1C3-02

概念の系統的理解を目的とした情報獲得支援システム

Support System of Information Acquiring to Understand Concept Systematically

西原陽子[*1] 砂山渡[*2] 谷内田正彦[*1]
Yoko Nishihara Wataru Sunayama Masahiko Yachida

[*1]大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University

[*2]広島市立大学情報科学部
Faculty of Information Sciences, Hiroshima City University

With the increased availability of search engines, we can obtain more information from the Web. Since a person doesn't have knowledge related to an unfamiliar concept, it is difficult to understand the meaning. Moreover, Web pages given by a search engine are not displayed in the order of their informational difficulty. So, he has to retrieve more easier Web pages again. In this paper, we propose an information acquiring support system for understanding unfamiliar concepts by systematically studying the related knowledge. First, the system obtains Web pages include the inputted keywords describing the concept. Next, it divides them into clusters. Then, it gives values of difficulty to the clusters and sequences them. Finally, it outputs clusters on a plane whose vertical axis denotes difficulty and whose horizontal axis denotes their sequences. Experimental results proved that the system can supports users in understanding unfamiliar concepts after grasping related knowledge.

1. はじめに

検索エンジンが発達して，たいていの情報は Web ページを使って調べられるようになった．知りたい情報を端的に表すキーワードを検索エンジンに入力すると，キーワードに関する Web ページを得ることができ，それを使ってキーワードの意味や関連事項を調べることができる．

だが，ユーザになじみのない概念を理解しようと検索エンジンを用いても，概念を理解することは難しい．例えば人工知能の分野に詳しくない人が「人工知能とはどういうものか？」を知ろうとしても，検索結果から人工知能の何たるかを理解することはできない．それは与えられる Web ページ集合は簡単なもの難しいものが入り混じっており，内容の難易度の順には出力されないためである．Web ページで使われている言葉がユーザになじみのないものばかりであると，内容を理解することができないため，ユーザは内容が分かる Web ページを求めて再検索をしなくてはならない．

また，なじみのない概念を理解しようとすると，目標概念に関連する多くの知識を学習する必要がある．人間は新しいことを学ぶ際には，前提となる知識を帰納的に学習していくことが多い．ところが，検索結果からでは必要な関連知識を明確に知ることはできず，それを理解する順番も知ることはできない．そのため関連知識を学習する効率が悪くなることがある．

そこで本研究では，ユーザが知りたい概念を端的に表すキーワードを入力とし，キーワードを含む Web ページを内容の類似度でクラスタリングした上で，学習の流れが自然となる学習順序と簡単な知識から学習するための情報の難易度を各クラスタに与えることで情報獲得を支援するシステムを提案する．

2. 関連研究と本研究の位置付け

WWW 上での代表的な情報獲得支援システムとしては Google[5] などの検索エンジンがある．日々のニュースは asahi.com[1] などのニュースサイトを使えば知ることができ，目的とする情報に応じて様々な情報獲得支援システムが置かれている．本研究では情報獲得分野の中の概念理解を対象とし，概念理解を支援するシステムの構築を目的としている．概念理解支援を行うシステムは多くの関連知識の学習から概念を理解させるシステム (AreaView2001[9]) と必要な最小限の知識の学習から概念を理解させるシステム（NaviPlan[11]）の２つに分類できる．AreaView2001 の長所は概念に関連する多くの知識を学習できることであるが，関連知識の学習効率が悪くなる恐れがあることもある．NaviPlan は最小限の知識から概念を理解することができるが，幅広い関連知識の把握までには至りにくいという欠点がある．提案システムはなじみのない概念の理解を支援するものであり，多くの関連知識を順序よく系統的に学習できるシステムが求められる．そこで本研究では，概念の系統的理解を「概念に関連のある知識の意味を順序立てて学習し，その上で目標概念の意味を理解すること」と定義し，多くの関連知識を順序立てて学習することで目標概念を理解できるシステムの構築を目的とする．

提案システムでは関連知識に学習する順序をつけるが，従来研究において最も有名なのは Gagne の学習階層による学習コンテンツの順序付け [4] である．この手法では学習コンテンツをノード，コンテンツ間の順序関係をアークに対応させた有効グラフを作成し，コンテンツの並び替えを支援している．二次元グラフから一次元の学習順序を作る際には，必要な関連知識を帰納的に学習できるような順序を作ることが多い．しかし，帰納的な順序付けを行うと目標概念を理解するまでに長い時間がかかり，学習者は何を学習しようとしていたのかが分からなくなってしまう恐れがある．そこで本研究では学習の過程を話の流れにたとえて，その流れが自然なものとなるような学習順序付けを行うことで，目標概念を念頭に置きながら長い時間をかけずとも関連知識を学習できるようにしている．学習順序付けには，語の流れの自然さを評価し文章作成支援を行っている砂山らの手法 [8] を応用した評価関数を用いている．

さらに提案システムでは簡単な情報を選択できるように，Web ページに情報の難易度を与えている．従来研究では文書中の単語長 [10] や一文の長さから文書に難易度を与えているものが多いが，測っているのは文章の読みにくさであり情報の難易度ではない．そこで本研究では，Web ページに含まれる単語の頻度情報から，Web ページの情報の難易度を測る．

連絡先: 西原陽子，大阪大学大学院基礎工学研究科，豊中市待兼山町１－３，tel: 06-6850-6363，fax: 06-6850-6341，yoko@yachi-lab.sys.es.osaka-u.ac.jp

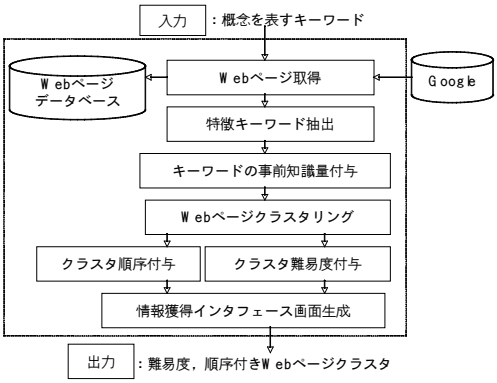

図 1: 情報獲得支援システム構成

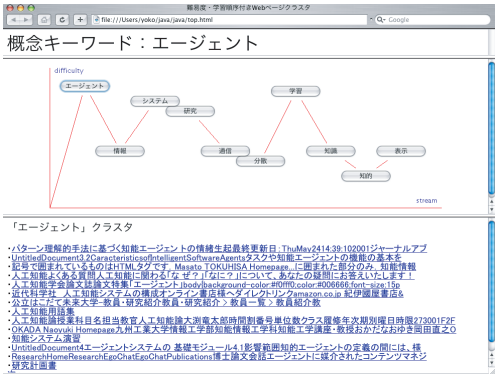

図 2: システム出力：難易度，学習順序付きクラスタ

3. 情報獲得支援システム構成

本章では図 1 に示す情報獲得支援システムの構成を説明する．本システムはユーザがその意味を知りたい概念を示す，名詞の概念キーワードを入力にとる．システムは概念キーワードを検索キーワードとして Google に入力し，検索結果から最大 1,000 件[*1]の Web ページを取得する．次に Web ページからその Web ページを特徴付ける特徴キーワードを抽出し，特徴キーワードに難易度を付与する．続いて，各 Web ページを含むキーワードによってクラスタリングし，クラスタに情報の難しさを示す難易度を付与する．また，学習過程が自然なものとなるようにクラスタに学習順序を与えて，図 2 のような難易度，学習順序付きクラスタを出力する．

3.1 特徴キーワード抽出モジュール

Web ページを特徴つける特徴キーワードの抽出には展望台システム [7] を用いる．展望台システムとは重要文抽出システムの 1 つで，文章中の各文が含む観点語，背景語，特徴語の 3 種類の語にスコアを与え，語の合計スコアの高い文から順に重要文として出力するシステムである．本システムでは Web ページを概念キーワードや関連する語の説明文であると仮定している．概念キーワードや関連する語を観点語とみなし，それに基づき抽出される特徴語を特徴キーワードと見なせるため展望台システムを用いている．展望台システムに Web ページのテキストとユーザが入力した概念キーワードを観点語として与え，抽出される最大２０個の特徴語を特徴キーワードとする．

3.2 特徴キーワード難易度付与モジュール

本研究で扱う Web ページ集合においては，概念キーワードや概念キーワードに関連する，一般的には専門語とされる語ほど出現頻度は高い．そのため出現頻度により特徴キーワードに難易度を与えることは一見不適切な方法に見えるが，概念キーワードによって限られた領域では，頻出している語ほど概念キーワードを理解するためには必要な基礎知識を表すと考えられる．そこで式 (1) で，特徴キーワード t に難易度を与える．

$$specialty(t) = \frac{anum - pnum(t)}{anum} \quad (1)$$

ただし，$pnum(t)$ は t を含む Web ページ数を表し，$anum$ は取得した Web ページ数を表す．

3.3 Web ページクラスタリングモジュール

提案システムで求められるクラスタの性質としては，(１) 目標概念を理解する上で重要な知識が分かること，(２) クラスタ内の Web ページの重複度が高くないことがある．以上の２点を考慮したクラスタリングアルゴリズムを以下に示す．

step0:初期クラスタの作成

取得した Web ページから抽出した観点語をラベルとする空のクラスタを作成し，クラスタ集合 S_c とする

step1:ノイズクラスタの除去

クラスタ集合 S_c 内のクラスタ C の独立率を計算する．独立率とは [2] で提案されている指標で，クラスタの独自性を示す評価値である．独立率が閾値以下[*2]のものは，概念キーワードを理解する上で重要な知識を表すクラスタではないとして削除し，独立率の昇順にクラスタに番号を与える．

step2:クラスタの統合

step1 で与えた番号順に統合するクラスタを選択する．$i < j$ なる任意の２つのクラスタ C_i, C_j 似ていし，以下の（ア）から（ウ）の条件を満たすクラスタ C_j を C_i との統合候補とする．**（ア）C_j 内の Web ページ数が 100 以下，(イ) C_i と C_j の類似度が閾値以上，(ウ) C_i 内の Web ページ数が C_j 内の Web ページ数の 5 倍以下．**ここで，クラスタ C_i, C_j の類似度は式 (2) で測る．

$$sim(Ci, Cj) = \frac{samenum(Ci, Cj)}{Multisize(Ci)} \quad (2)$$

ただし，$samenum(C_i, C_j)$ は 2 つのクラスタ C_i, C_j に共通に含まれる Web ページ数である．上記の条件（ア）では，本システムが取得できる Web ページの最大件数は 1,000 件であることから，その 10%以上の Web ページを含むクラスタを統合すると，関連知識の切り分けができなくなるため設けた．条件（ウ）は Web ページ数に差があるとクラスタ間の類似度を適切に測ることができないと考えたため設けた．

以上，3 つの条件を満たすクラスタ統合候補の中から，式 (3) の値が最大となる C_j を C_i と統合する．

$$uni(C_i, C_j) = Rank(idrate(C_i)) + Rank(sim(C_i, C_j)) \quad (3)$$

*1 Google で表示される検索結果のうちで，URL が取得可能な Web ページが最大 1,000 件であるため

*2 今回は経験上閾値を 0.3 とした．

ただし，$Rank$ は値の順位を返す関数であり，統合されてできた新しいクラスタは，C_i, C_j のラベルをコンマでつないだものをラベルとする．

step4:クラスタリング終了条件の判定

step3 で統合するクラスタが無く，類似度の閾値が 0.35 以上の時は閾値を 0.05 下げ step2 へ戻る．0.35 の時はクラスタリングを終了する*3．また，クラスタ数が初期クラスタ数の 10 分の 1 になった時も終了する．そうでないときは step1 へ戻る．

本モジュールでは得られたクラスタの中から Web ページ数が最も多いクラスタを選択し，その中の Web ページを再度クラスタリングする．システム出力には第二段階のクラスタリングで得られた，Web ページ数の上位 10 クラスタを用いる．

3.4 クラスタ難易度付与モジュール

本研究ではクラスタの難易度を「クラスタラベルや含まれる Web ページによって表される知識を理解するために必要な事前知識量」と定義する．クラスタに含まれる特徴キーワードの種類が多いほど，また含まれる Web ページが難しいほどそのクラスタの理解に必要な事前知識量が多く，難しい情報であると考えられるため，この 2 つからクラスタの難易度を測る．

Web ページ P の難易度は式 (4) で評価し，クラスタ C の難易度は式 (5) で評価する．

$$d(P) = \sum_t specialty(t) p(t|\overline{t_{max}}) \tag{4}$$

$$difficulty(C) = Rank(d(P)_{max}) + Rank(num(C)) \tag{5}$$

ただし，t_{max} は P 中の最大難易度の特徴キーワード，$d(P)_{max}$ はクラスタ内の Web ページの最大難易度を表し，num はクラスタ C 内の特徴キーワードの種類数を表す．

3.5 話の流れによるクラスタ学習順序付与モジュール

本研究では話の流れを「概念に関連する知識を表す複数の語の連続的な出現状態」と定義する．多くの関連語が連続して出現するクラスタ順序ほど，話が分かりやすい自然な流れになっていると考え，クラスタを一次元に並べたクラスタ列において特徴キーワードの出現密度からクラスタ列の話の流れを評価する．式 (6) でクラスタ列における特徴キーワード t の出現密度を測り，式 (7) でクラスタ列 $order$ の話の流れの評価値を算出し，評価値の最も高いクラスタ列をシステム出力とする．

$$continuity(t) = \frac{clusnum(t)}{end(t) - start(t)} \tag{6}$$

$$stream(order) = \sum_t continuity(t) \tag{7}$$

ただし，$end(t)$ は t が出現する最後のクラスタの順番を示す数，$start(t)$ は t が出現する最初のクラスタの順番を示す数，$clusnum(t)$ は t の出現クラスタ頻度を表す．

4. 難易度付与，学習順序付与の評価実験

クラスタ難易度付与と学習順序付与の有効性を調べる予備実験を行った．実験の目的は式 (5) でクラスタに適切な難易度が与えられること，および式 (7) で話の流れの自然な学習順序をクラスタに付与できることを確認することである．

*3 類似度の初期閾値，最終閾値，減らす閾幅は予備実験において得られた最も良い値を設定している

表 1: 回答クラスタ難易度と評価値クラスタ難易度の順序相関

概念キーワード	ケンドール係数	有意水準
形態素解析	0.467	0.079
アルゴリズム	0.333	0.108
検索エンジン	0.467	0.079
フーリエ変換	0.333	0.225
電子署名	0.429	0.022

表 2: 比較システムの名称と説明

比較システム	説明
Google	Google の検索結果
クラスタリング	Google の検索結果をクラスタリング
難易度付与	各クラスタに難易度を付与
順序付与	各クラスタに学習順序を付与
提案システム	各クラスタに難易度と学習順序を付与

4.1 実験条件

難易度付与では，2 つのクラスタから Web ページの内容を理解しやすい方を選択してもらい，学習順序付与ではクラスタ内の Web ページを読み，話の流れの自然な順にクラスタを並べてもらう実験を行った．被験者は理系の大学生，大学院生で，難易度付与ではクラスタ一組につき 3 人，クラスタ学習順序付与ではクラスタセット一組につき 7 人を割り当てた．

概念キーワードは表 1 の左列に示す 5 つの自然科学分野の語で，Web ページは Wikipedia[3] より用意した．Wikipedia は Web 上の百科事典であり，言葉や事物が関連事項も含めて章，節形式で説明されている．Wikipedia の概念キーワードの説明がされている Web ページを章ごとに分け，章のタイトルをラベルとしたクラスタを作成した．

4.2 クラスタ難易度付与モジュールの評価

被験者によるクラスタ難易度と評価式 (5) によるクラスタ難易度の順序をケンドールの順位相関係数 [6] を用いて比較した．表 1 にケンドールの相関係数と有意水準を示す．

表 1 においては「形態素解析」,「検索エンジン」,「電子署名」の 3 つのキーワードで中程度の相関があることが確認でき,「アルゴリズム」,「フーリエ変換」の 2 つのキーワードで弱い相関があることが確認できた．人間が感じる難易度と同じ順序で難易度を付与することはできないが，弱いながらも相関が見られることから，評価関数によって人間が感じる難易度に近い順序でクラスタ難易度を与えられることを確認した．

4.3 クラスタ学習順序付与モジュールの評価

被験者の回答で最も特徴的だったことは，35 人中 28 人がそれぞれの概念キーワードの概要が書かれたクラスタから始まる順序を回答したことである．概要のクラスタを先頭とした理由は，概要のクラスタを読むことで何を理解する必要があるのかと概念の全体像を把握したかったためと考えられる．被験者は自分に足りない知識は何であるかを考え，それをできるだけ効率よく埋めていく順序が自然な話の流れであると考えた．この結果から概念キーワードやその一部をラベルに含むクラスタを先頭に固定した学習順序に式 (7) で評価値を与え，評価値が最も高い学習順序を採用することにした．

続いて，被験者が回答した順序が式 (7) で何位に順位付けされるかを調べた．被験者の 28 人中 21 人によって回答された順序が評価値の上位 20% に入っていた．このことから評価関数によって得られる学習順序は人間にとって自然な，話の分かりやすい流れになっているといえる．

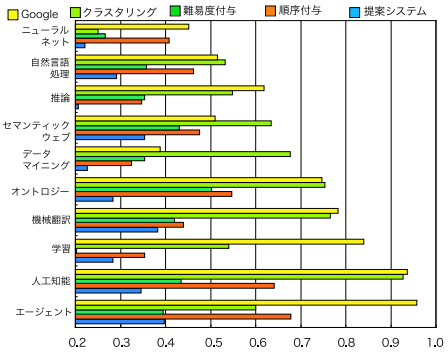

図 3: 読んだ Web ページ数に対する分かりにくかった Web ページ数の割合

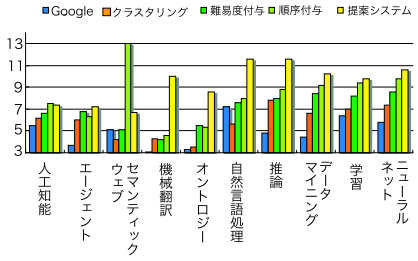

図 4: 被験者から得られた関連キーワードの平均数

5. 情報獲得支援に関する評価実験

提案システムを用いて概念キーワードの意味を理解してもらう実験を行った．概念キーワードは図 3 に示す人工知能に関する 10 個のキーワードである．比較システムは表 2 に示す 4 つを用意し，それぞれ 5 人の理系大学生，大学院生を被験者として割り当てた．被験者には読んだ Web ページ数，内容が分かりにくかった Web ページ数，概念キーワードに関連すると思ったキーワード，理解が深まったキーワード（以下，関連キーワード，理解キーワードと呼ぶ）を回答してもらった．

5.1 実験結果

読んだ Web ページ数に対する分かりにくかった Web ページ数の割合を，システム間で比較した．図 3 を見ると，提案システムにおける割合が最も少なかったことが分かる．このことはクラスタや Web ページに与えた難易度によって被験者は簡単な情報から選択できたことを意味し，提案システムを用いることで概念理解の効率は良くなることが確認できた．

また，図 4 を見ると，提案システムを用いたときに得られた関連キーワード数が最も多かったことが分かる．このことは提案システムが話の流れの自然さから関連知識の学習順序を付けたことで，関連知識が明確になったために，被験者は幅広い関連知識を把握できたことを意味する．したがって，提案システムを用いる方が幅広い関連知識の把握とともに概念を理解できることが分かる．

さらに，図 5 を見ると，提案システムを用いると最も多くの関連キーワードを理解できたことが分かる．このことは難易度付与によって簡単な情報から選択できるようになり，学習順序を付与によって多くの関連知識を把握できるようになったが，双方を組み合わせることで多くの関連知識の意味を理解しやすくなったことを意味する．したがって，提案システムを用いることで情報獲得の効率がよくなり，幅広い関連知識の把握，深い理解とともに概念を理解できることが分かり，提案システムによる情報獲得支援の有効性を確認できた．

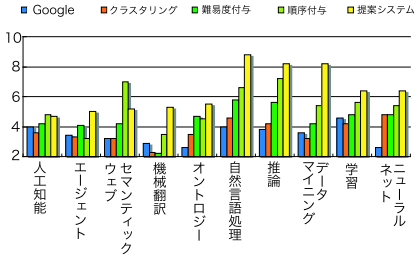

図 5: 被験者から得られた理解キーワードの平均数

6. まとめ

本研究では，ユーザになじみのない概念の理解を支援する情報獲得支援システムを提案した．評価実験により多くの関連知識の把握とともに概念を理解できることを確認した．

参考文献

[1] (URL)http://www.asahi.com/

[2] 井山，砂山，谷内田：多角的な話題の収集を目的とした話題の独自性に基づく Web ページ分類システム，人工知能学会論文誌，Vol.19，No.6，pp.561 - 570（2004）．

[3] (URL)http://ja.wikipedia.org/wiki/メインページ

[4] Gagne：Principle of Instruction Design，Holt Rinehart and Winston，New York（1979）．

[5] (URL) http://www.google.co.jp/

[6] 加藤，石村：Point 統計学 相関係数と回帰直線，東京図書 (2003).

[7] 砂山，谷内田：観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装，人工知能学会論文誌，Vol.17，No.1，pp.14-22，(2002).

[8] 砂山，橘：サブストーリーモデルに基づく文章の流れの抽出，第 164 回自然言語処理研究会資料, pp.153 – 158, (2004).

[9] 平，福島，大澤，伊庭，石塚：AreaView2001:WWW からの構造化した領域総覧提示システム，人工知能学会論文誌，Vol.17，No.3，pp.268 - 275（2002）．

[10] How to Write Plain English：Barnes & Noble Books, New York（1979）．

[11] 山田，大澤：WWW における概念理解のためのナビゲーションプラニング，人工知能学会論文誌，Vol.14，No.6，pp.1125 - 1133（1999）．