クエリ「ゆとり教育」で検索した 1,000 ページに対するクラスタリング結果 (図 1) について，生成されたクラスタのラベルの考察を行い，クラスタリングシステムの有効性を確かめる．ラベルの評価基準は Geraci ら [12] のものを参考にした．

(a) ラベルは統語的に正しく構成されているか　いずれのラベルも統語的に正しく構成されていると言える．

(b) ラベルからクラスタの内容が推測できるか　ほとんどのラベルは具体的であり，クラスタの中身を想像しやすい．例えば「学力低下」というクラスタでは，ゆとり教育の施行と学力低下との関係性について述べたページが存在していることが推測される．

(c) ラベルはよくクラスタの内容を表していると言えるか　どのクラスタも，内容がラベルと関連のあるページから構成されており，ラベルがクラスタの内容をよく表していると言える．また，表示の際に行っているクラスタ内のページの並び替えの効果により，特にラベルと深く関連するようなページが上部に表示されることが確認できた．

これより，興味のあるページに効率よくアクセスするのにラベルが有効であることがわかる．

5.3 今後の課題

この節では今後の課題について述べる．

5.3.1 より高精度なラベル抽出

同義表現の拡充 本システムで用いた同義表現データベースには，「髪型」と「ヘアスタイル」を同義とみなすエントリが存在しないなど不十分であり，さらなる同義表現の獲得を目指す必要がある．

サ変名詞のラベル 「見直し」や「削除」などサ変名詞が単独で現れるラベルは，対象格の情報を欠いているため不明瞭である．このような場合，動詞の対象格にあたる語の情報も合わせて表示し，分かりやすいラベルを提供する予定である．

ラベルの選択に用いる指標の検討 ラベルの選択には 3.6 節で定義したスコアを用いたが，このスコアではラベルの適切さをうまく測れないことがある．例えば，「注目情報」と「宮沢賢治」は，1 億件のウェブページにおいてほぼ同じ文書頻度を持つので，検索結果中の文書頻度も同じならば，これらのラベルはほぼ同じスコアを持つことになる．しかし直観的には，後者の方がラベルとしての適切さは高い．このことから，ラベルとしての適切さを測るには，文書頻度のみでは不十分であると考えられる．このため，ラベルの選択に用いる指標として，例えば Zeng ら [3] による方法や，久光ら [13] の方法との比較検討を行う予定である．

5.3.2 選択したクラスタの再クラスタリング

ユーザが，自分の興味を持ったクラスタをさらに詳しく調べたいときのために，選択したクラスタを高速に再クラスタリングする機能を追加する予定である．

5.3.3 正解セットを用いた評価

擬似的な正解セットを構築し，システムの出力の評価を行う必要がある．具体的には，クエリに対して関連のあるキーワードを適当な数用意し，クエリとキーワードを AND 検索して得られるページ集合を正解のクラスタとみなし評価することを予定している．

6 おわりに

本稿では，検索エンジン基盤 TSUBAKI を使って検索されたウェブページを，ページ中の複合名詞に注目して自動的にクラスタリングするシステムについて述べた．本システムでは検索エンジン TSUBAKI の提供する構文解析済みのウェブページを利用し，大量の検索結果を対象に，深い言語処理に基づいた複合名詞抽出を行う仕組みを実現した．簡単な評価の結果，本システムを用いることで検索結果の閲覧性が向上することが確かめられた．

参考文献

[1] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, Vol. 36, No. 2, pp. 3–10, 2002.

[2] O. Zamir and O. Etzioni. Grouper: A Dynamic Clustering Interface to Web Search Results. *Proc. of the 8th International World Wide Web Conference*, Vol. 31, No. 11-16, pp. 1361–1374, 1999.

[3] H.J. Zeng, Q.C. He, Z. Chen, W.Y. Ma, and J. Ma. Learning to cluster web search results. *Proc. of the 27th annual international conference on Research and development in information retrieval*, pp. 210–217, 2004.

[4] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. *International World Wide Web Conference*, pp. 801–810, 2005.

[5] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Proc. of IJCNLP2008*, 2008.

[6] 新里圭司, 橘本力, 河原大輔, 黒橋禎夫. 自然言語処理基盤としてのウェブ文書標準フォーマットの提案. 言語処理学会第 13 回年次大会論文集, 2007.

[7] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of japanese morphological analyzer juman. In *The International Workshop on Sharable Natural Language*, pp. 22 – 28, 1994.

[8] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, No. 4, pp. 507–534, 1994.

[9] Tomohide Shibata, Michitaka Odani, Jun Harashima, Takashi Oonishi, and Sadao Kurohashi. SYNGRAPH: A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *Proc. of IJCNLP2008*, 2008.

[10] Toshiaki Nakazawa, Kun Yu, Daisuke Kawahara, and Sadao Kurohashi. Example-based machine translation based on deeper nlp. In *Proc. of International Workshop on Spoken Language Translation (IWSLT'06)*, pp. pp.64–70, 2006.

[11] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. 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.

[12] Filippo Geraci, Marco Pellegrini, Paolo Pisati, and Fabrizio Sebastiani. Cluster generation and cluster labelling for web snippets. In *Proc. of the 13th Symposium on String Processing and Information Retrieval (SPIRE 2006)*, Vol. 4209, pp. 25–36, 2006.

[13] 久光徹, 丹羽芳樹, 辻井潤一. タームの representativeness を測る. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 99, No. 73, pp. 115–122, 1999.