Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

多視点融合型クラスタリング検索エンジンの開発と評価について

村松亮介[†1]　横山昌平[†2]
福田直樹[†2]　石川博[†2]

検索エンジンは，Web 上の膨大な情報を検索する手段の１つとして，重要な役割を果たしている．代表的な検索エンジンである Google や Yahoo!が提供している検索結果の提示手法はランキングに基づくリスト表示に基づいているため，必要な情報をすぐに探し出すことが難しい場合がある．そこで，我々は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．クラスタ閲覧の手がかりとなるクラスタラベルが単一の手法で生成されている場合，ラベル内にユーザの求めている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生じる．本論文では，この問題を踏まえ，さらなる閲覧性向上を目的として，用語の専門性や一般への認知度の違いを反映するような複数のラベリング手法を用意する．また，クラスタリングによる検索結果表示に対するユーザビリティ評価を行い，提案システムを用いた場合においてページ収集を有効に行える場合があることを示す．

Development and evaluation of a clustering search engine with harmonized multiple viewpoints

Ryosuke Muramatsu,[†1] Shohei Yokoyama,[†2]
Naoki Fukuta[†2] and Hiroshi Ishikawa[†2]

Web search engines have important roles as tools to get necessary information from the web. Since search results are usually provided as ranking based listings, there are some difficult cases when the users seek find necessary information from them quickly. We have developed “ SearchLife ” that clusterizes search results and give labels to clusters to help the users find demanded pages and catch an overal view of search results. When clusters’ labels are made by a simple way, the users may need to read search results linearly or search again by different queries since sometimes there are no helpful words in the labels. In this paper, we provide a multi way labeling for clusters that reflect differences of domain locality and commonality of words for improved view of results. We show that the usability of our labeling strategies for clustering search system has advantages in some complex searching tasks.

1. まえがき

検索エンジンは，Web 上の膨大な情報を検索する手段の１つとして，重要な役割を果たしている．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，ランキングに基づくリスト表示に基づいているため，検索対象やそれに対する検索クエリの与え方によっては，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ことが難しい場合がある．検索エンジンにおける検索結果提示の改善方法として，クラスタリングに基づくアプローチがある[3)–9)]．しかしながら，クラスタリングに基づく検索結果の提示手法には，クラスタリングを高精度に行うということ以外にも，クラスタへの適切なラベル付けや，そのラベル付け手法の目的に応じた選択を行えるようにすることで，改善が行える余地がある．我々はこれまでに Web ページ検索結果の閲覧性向上のための，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[9)] を実装した．しかし，クラスタ閲覧の手がかりとなるクラスタラベルが単一の手法で生成されているため，ラベル内にユーザの求めている語が存在しない場合にはクラスタラベルが閲覧のための有用な手がかりとならず，検索結果の線形的な閲覧，もしくは再検索をする必要が生じる．本論文では，この問題に対し，閲覧性の向上のためにクラスタに対するラベリング手法を複数用意し，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したラベリング手法を用意することで，検索結果中の単語の専門性，一般的認知度の違いを反映するような，複数のラベル付けを同一のクラスタリング結果に対して行えるようにする．なお，本論文における上記のアイデアは文献[10)] が前身となっている．文献[10)] では文章によるクラスタに対するスニペットの提示を目的としていたが，本論文では，単語によるクラスタに対するラベリングを行うものである．本論文では 4 種類のラベリング手法を用意する．1 種類目と 2 種類目はクラスタ内文書

†1 静岡大学大学院情報学研究科
Graduate School of Informatics, Shizuoka University
†2 静岡大学情報学部情報科学科
Department of Computer Science, Faculty of Informatics, Shizuoka University

Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

のいくつかに共通して出現する単語である．前者は後者と比較して Web 上で使用頻度が高く，後者は前者と比較して Web 上で使用頻度が低い単語であるという特徴を持つ．3 種類目は 4 種類目と比較して Web 全体集合では使用頻度は低いが，検索クエリに対する検索結果集合中では使用頻度が高くなる単語である．4 種類目は 3 種類目と比較して Web 全体集合では使用頻度は高いが，検索結果集合中では使用頻度が低くなる単語である．本論文では上記 4 種のラベルを導入することで，生成クラスタに対する多視点融合ラベルを構築する．また，提案システムの評価に関して，クラスタリングの定量的評価および Broder[11] が提案するユーザの検索意図に基づく検索クエリの分類モデルを本研究で提案する検索システム SearchLife に適用し，ユーザビリティ評価を行い本システムの有効性を確認した．

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 Zamir ら[6] や成田ら[8] の研究がある．Zamir らは Suffix Tree と呼ばれるデータ構造から共通して現れる単語を容易に発見し，その単語が出現する検索結果をまとめてクラスタを形成する手法である．成田らは検索結果の階層型排他的クラスタリングを行うシステムとして METAL を開発した．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るという課題がある．ストラクチャマイニングを行う研究として大野ら[7] の研究がある．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検索エンジンとして Clusty[5] 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．本論文では，クラスタ構成は一定のまま，生成クラスタに対して性質の異なるラベリング手法を複数用意することで，ユーザが検索結果から目的とする Web ページ発見の支援を行う．

2.2 Web ページに対する複数スニペットの提示

検索エンジンから返されるスニペットを複数の視点から構築し，ユーザに提示する研究として Jae-wook ら[12] や高見ら[13] の研究がある．Jae-wook 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．Jae-wook らの手法では，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．タスクの干渉度を考慮して，3 種のスニペットを構築し，ユーザが状況に応じてスニペットを選択できる．高見らは，スニペットを，その生成方法により 2 種類の軸で分類している．さらに，Web ページに対して 4 種類のスニペットを生成できるようにすることで，ユーザの検索目的に適したスニペットを提示できるようにした．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．本論文では検索結果クラスタという，ある類似性を持った Web ページ群に対するラベルを，複数の視点から構築できるようにすることを目的とする．

2.3 検索クエリの分類モデル

Broder[11] や Daniel ら[14] はユーザの検索意図に基づく検索クエリの分類モデルを提案した．Broder は検索エンジン利用者の検索クエリを navigational，informational，transactional の 3 種に分類した．また，Daniel らは Broder が提案した informational をさらに階層的に再定義し，また，transactional の代わりに resource を定義した．本研究では Broder が提案した分類モデルに従って，検索タスクを作成し，提案システムのユーザビリティ評価を試みる．

3. 多視点融合ラベルの構築

本節では多視点融合ラベルの構築手法に関して述べる．以下に本論文で提案する 4 種類のラベルの特徴を記す．

ラベル 1 はクラスタ内文書のいくつかに共通して出現し，1 つの名詞からなる．ラベル 1 はラベル 2 と比較して Web 上で使用頻度が高い単語である．ラベル 1 は検索対象においてよく使われる単語の把握やそれらの単語を手掛かりとして必要とする文書を迅速に発見できることが期待される．

ラベル 2 はクラスタ内文書のいくつかに共通して出現し，複数の名詞により構成される．ラベル 2 はラベル 1 と比較して Web 上で使用頻度が低い単語である．ラベル 1 と比較して

Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

Web 上で使われない単語であるため，検索対象に関して情報量が多い場合にはラベル 2 を見ることでより迅速に必要となる文書を発見できることが期待される．

ラベル 3 は各文書より断片的に抽出される単語である．ラベル 3 はラベル 4 と比較して Web 全体集合では使用頻度は低いが，検索クエリに対する検索結果集合中では使用頻度が高くなる単語である．例えば，ある分野に詳しいユーザなら常識的に知っているが，その分野に関して事前に触れたことのないユーザには理解の難しい語句を含む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，検索クエリ集合においてある種「常識的」な単語であると期待されるラベル 3 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事項をユーザが把握するのに役立つ可能性がある．

ラベル 4 は各文書より断片的に抽出される単語である．ラベル 4 はラベル 3 と比較して Web 全体集合では使用頻度は高いが，検索結果集合中では使用頻度が低くなる単語である．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，他のクラスタとの差異を際立たせるのに利用できる可能性が高い．

上記 4 つのラベル生成手法を用いることにより，ユーザは，各自の検索要求や検索対象に関連する事項の理解の程度に応じて，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，閲覧性を向上させることができるのではないかと考えられる．

3.1 ラベル 1 およびラベル 2 の生成

本節ではラベル 1 およびラベル 2 の生成手法について，その概要を示す．なお，ラベル 1 およびラベル 2 の生成手法は本システムにおけるクラスタリング過程と密接に関わる部分があるため，あわせて本節では文献[9] で述べた本システムでのクラスタリング手法についても述べる．本システムの概要を**図 1** に示す．

3.1.1 検索結果の取得および形態素解析

本システムでは最初に，Yahoo!Japan デベロッパーネットワーク[15] 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．

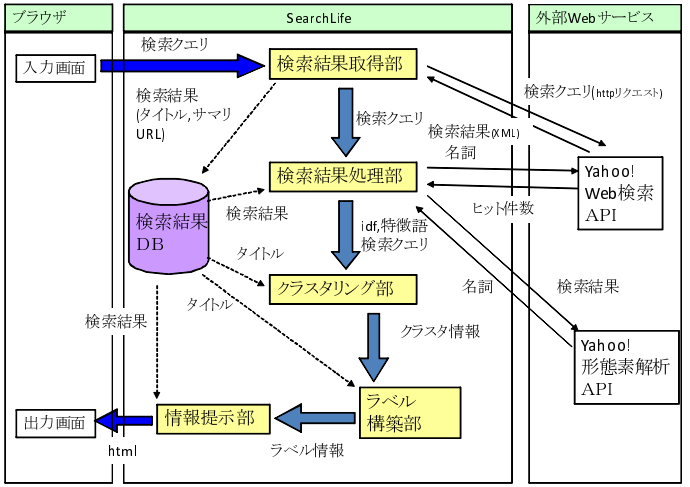

図 1 システム概要

3.1.2 idf 算出およびラベル 2 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．

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

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．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し，これをラベル 2 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．

Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

3.1.3 クラスタリングおよびラベル 1 の抽出

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．

$$tfidf = tf \times idf \tag{2}$$

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，これをラベル 2 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．

手順 (2)　先の手順 (1) 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 3 つ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示順序は手順 (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，これをラベル 1 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．

3.2 ラベル 3 とラベル 4 の生成

3.2.1 条件付き特徴量 idf(t,q)

3.1 節までに生成されたラベル 1 とラベル 2 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，ラベル 3 とラベル 4 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．

$$idf(t,q) = log\frac{dt(q)}{dt(q+t)} \tag{3}$$

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 ラベル 3 とラベル 4 の抽出

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 aveidf(t,q)，aveidf(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．

ラベル 3

$idf(t) > aveidf(t) \cap idf(t,q) < aveidf(t,q)$

ラベル 4 候補語

$idf(t) < aveidf(t) \cap idf(t,q) > aveidf(t,q)$

手順 (2)　ラベル 4 候補語の絞り込み

手順 (1) の時点では内容的に関係のない単語もラベル 4 候補語として設定されている．そこでラベル 4 候補語が本文の html の body タグ内で使われているときに限りラベル 4 とする．

Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

表 1 5つの検索クエリについての METAL と本システム (SearchLife) の比較

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
METAL	0.287	0.833	42.7	0.685
SearchLife	0.409	0.678	51.0	0.750

4. 提案システムの評価

4.1 クラスタリングの定量的評価

本研究では，文献[9] において検索結果のクラスタリングに関する定量的評価を行った．以下にその実験内容および実験結果の概要を記す．本研究では成田ら[8] 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ラベル 1 の妥当性を評価した．成田ら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 ’に対する本システムと成田らのシステムにおけ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 1** に示す．表 1 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 値に関しては我々の手法のほうが高くなった．

4.2 ラベル 3 およびラベル 4 の定性的評価

本節ではラベル 3 とラベル 4 に関して，定性的な評価を行う．**表 2** に検索クエリ「Java MySQL」に対するラベル 3 とラベル 4 をそれぞれ 5 件づつ示す．まず，検索クエリ「Java MySQL」に対するラベル 3 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,「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,「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，ラベル 3 に関して良好な結果を得ていると言える．次にラベル 4 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と思われるが，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，ラベル 4 として抽出成功している例を挙げる．まず,「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，特定の文書あるいはクラスタの他との違いを際立たせるのに役立つ可能性が高いと考えられる．また,「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他の文書あるいはクラスタとの違いを際立たせるような情報であると言える．次にラベル 4 として抽出失敗している例を挙げる.「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，ラベル 4 には該当しない.「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，内容が他のページと際立って異なるものとはいえない．このようにラベル 4 に関して，単純に検索クエリとラベル 4 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．

表 2 ラベル 3 とラベル 4 の例

ラベル 3	ラベル 4
PostgreSQL	掲示板
JDBC	影響
Tomcat	竹
FreeBSD	道
MACOS	求人情報

4.3 検索モデルに基づくユーザビリティ評価

本実験の目的は，実装したシステムで期待される機能が効果的に働いているかを確認することである．本提案システムは検索結果のクラスタリングを行うことでユーザの目的のページ発見や検索結果の概観把握の支援を目的としており，それらが効果的に機能しているかを確認する．本研究での単語の難易度別に複数の視点でラベリングを行う手法について，ユーザの検索対象に関する知識量の違いに応じた最適なラベルを提示できるかを実験により確認する．本実験では Broder[11] が提案した検索クエリの分類モデルを参考にタスクを作成する．Broder の検索クエリの分類モデルを**表 3** に示す．本実験では分類モデルのうち，Navigational および Informational に関する検索タスクを作成し，実験を行う．

実験の手順は以下の通りである．まず，被験者を提案システムのラベル 1 使用ユーザ，ラベル 2 使用ユーザ，およびリスト表示使用ユーザの 3 つに分類する．それぞれのユーザに

Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

表 3 Broder の検索の種別

種別	説明
Navigational	過去に訪問済みで，明らかに知っているサイトの閲覧．検索理由はサイトの URL を入力するよりも便利であるから，または，URL を覚えていないなどの理由から検索がなされる．
Informational	Web ページを読むことや眺めることで何かを学ぶこと
Transactional	Web ページ上で利用できるリソースを手に入れること

は，**図 2** の左側，図 2 の右側，および**図 3** で示される検索結果画面を提示する．提案システムは画面左にラベル一覧が表示され，ラベルをクリックすると画面右側にそのクラスタ内文書が表示される仕組みになっている．なお，リスト表示は Yahoo!Japan デベロッパーネットワーク[15] が提供するウェブ検索 Web サービスを利用したときの検索結果上位 100 件を順番に 10 件ずつ提示する．ユーザは画面下部のリンクをクリックすることで 10 件ずつブラウジングを行う．リスト表示はランキングアルゴリズムに基づいて Web ページが順番に表示されているが，提案システムであるクラスタリング検索エンジンは内容的に類似するであろうページ群であるクラスタに対する名前つまりラベルが順番に表示される．また，図 2 で示されるように本論文で提案するクラスタリング検索エンジンは生成されたクラスタに対して，複数のラベル付け方法を用意している．また，3 つの検索結果は共通して各文書のタイトル，サマリ，URL が表示されている．次に被験者に対して**表 4**, **表 5** で示されるタスクを提示する．次に実験者が設定した検索クエリに対する処理済みの結果を提示する．再検索は行わず，提示した検索結果内でタスクに対する解答を探す．被験者は「タスク開始」ボタンをクリックして，タスクを開始し，タスクが終了したら画面左の「タスク終了」ボタンをクリックする．評価指標はタスク達成時間，文書クリック回数，全クリック回数である．**表 6**, **表 7**, **表 9** は各被験者の結果を提示し，**表 8** に関しては，それぞれの表示型について 4 人の被験者の平均値，ラベル 1 およびラベル 2 使用ユーザ合計 8 人の平均値を使用表示型クラスタリングとして提示する．**図 4** はタスク 2.1 における各正解ページまでの到達時間を表したグラフである．

4.3.1 実験結果と考察

タスク 1.1 は同名の対象が複数ある場合を想定している．このとき，目標となるページはリスト表示において 3 番目に表示されており，過去に訪問したことのあるユーザであるならば，検索結果のサマリを見ることで目標のサイトであることが確認できるため，タスク達成時間がリスト表示使用被験者のほうが提案システムのラベル 1 使用被験者より短くなっている．提案システムではラベル一覧に目標のサイトを特徴付ける単語が提示されておらず，

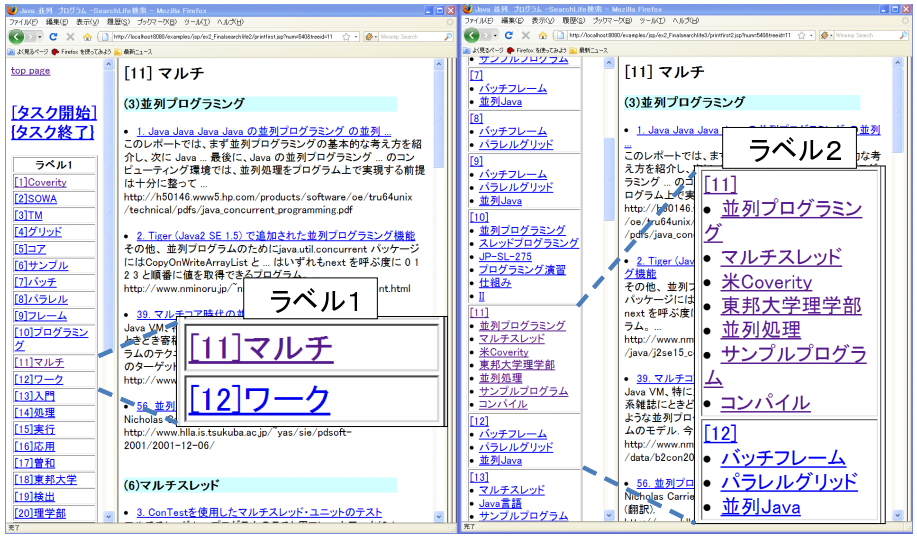

図 2 異なるラベル付け戦略によるクラスタリング結果提示

表 4 設定した Navigational 検索タスク

タスク番号	実験タスク	検索クエリ	タスクの特徴
1.1	静岡大学情報学部石川研究室のサイトを発見せよ	「石川研究室」	同名の対象が複数ある場合の検索タスク
1.2	学会 DEIM2009 の公式サイトを発見せよ	「データ工学　学会」	検索クエリが間接的に検索対象を示すようなタスク

クラスタ内での探索が必要になってしまった．

タスク 1.2 は，ユーザが目標となるサイトを表示させるための重要単語を出力できない場合を想定している．この場合，ユーザは目標となるサイトの関連単語で検索することになる．このとき，提案システムのラベル 1 を使用したユーザはラベル「2009」より目標となるサイトが格納されているクラスタを発見した．これに対してリスト表示において，目標サイトは 31 番目に存在しており，ユーザは結果の線形的閲覧を余儀なくされ，タスク達成時間が長くなっている．

タスク 2.1 は，サマリを見るだけでは正解文書であるかの判断が行い難く，実際にページ内の閲覧を行わないと正解に辿り着けない場合を想定している．図 4 は，タスク 2.1 にお

Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

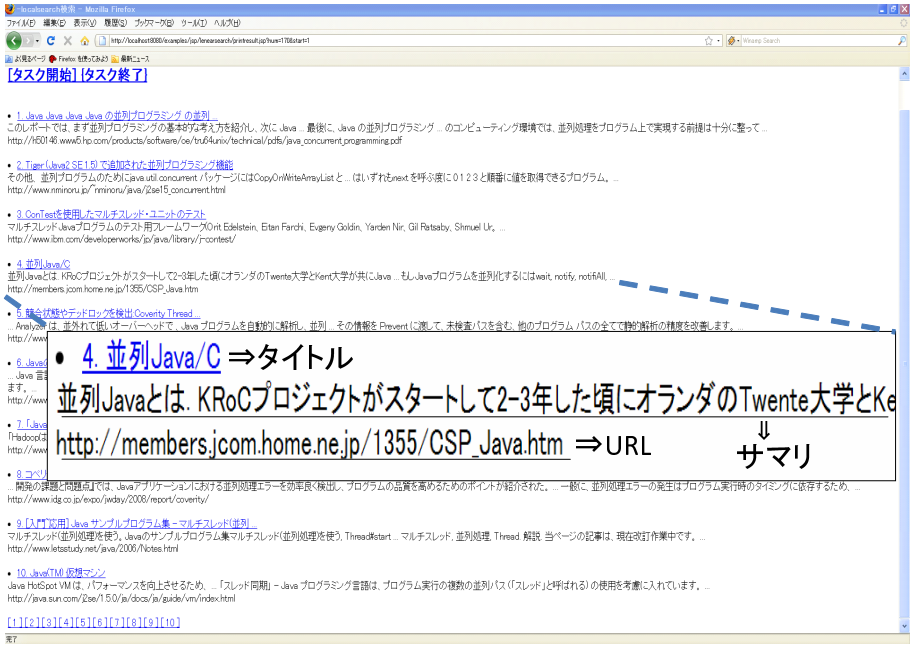

図 3 検索結果のリスト型表示の例

表 5 設定した Informational 検索タスク

タスク番号	実験タスク	検索クエリ	タスクの特徴
2.1	Java の並列処理のプログラムが記述されているページ 5 つ発見せよ	「Java 並列 プログラム」	検索結果のサマリだけでなくサイトの閲覧が必要
2.2	IT に関する記事を 7 つ発見せよ	「サーズ」	検索クエリが多義語である

ける各正解ページまでの到達時間を表したグラフである．本タスクの最終的な達成時間はクラスタリング表示使用被験者とリスト表示被験者の間に，それほどの差は見られないが，4 ページ目を発見するまでの到達時間はクラスタリング表示使用被験者のほうが短くなっている．これは 1 つのクラスタを見ることで複数の正解ページが発見できた結果であると考えられ，本システムの有効性が示されている．本タスクでは使用表示型としてラベル 2 を追加した．ラベル 2 はラベル 1 と比較して Web 上で使われない，つまり，単語の難易度が

表 6 同名の複数対象がある場合のタスク 1.1 の実験結果

被験者	使用表示型	タスク達成時間 (秒)	文書クリック回数	全クリック回数
A	ラベル 1	159	0	16
B	リスト	32	1	1

表 7 検索クエリが検索対象を間接的に示すタスク 1.2 の実験結果

被験者	使用表示型	タスク達成時間 (秒)	文書クリック回数	全クリック回数
A	ラベル 1	20	0	4
B	リスト	41	0	3
C	リスト	114	5	8

表 8 ページ内文書の閲覧を必要とするタスク 2.1 の実験結果

使用表示型	タスク達成時間 (秒)	文書クリック回数	全クリック数
ラベル 1	605	26	49
ラベル 2	507	20	36
クラスタリング	556	23	42
リスト	571	28	38

表 9 多義性を持つ対象を含むタスク 2.2 の実験結果

被験者	使用表示型	タスク達成時間 (秒)	文書クリック回数	全クリック回数
A	ラベル 1	219	0	11
B	ラベル 1	136	7	9
C	リスト	166	2	9
D	リスト	352	0	19

高い可能性がある単語であり，検索対象に関して情報量があるユーザに適していると考えられるラベルである．本タスクのラベル 1 およびラベル 2 使用被験者の合計 8 人は全員 Java の使用経験があり，ラベル 1 使用被験者およびラベル 2 使用被験者の平均 Java 経験歴はそれぞれ 13 カ月と 16 カ月である．ほぼ同様な Java 使用経験を持つ両被験者の結果を比較すると，タスク達成時間および 1 ページ目を除いた各正解ページまでの到達時間はラベル 2 使用被験者のほうが短くなっていることから，検索対象に関して情報量があるユーザにはラベル 2 のほうが扱いやすかったものと考えられる．

タスク 2.2 は，検索クエリが多義性を持つ単語である場合を想定している．この検索クエリの場合，検索結果中には病気の SARS と IT 用語の SAAS に関する記事が含まれており，提案システムのラベル 1 には例えば「SARS」,「感染症」などの病気に関連したラベル

Vol.2009-DBS-148 No.13
Vol.2009-FI-95 No.13
2009/7/28

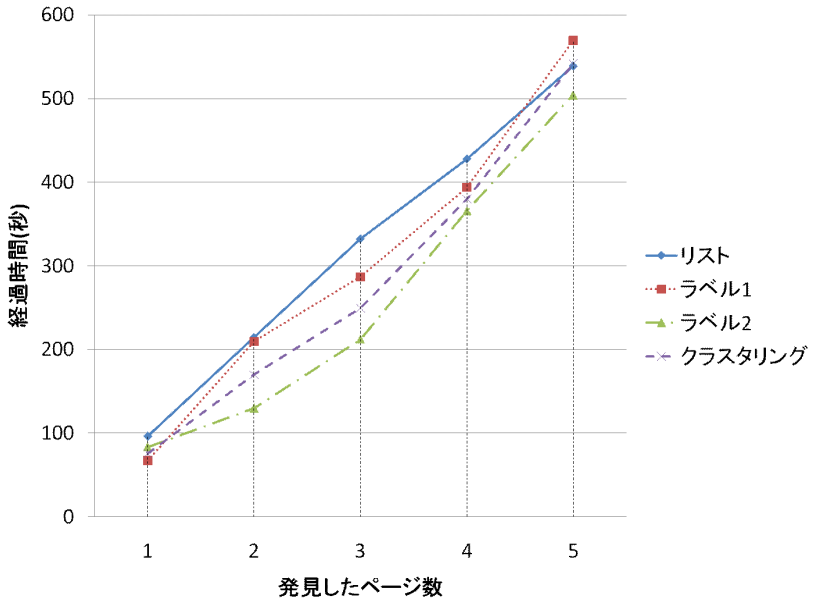

図 4 タスク 2.1 でのページ発見までの経過時間

や「SAAS」,「IT ソリューション」などの IT に関連したラベルとするクラスタが生成された．ラベル 1 使用被験者 B はクラスタラベル「SAAS」を早期に発見したことで，リスト表示使用被験者よりタスク達成時間が短くなったと考えられる．

5. おわりに

本論文では，ユーザが特定タスクのための Web ページ発見の迅速化を目的として検索結果の概観把握が容易となるような検索結果をクラスタリングして提示し，生成クラスタに対して複数の手法でラベリングを行うシステムとして SearchLife を提案した．また，クラスタリングによる検索結果表示に対するユーザビリティ評価を行い，提案システムを用いた場合においてページ収集を有効に行える場合があることを確認できた．今後の課題としてはユーザの状況に応じてクラスタリング手法やラベリング手法を変化させるパーソナライズ機能の検討が挙げられる．

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．

参 考 文 献

1) : Google, `http://www.google.com/`.
2) : Yahoo!, `http://www.yahoo.com`.
3) Ferragina, P. and Gulli, A.: A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering, WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web, ACM, pp.801–810 (2005).
4) Xu, S., Jin, T. and Lau, F.C.: A New visual Search Interface for Web Browsing, *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ACM, pp.152–161 (2009).
5) : Clusty, `http://www.clusty.com/`.
6) Zamirl, O. and Etzioni, O.: Grouper: A Dynamic Clustering Interface to Web Search Results, *WWW '99: Proceedings of the 8th international World Wide Web Conference*, Elsevier North-Holland , Inc., pp.1361–1374 (1999).
7) 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論文誌 (DBSJ Letters)， Vol.4, No.2, pp.13–16 (2005).
8) 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検索エンジンの提案，技術報告，電子情報通信学会技術研究報告 DE2002-61 (2002).
9) 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，電子情報通信学会第 19 回データ工学ワークショップ，B6-3 (2008).
10) 村松亮介，横山昌平，福田直樹，石川 博：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，第 146 回データベースシステム研究発表会（iDB フォーラム 2008），pp.301–306 (2008).
11) Broder, A.: A taxonomy of web search, *SIGIR Forum 36*, Vol.36, ACM, pp.3–10 (2002).
12) wook Ahn, J., Brusilovsky, P., He, D., Grady, J. and Li, Q.: Personalized Web Exploration with Task Models, *Proceeding of the 17th international conference on World Wide Web*, ACM, pp.1–10 (2008).
13) 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論文誌 (DBSJ Letters)， Vol.6, No.2, pp.33–36 (2007).
14) E..Rose, D. and Levinson, D.: Understanding user goal of search, *WWW '03: Proceedings of the 13th international conference on World Wide Web*, ACM, pp.13–19 (2004).
15) : Yahoo!Japan デベロッパーネットワーク，`http://developer.yahoo.co.jp/`.