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

2008－DBS－146 (51)
2008／9／23

単語の特徴量を考慮した検索結果クラスタに関する多視点融合型スニペットの構築

村松 亮介† 横山 昌平†† 福田 直樹†† 石川 博††

†静岡大学大学院情報学研究科 〒432-8011 静岡県浜松市中区城北 3-5-1

††静岡大学情報学部情報科学科 〒432-8011 静岡県浜松市中区城北 3-5-1

E-mail: †gs08062@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta, ishikawa}@inf.shizuoka.ac.jp

あらまし 我々はこれまでにユーザが目的とするWebページ発見の支援や検索結果の概観把握を目的として，検索結果のクラスタリングとラベリングを行うシステムとしてSearchLifeを実装した．しかし，クラスタ閲覧の際の指針となるクラスタラベルが単一の観点で生成されるものであるため，ラベル一覧にユーザの求めている関連語が存在しない場合は，検索結果の線形的な閲覧，もしくは再検索をする必要があった．本研究では，この問題を踏まえ，さらなる閲覧性向上にむけて，Web全体集合と，あるクエリに対する検索結果集合という異なる２種の集合に対して，それぞれの単語の特徴量の違いを考慮することで，検索結果中の単語の専門性，一般的認知度に基づく分類を行い，それらを用いて生成クラスタに対する多視点融合型スニペットを構築する．

キーワード 情報検索，Webとインターネット，データマイニング

Architect snippets with harmonized various view point about search result cluster with consideration of word's characteristic volume

Ryosuke MURAMATSU†, Shohei YOKOYAMA††, Naoki FUKUTA††, and Hiroshi ISHIKAWA††

†Graduate School of Informatics, Shizuoka University Johoku 3-5-1 , Naka-ku, Hamamatsu-shi, Shizuoka, 432-8011 Japan

††Department of Computer Science, Faculty of Informatics, Shizuoka University Johoku 3-5-1, Naka-ku, Hamamatsu-shi, Shizuoka, 432-8011 Japan

E-mail: †gs08062@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

Abstract We have developed "SearchLife" that clusterizes search results and labels cluster to help the user find destination pages and catch a overal view of search results.But because clusters' labels that are guides to watch clusters are made by simple viewpoints, the users have to check search results linearly or search by different queries when there are no relative words in labels. In this paper, we classify words in search results based on both specialty and commonality with respect to word features in a set of web and a set of search results and create snippets with harmonized various viewpoints about clusers of search results.

Key words Information Retrieval, Web and Internet , Data Mining

1. はじめに

Web上の情報量は増加の一途を辿っている．そのような膨大な情報の中から必要な情報を取得するツールとして検索エンジンが一般的に利用されている．代表的な検索エンジンであるGoogle[1]やYahoo[2]が提供している検索エンジンの検索結果は，独自のランキングに基づくリスト表示に基づいているため，検索結果が膨大な場合，ユーザが検索結果の概観を捉えることや必要な情報をすぐに探し出すには限界がある．そこで，我々はこれまでにユーザが目的とするWebページ発見の支援や検索結果の概観把握を目的として，検索結果のクラスタリングとラベリングを行うシステムとして

SearchLife[3]を実装した．しかし，クラスタ閲覧の際の指針となるクラスタラベルが単一の観点で生成されるものであるため，ラベル一覧にユーザの求めている関連語が存在しない場合，検索結果の線形的な閲覧，もしくは再検索をする必要があった．本研究では，この問題を踏まえ，さらなる閲覧性向上にむけて，既存のラベリング手法に加えて，Web 全体集合と，あるクエリに対する検索結果集合という異なる２種の集合に対する，それぞれの単語の特徴量の違いを考慮することで，検索結果中の単語の専門性，一般的認知度に基づく分類を行い，検索クエリに対する常識語ラベルと分野特徴語ラベルを生成する．

本論文での常識語とは，Web 全体においては使用頻度が低く，特徴的であると考えられるが検索クエリに対する検索結果集合中では一般的である単語を指す．常識語は，検索結果集合において，最低限知っておくべき単語であり，検索クエリに関して知識が少ない人の学習支援に役立つことが期待される．

本論文での分野特徴語とは，Web 全体においては使用頻度が高く，一般的であると考えられるが検索クエリに対する検索結果集合中では使用頻度が低く特徴的である単語を指す．分野特徴語は，検索クエリに関して，より情報の焦点を絞った，ある種マイナーな事柄を得たい場合に役に立つことが期待される．

本論文では既存のラベリング手法に加えて，上記２種のラベルを新たに導入し，生成クラスタに対する多視点融合型スニペットを構築する．

2. 関連研究

2. 1 検索結果のクラスタリング

検索結果のクラスタリングに関する研究は大きく二つに分類できる．１つは，Web ページの内容に着目してクラスタリングを行うコンテンツマイニングであり，もう１つは，Web ページのリンク情報に基づいてクラスタリングを行うストラクチャマイニングである．コンテンツマイニングを行う研究として，例えば成田ら[4][5]の研究がある．ストラクチャマイニングを行う研究として大野ら[6]の研究がある．成田らの研究では生成されたクラスタ，ラベルの有用度に関して未評価であり，大野らの研究ではクラスタに分類されないページが多いという課題がある．

また，現在 Web 上に公開されているクラスタリングサーチエンジンとして Clusty[7]がある．Clusty はメタ検索エンジンの一種で，検索結果を階層的にクラスタリングして，画面左にクラスタをツリー型メニューとして表示し，画面右に選択したクラスタに属する Web ページがリスト表示される．Clusty は“Velocity”と呼ばれる独自クラスタリングエンジンを利用しており，文書を意味のあるグループに自動組織化する．また，2008 年１月より，新機能として remix 機能が追加された．remix とは提示されたクラスタリング結果がユーザの意図と一致しなかった場合，つまり，ラベル一覧に求めている情報が存在しなかった場合に，別の観点で再クラスタリングを行う機能である．remix はユーザにとって適切なクラスタが生成されるまで繰り返される．本研究ではクラスタ構成は一定のまま，生成クラスタに対するラベリングを複数の観点で行うことで，ユーザの目的 Web ページ発見の支援を行う．

2. 2 スニペットの生成

検索エンジンから返される Web ページのスニペット構築に関して，Jae-wook ら[8]や高見ら[9]の研究がある．

Jae-work らは，Web ページに対するパーソナライズ化されたスニペットをユーザに提示する手法を提案している．ユーザの現在直面している課題からタスクモデルを構築し，そのモデルに基づいて，検索結果に表示される各 Web ページのスニペットを構築する．スニペット構築では，タスクの干渉度を調節し，３種のスニペットを構築し，ユーザは状況に応じて，スニペットを選択する．

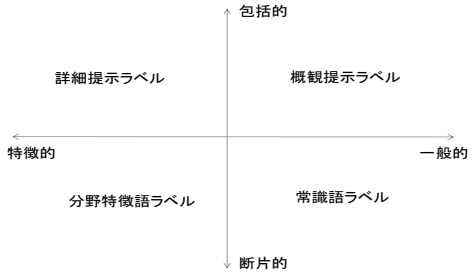

図１：多視点融合型スニペット

高見らは，スニペットを，その生成方法により２種類の軸で分類した．そして，web ページに対して４種類のスニペットを生成することで，ユーザの検索目的に適したスニペットを提示した．これら２つの研究は，１つの web ページに対する多角的なスニペットの提示を目指したものであるが，我々は，検索結果クラスタという，ある類似性を持った web ページ群に対するスニペットの構築手法の実現を目的とする．

2.3 関連語抽出

野田[10]らは検索結果のクラスタリングに用いるためのキーワードとして，質問キーワードに対する話題語の抽出を行った．本研究では単一の種類の単語ではなく複数の種類の単語を抽出し，それらを用いることで検索結果クラスタに対するスニペットを構築する．

3 多視点融合型スニペットの構築

本節では多視点融合型スニペットの構築手法に関して述べる．本システムでは生成された検索結果クラスタに対して図１に示すように以下の４種類のラベルを付与する．

①概観提示ラベル

生成されたクラスタ全体を包括する単語であり，１つの名詞からなる．特徴的な単語ではなく２タイトル以上で共起する単語をラベルとしており，検索結果集合においての概観の把握に役に立つことが期待される．

②詳細提示ラベル

各クラスタ内での特徴的な単語であり，複数の名詞により構成される．より細かいクラスタ内の文書内容を知ることができる．ある明確な検索対象が存在する場合，本ラベルが利用できる．

③常識語ラベル

各文書より断片的に抽出される単語である．常識語を見ることで，検索クエリ分野に関して知識のないユーザのナビゲートを行う．その分野に詳しい人なら常識的に知っているが，その分野に関して知識の少ない人には理解の難しい内容が検索結果に含まれることがある．そのような場合に，検索クエリに対する常識語を閲覧することで，検索クエリの結果の閲覧に関してあらかじめ知っておくべき単語をユーザが学習することができる．

④分野特徴語ラベル

各文書より断片的に抽出される単語である．検索クエリに関して，マイナーな情報を得たいとき，このラベルを手がかりにすることで，ユーザの閲覧行動を補助できる．この単語

自体は Web 全体集合ではよく使われる単語であるが，検索クエリに対する検索結果集合においては，あまり使われなくなる単語であり，マイナーな事柄である可能性が高い．

ユーザは，各自の検索要求や検索対象に関する知識レベルに応じて，閲覧するラベルを選択することで，より早く目的の Web ページを発見でき，閲覧性が向上すると考えられる．

3．1　概観提示ラベルおよび詳細提示ラベルの生成

本節では文献[3]で述べたクラスタリングおよびラベル生成手法について，概要を示す．

3．1．1　検索結果の取得および形態素解析

本システムの概要を図２に示す．本システムでは最初に，Yahoo!Japan デベロッパーネットワーク[11]が提供するウェブ検索 Web サービスを利用して検索結果上位 100 件のタイトル，サマリ，URL を取得する．

次に，同デベロッパーネットワークが提供する日本語形態素解析 Web サービスを利用して，上記で取得した検索結果 100 件のタイトル，サマリ，URL の形態素解析を行い，名詞のみを抽出する．ここで，本サービスを用いて例えば人名‘村松亮介’を形態素解析した場合，‘村松’，‘亮介’のように２つの名詞として抽出されてしまう．そこで２回連続して名詞が出現した場合には１つの名詞として抽出した結果をテーブル１として保存し，サービスからのそのままの返却結果をテーブル２として保存する．

3．1．2　*idf* 算出および特徴語の抽出

各文書においてその特徴を表すと思われる単語を抽出するためタイトルに出現する名詞の *idf* 値を算出する．単語 t が出現する文書数を $dt(t)$とし，Nを比較文書数とすると，式(1)のように表すことができる．ここで比較文書数 N は 548 億に設定する．また，$dt(t)$はウェブ検索 Web サービスを利用したときの検索クエリ t に対する検索結果ヒット件数とする．

$$idf = \log \frac{N}{dt(t)} \qquad (1)$$

上記式(1)によって形態素解析結果であるテーブル１に保存される全名詞の *idf* を求め，各タイトルにおいて以下の２条件を満足する名詞を特徴語とする． タイトル内に条件を満足する名詞が存在しない場合はサマリ，URL の順で同様の処理を行い，条件を満足する名詞を探索する．

条件１：*idf* 最大値

条件２：検索クエリの部分文字列ではない

以上の条件を設定した理由は３．１．３節手順(1)で述べる．

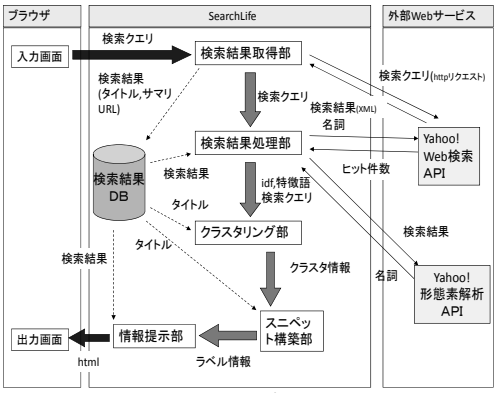

図２：システムの概要

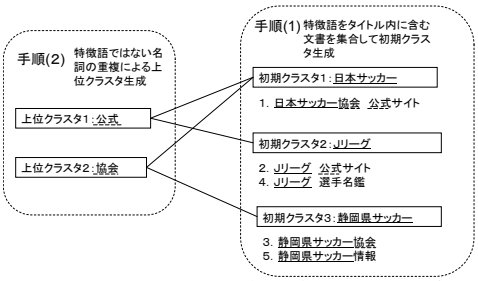

図３：クラスタサイズの平均化とラベリング

3．1．3　クラスタリングおよび概観提示ラベルの抽出

本提案手法における検索結果のクラスタリングとラベリング手法の概略を図３に示す．

手順(1)　３．１．２節で求めた特徴語集合内における特徴語の出現回数を計測する．その出現回数 *tf* と特徴語の *idf* を用いて式(2)で表される *tfidf* を算出し，検索結果集合における重要単語のランキングを行う．このランキングは手順(2)の処理によって生成されるクラスタ内における表示順序を示す．

$$tfidf = tf \cdot idf \qquad (2)$$

重要単語をタイトルに含む文書を集めて，クラスタを形成する．ここでは非排他的クラスタリングを行い，各文書が２個以上のクラスタに含まれることを許す．以下では，ここで作成されるクラスタを初期クラスタと呼ぶこととする．また，クラスタリングの指標とした重要単語を各初期クラスタのラベルに設定し，これを詳細提示ラベルとする．このとき３．１．２節の条件２を付加しない 場合，例えば検索クエリが“静岡大学”のとき特徴語として“静岡大学”や“静岡”が選択される可能性がある．例えばこの例の場合，検索クエリ“静岡大学”に対する取得検索結果 100 件中 59 件がタイトル内に“静岡大学”を含み，76 件が“静岡”を含んでいた．同様に他の５個の検索クエリで行なった結果，取得検索結果 100 件中平均 72 件がタイトル内に検索クエリを含んでいた．このような検索対象に対して上記のクラスタリングを行うと，タイトル内に検索クエリが存在する文書を１つのクラスタに集合させることになり１クラスタに膨大な文書が含まれてしまい，閲覧性が低下する．また，初期クラスタラベルとして検索クエリが設定されることになり，クラスタとしての有効性が低下する．そこで，我々の手法では，条件２を付加することで，クラスタ内文書数の平滑化を図り，意味のあるラベルが設定されるようにする．作成される初期クラスタの例を図４に示す．

手順(2)　手順(1)の手法では得られなかったタイトル間における名詞のつながりを発見するため，形態素解析結果であるテーブル２に保存される名詞で以下の条件を満足する名詞を発見する．

条件１：特徴語ではなく，２タイトル以上に出現する名詞

条件２：その名詞が検索クエリの部分文字列ではない

条件３：その名詞の *idf* が１．５以上

条件２については，手順(1)での理由と同じである．条件３については，ラベルとして意味を成さないと思われる語，例えば com，jp，co など多くの Web ページで使用される名詞を排除するため，経験的に設定した．以上の条件を満たす名詞が使われているタイトルを含む初期クラスタを併合し，新たにクラスタを作成する．以下では，このクラスタを上位クラスタと呼ぶこととする．上位クラスタ内の初期クラスタの表

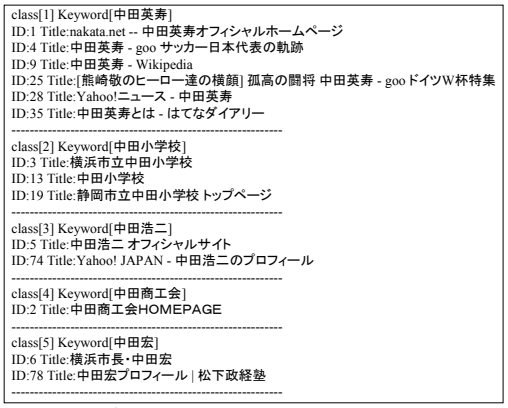

図４:検索クエリ“中田”の初期クラスタの例

示順序は(1)で求めた *tfidf* によるランキングに従うこととし，その名詞を上位クラスタのラベルに設定し，これを外観提示ラベルとする．併合が行なわれなかったクラスタに関して，初期クラスタ内文書が１個のクラスタに関しては“その他”のクラスタに分類する．例えば図４のような初期クラスタに対して上記の処理を行う場合，class[2]に所属する ID3 のタイトル内の“横浜”という名詞は class[5]の ID6 のタイトル内にも出現する．この場合，これら２個の初期クラスタを併合して“横浜”をラベルとする上位クラスタを作成する．

３．２　常識語ラベルおよび分野特徴語の生成

本節では，検索結果文書内からの，常識語および分野特徴語の抽出手法に関して述べる．

３．２．１　条件付き特徴量 idf(t|q)

3.1節までに生成された概観提示ラベルと詳細提示ラベルは，Web 全体集合において各単語が一般的もしくは特徴的な単語であるかを考慮して生成された．しかし，その同じ単語が検索クエリに対する検索結果集合内においても同じように一般的もしくは特徴的であるとは限らず，その特性が逆転することがある．本研究では，２個の異なる集合における特徴量の差異を考慮することで，常識語と分野特徴語の抽出を試みる．ここで，検索クエリ q に対する検索結果集合における単語 t の特徴量を，条件付き特徴量 idf(t|q)と呼び，式(3)により求める．

$$idf(t \mid q) = \log \frac{dt(q)}{dt(q+t)} \qquad (3)$$

dt(q)は単語 q に対する検索結果ヒット件数，dt(q+t)は単語 q と t の and 検索による検索結果ヒット件数を表す．

３．２．２　常識語と分野特徴語の抽出

本節では以下の仮説のもとに常識語と分野特徴語の抽出を行う．

仮説（１） Web 全体集合では特徴的であり式(1)idf(t)の値が高いが，検索クエリ q に対する検索結果集合において一般的であり式(2)idf(t|q)が低い単語は，検索結果集合において，知っているべき単語，つまり常識語である．

仮説（２） Web 全体集合では一般的であり式(1)idf(t)の値が低いが，検索クエリ q に対する検索結果集合において特徴的であり式(2)idf(t|q)が高い単語は，検索結果集合において，マイナーな情報であることを示し，該当する単語を分野特徴語とする．

以下に常識語と分野特徴語の抽出手順を示す．

手順（１）条件付き特徴量の算出および単語の分類

まず，各文書のタイトル内の全名詞の条件付き特徴量を式(3)により求める．次に，それぞれの集合における特徴量の平均値によって，各単語が一般的もしくは特徴的であるかの判定を行う．手順(1)で求めた idf(t|q)と 3.1.2 節で求めた idf(t)の，それぞれの平均値 ave_idf(t|q)，ave_idf(t)を計算する．そして，平均値より高い単語を特徴的，低い単語を一般的であると設定し，仮説に沿って以下のように単語を分類する．

・常識語の候補語
　idf(t)>ave_idf(t)かつ idf(t|q)<ave_idf(t|q)
・分野特徴語の候補語
　idf(t)<ave_idf(t)かつ idf(t|q)>ave_idf(t|q)

手順（２）各候補語の絞り込み

Web ページ本文中のどこかに検索クエリが含まれているだけで，内容的に検索クエリとは無関係な文書に関しても手順（1）においては候補語と判定される可能性がある．本研究では検索クエリに対する常識語もしくは分野特徴語の抽出を目的としているため，タイトル内に検索クエリが含まれている場合に限って，常識語もしくは分野特徴語とする．

手順（３）傾きによる単語のランキング

ある単語の Web 全体集合と検索結果集合での特徴量のギャップの大きさに基づくランキングを行う．X 軸を idf(t|q)，Y 軸を idf(t)とみなしたときの常識語と分野特徴語の傾きを式(4)により計算する．常識語に関しては傾きが大きいほどランキングは高くなり，分野特徴語に関しては傾きが低いほどランキングが高くなる．ここで求めたランキングは，実装における単語の提示順序に影響を与える．

$$dividf(t) = idf(t) / idf(t \mid q) \qquad (4)$$

４．多視点融合型スニペットの評価

４．１　概観提示ラベルおよび詳細提示ラベルの評価

本節では概観提示ラベルおよび詳細提示ラベルに関して，クラスタ内文書とラベルとの定量的な評価を行う．

（ｉ）評価方法

本論文では，成田ら[4]の実験で提案されているクラスタ再現率，クラスタ適合率，クラスタリング率と一般的な F 値(調和平均)を用い，提案手法によって生成された上位クラスタと概観提示ラベルの妥当性を評価する．まず，各用語の定義を表１に示す．

検索結果全体において，あるクラスタに含まれるべき検索結果のうち，実際に含まれている検索結果の割合を示す指標を，クラスタ再現率(*recall*)として式(5)により定義する．

$$recall = \frac{|\beta|}{|\alpha|} \times 100 \qquad (5)$$

あるクラスタに含まれている検索結果のうち，そのクラスタに含まれるべき検索結果の割合を示す指標を，クラスタ適合率(*precision*)として式(6)により定義する．

$$precision = \frac{|\beta|}{|clst|} \times 100 \qquad (6)$$

全検索結果のうち，どれだけの検索結果が「その他」以外のクラスタに振り分けられたかを示す指標として，クラスタリング率(*clster*)を式(7)により定義する．

$$clster = \frac{|org| - |etc|}{|org|} \times 100 \qquad (7)$$

表１：用語の定義

被評価検索エンジン	本システム
判定用検索エンジン	Yahoo!Japan デベロッパーネットワークウェブ検索 Web サービス
Org	被評価検索エンジンで検索したときの検索結果全体(100 件)
Clst	org のうち「クラスタラベル」のクラスタに含まれている検索結果
α	org のうち判定用検索エンジンで「検索キーワード」＋「クラスタラベル」で AND 検索を行なったとき，その上位１００件の検索結果に含まれている検索結果
β	αのうち，*clst* に含まれる検索結果
Etc	「その他」クラスタ

注：各記号を｜｜で括ったものはその要素数を示す.

表２：５つの検索クエリについての実験結果

検索クエリ	平均再現率	平均適合率	クラスタリング率
無料	49．8%	71．5%	87．0%
壁紙	41．2%	75．1%	68．0%
アイドル	57．0%	71．4%	65．0%
ワールドカップ	39．1%	58．0%	75．0%
チケット	37．6%	71．8%	80．0%

表３：５つの検索クエリについての集計と比較

システム名	平均再現率	平均適合率	F 値	クラスタリング率
成田ら[4]	28．7%	83．3%	42.7	68．5%
提案手法	40．9%	67．8%	51.0	75．0%

再現率と適合率のＦ値(調和平均) を式(8)により定義する.

$$F値 = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (8)$$

（ii） 実験結果と評価

成田ら[4]の研究における実験用検索クエリ‘無料’,‘壁紙’,‘アイドル’,‘ワールドカップ’,‘チケット’を本システムでクラスタリングしたときの平均再現率と平均適合率, F 値，クラスタリング率を表２に示す．本システムと成田ら[4]のシステムにおける５つの検索クエリに対する平均再現率と平均適合率,F 値,クラスタリング率の集計を表３に示す．表3から成田らの手法と比較するとクラスタ再現率とクラスタリング率は上昇，クラスタ適合率は低下していることが分かる．クラスタ再現率に関しては非排他的クラスタリングを行なったことで上昇した．クラスタリング率に関しては初期クラスタに対してクラスタリング結果の改善手法を適用したことで上昇した．適合率に関しては 3．1．3 節手順(2)における初期クラスタ併合の際にラベルとは関係のない文書が属してしまっていることが低下の要因であると考えられる．F 値に関して，上昇していることから概ねクラスタ内文書とそのラベルの適切性は保障されていると言える．

４．２　常識語ラベルと分野特徴語ラベルの評価

本節では常識語ラベルと分野特徴語ラベルに関して，定性的な評価を行う．表４に検索クエリ「Java MySQL」に対する常識語全５件，表５に分野特徴語 35 件中上位 20 件を示す．

まず，検索クエリ「Java MySQL」に対する常識語の結果を考察する．個々の単語の意味は Java や MySQL の使用経験のある人にとっては概ね理解できるが，使用経験がない人にとっては理解し難い単語が列挙されている．まず，Rank2 の「JDBC」は Java と MySQL の接続に用いるドライバであり，検索クエリとの関連性は高いと考えられ，妥当である．また，Rank3 の「Tomcat」は Java を用いたアプリケーションサーバーであり，データベースとして MySQL を用いることも多いため，妥当である．その他の単語に関しても何かしら検索クエリとの関連が想像し得るものであり，常識語に関して良好な結果を得ていると言える．

しかし，常識語と認定されなかった単語に関して「JSP」,「Apache」などがあり，クエリと関連性のある単語であり，Java や MySQL を知っている人であれば，これらの単語も知っている可能性が高い．よって，常識語の抽出に関して，より幅広い抽出手法を考えていきたい．

次に分野特徴語について考察する．個々の単語の意味はおそらく多くの人が理解できるのだが，検索クエリとの関連性は想像し難い単語が列挙されている．まず，分野特徴語として抽出成功している例を挙げる．まず，Rank2 の「掲示板」という単語の元文書のタイトルは「Java + MySQL + Tomcat で作る掲示板とブログ」であり，限定性があり内容的にマイナーな情報である可能性が高いと考えられる．また，Rank6 の「影響」という単語の元文書のタイトルは「InfoQ: Sun が MySQL を買収:その展望と，影響の分析」であり，「Java MySQL」という検索クエリに対しては内容的に技術的な情報が多い中にあって，本ページは企業の経営的な情報を記述しており，検索結果集合においてはマイナーな情報であると言える．

次に分野特徴語として抽出失敗している例を挙げる．Rank5 の「竹」は人名「竹形誠司」の１部であり，正しく形態素解析されれば idf(t)の値は高くなると考えられ,分野特徴語には該当しない．Rank8 の「道」という単語の元文書のタイトルは「Java の道」であり，道という単語が Java という単語と共に使われることは珍しいように思われるが，実際の内容は Java に関する質問掲示板であり，あまりマイナーな情報であるとは言えない．このように分野特徴語に関して，単純に検索クエリと分野特徴語の共起の少なさによって抽出されてしまう単語が多くなってしまった．よって，分野特徴語の抽出手法に関して，分野特徴語と本文の内容を照らして，本当に内容においても貴重な情報であるかを確認する処理を追加する必要がある．

５．検索結果提示画面

図５に Web ブラウザとして Mozilla Firefox[12]を用いたときの検索結果提示画面を示す．画面左に各種ラベルが表示され，ユーザは自分の検索要求にあったラベルを選択する．また，画面右にはユーザが選択したラベルのクラスタ内文書が表示される．

６．おわりに

本論文では検索エンジンの検索結果のクラスタリングによって生成されるクラスタに対する多視点融合型スニペットの構築手法を提案した．4 種類のラベルのうち，概観提示

表 4：「Java MySQL」に対する常識語全５件

Rank	dividf(t)	idf (t|q)	idf (t)	常識語
1	3.205	2.199	7.047	PostgreSQL
2	2.249	3.419	7.689	JDBC
3	2.117	3.418	7.235	Tomcat
4	1.846	3.537	6.528	FreeBSD
5	1.426	4.748	6.771	MACOS

表 5：「Java MySQL」に対する分野特徴語の上位２０件

Rank	dividf(t)	idf(t|q)	idf(t)	分野特徴語
1	0.432	6.465	2.795	FLUSH
2	0.657	6.64	4.363	掲示板
3	0.676	7.705	5.208	楽天
4	0.687	5.099	3.503	記事
5	0.701	8.601	6.029	竹
6	0.705	6.986	4.925	影響
7	0.711	6.604	4.699	登場
8	0.72	5.82	4.189	道
9	0.733	7.638	5.598	サン
10	0.746	7.352	5.486	Jungle
11	0.76	6.388	4.852	フリー
12	0.762	6.661	5.077	アーカイブ
13	0.764	5.406	4.13	アクセス
14	0.79	6.4	5.056	編
15	0.802	7.341	5.888	Neon
16	0.802	7.409	5.942	求人情報
17	0.804	6.326	5.085	使い方
18	0.829	5.933	4.917	ダウンロード
19	0.842	5.686	4.789	サポート
20	0.842	5.383	4.533	207

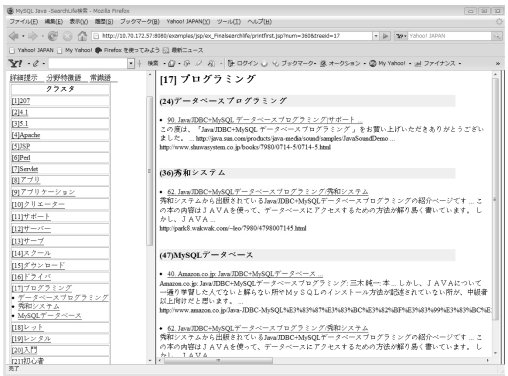

図 5：検索結果提示画面

ラベル，詳細提示ラベルの２種のラベルに関しては，定量的評価の結果，既存の研究と比較して良好な結果が得られた．また，常識語ラベルに関しても，定性的な評価の結果，良好な結果が得られた．分野特徴語ラベルに関しては，分野特徴語として適切な単語より，不適切な単語のほうが多く抽出されてしまう結果となった．

今後の課題として，常識語と分野特徴語に関して，その抽出手法のさらなる改善やその定量的評価手法の検討が挙げられる．定量的評価手法としては，大規模な Web 検索クエリログを用いた評価手法の適用が挙げられる．つまり，検索クエリとして多く使用される単語は常識語と考え，逆にあまり使われない単語を分野特徴語と考えて，使用回数等により，数値化できるのではないかと考えられる．また，本論文で提案した多視点融合型スニペットを実際に検索の際に用いたときのユーザビリティに関する実験に基づく評価を行うことも，この後の課題である．

謝辞

本研究の一部は科研費基盤 B(19300026)の助成を受けたものである．

７．参考文献

[1] Google
http://google.com
[2] Yahoo!Japan
http://www.yahoo.co.jp/
[3] 村松亮介，福田直樹，石川博 “分類階層を利用した検索エンジンの検索結果の構造化とその提示方法の改良”DEWS2008 B6-3
[4] 成田宏和，太田学，片山薫，石川博 “階層的クラスタリングを利用したメタサーチエンジンの提案”研究報告「データベースシステム」アブストラクト No.128-050,pp.375-382,July.2002.
[5] 成田宏和，太田学，片山薫，石川博 “Web 文書検索のための非排他的クラスタリング手法の提案”DEWS2003 2-p-01
[6] 大野成義，渡辺匡，片山薫，石川博，太田学 “Max Flow アルゴリズムを用いた Web ページのクラスタリング方法の提案”日本データベース学会 Letters Vol.4,No.2,September.2005
[7] Clusty
http://www.clusty.com/
[8] Jae-wook Ahn, Peter Brusilovsky, Daqing He, Jonathan Grady, Qi Li “Personalized Web Exploration with Task Models”WWW2008/Refereed Track:Browsers and User Interface,April 21-25,2008・Beijing ,China
[9] 高見真也，田中克己 “検索目的に基づくスニペットの動的再生成によるウェブ検索結果の個人適応化”日本データベース学会 Letters Vol. 6, No. 2
[10] 野田武史,大島裕明,手塚太郎,小山聡,田中克己 “Web 検索結果のクラスタリングに用いる話題語の質問キーワードからの自動抽出” DEWS2006,2C-i8
[11] Yahoo!Japan デベロッパーネットワーク
http://developer.yahoo.co.jp/
[12] Mozilla Firefox
http://www.mozilla-japan.org/