情 報 学 基 礎 61－6
自 然 言 語 処 理 142－6
（2001．3．5）

AreaView2001: KeyGraphを用いたWWW構造化システム

平 博司† 大澤幸生††
伊庭斉志† 石塚 満†††

WWWは，その誕生から10年以上経った今なお爆発的なスピードで増えつづけている．このWWWの驚異的な進化は，WWWが基本的にオープンで，誰もが容易に情報発信できるという点によるところが大きい．しかし，オープンであるがゆえにそれぞれのページは分量も掲載目的もまったくばらばらであり，なかにはほとんど意味を持たないページさえあるのが現状である．

本稿では，このWWWの弱構造化システムであるAreaViewの最新版にあたるAreaView2001について紹介する．AreaView2001は，従来のAreaViewシステムをキーワード抽出システム「KeyGraph」を元に根本的に改良したものであり，本システムを利用することによって，ユーザは，自分が知りたいと思うクエリー分野に対して，「あたかも本を読むように」自然にページを読み進められ，その分野に関する大まかな知識を獲得することが出来るようになることを目標としている．これは，「ある分野の知識領域を総観したい」「初めて触れる分野なので、ざっと大雑把な知識を獲得したい」と考えているユーザにとって大いに有用となる．

AreaView2001: A WWW Organization System with KeyGraph Technology

Hiroshi Taira, Yukio Osawa, Hitoshi Iba and Mituru Ishizuka

Information on World Wide Web(WWW) is increasing day by day because of its open characteristics. It becomes difficult for users to find useful information in this huge WWW Information space. Even if the user can find the useful pages fortunately, It is difficult for him to understand all of the space of knowledge which he requests because even the 'useful' page is a huge amount.

In this paper, We propose the system called 'AreaView' which weakly structurizes WWW, and displays it. The latest version of AreaView, AreaView2001, takes out an appropriate keywords from among pages collected based on a certain query, and, as if the chapters of the book, sturucturize pages again using those keywords. When the keywords are extracted or the relations between pages are calculated, AreaView is helped by KeyGraph. This AreaView system quickens understanding to the knowledge field concerning users' query, and users come to understand even the field touched for the first time easily.

1. はじめに

1989年，スイスのジュネーブにあるヨーロッパ粒子物理学研究所（CERN）で，ティム・バーナーズ＝リー氏によって生み出されたWWWは，わずか10年間で約17.8億ページを数えるまでに成長した[1)]．このWWWの驚異的な進化は，ブラウザやネットワーク環境の急速な進化のほかに，WWWが基本的にオープンで誰もが容易に情報発信できるという点によるところが大きいといえるだろう．

しかし，オープンであるがゆえにWWW情報空間は組織的でも構造的でもなくなってしまった．WWWのページはいろいろな経歴，教育，文化，興味，動機を持つ人が，様々な言語，方言，そしてスタイルで書いている．そしてそれぞれのページは分量も掲載目的もまったくばらばらであり，なかにはほとんど意味を持たないページさえある．

この情報の大海原の中から，ユーザが求める情報に関するページを抽出するのが検索エンジン（search engine）である．実際，Infoseekや AltaVista，そしてGoogleなどの検索エンジンは，タグやリンクによる各

† 東京大学新領域創成科学研究科
Graduate School of Frontier Sciences, University of Tokyo
†† 筑波大学経営システム科学専攻
Graduate School of System management, University of Tsukuba
††† 東京大学工学部
School of Engineering, University of Tokyo

ページのランク付けを駆使して，「ユーザのクエリーに最も適した1ページ」をかなり高い精度で見つけてくることが出来るまでになっている．しかしこれらのシステムでは，Yahoo!などのディレクトリサービスと違い，ページ収集からインデクシング，ランキングにいたるまでを全て自動的に行うため，検索結果はおのずと1位から最終位までの単リスト上にならざるを得ず，関連深いページを自然な流れで追ったり，ユーザのクエリーに関する知識分野を概観したりすることは非常に難しい．

本論文では，WWW の構造化システムである AreaView の最新版にあたる AreaView2001 について紹介する．AreaView2001 システムを利用することによって，ユーザは，自分が知りたいと思うクエリー分野に対して，「あたかも本を読むように」自然にページを読み進められ，その分野に関する大まかな知識を獲得することが出来る．これは，「ある分野の知識領域を総観したい」「初めて触れる分野なので，ざっと大雑把な知識を獲得したい」と考えているユーザにとって大いに有用となる．

2. WWW 情報検索

ここで，私たちが常日頃何気なく行っている「WWW ブラウジング行動」についてまず考えてみよう．私たちが WWW でネットサーフィンを行う状況は大きく分けて2つある．1つは「特に何が知りたいというわけでもなく，ポータルサイトや掲示板などを見て回ることで情報の獲得そのものを楽しんでいる状況」であり，2つ目は「知りたい情報があり，それに基づいてネットサーフィンを行っている状況」である．前者に対しては，WebWatcher[3] などの Web ページ推奨ツールがあるが本稿の主旨と外れるのでここでは省略する．後者の状況は，以下のようにさらに二つに細分化される．

- 問題を具体的な言語表現で言語化は出来るが，質問対象の概要はまだよくわかっていない状態（ex.「人工知能のことが知りたい」）
- 問題・および質問対象の概要をはっきり言語化できる状態（ex.「遺伝的プログラミングの J. コーザが書いた最新の論文がほしい」）

前者は情報欲求（information need）の第3段階，後者は第4段階とそれぞれよばれる[4]．

WWW の代表的な検索システムの型であるロボット型検索システム（以下サーチエンジンと呼ぶ）は，情報欲求の第4段階にいるユーザにはほぼピンポイントに該当のページを提供することが出来る．というのも，多くのサーチエンジンは tf/idf[6] やベクトル空間モデル[6] などの手法を用いてシステム構築されており，「遺伝的プログラミング コーザ 論文」など複数のクエリーを入力することで急速に対象ページを絞り込むことが出来るからである．

さらに，近年最も注目されている検索システム Google[7] では，「多くの良質なページからリンクされているページはやはり良質なページである」という再帰的な関係から，被参照リンク関係を用いたページの重要度定義を行っている．この重要度のことを PageRank と呼ぶ．各ページの PageRank 度を表す式を以下に示す。

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \cdots + PR(Tn)/C(Tn)) \quad (1)$$

ここで、
T1･･･Tn:A にリンクを張っている Web ページ
PR(X):ページ X の PageRank
d:定数（Google では主に 0.85 を代入している）
C(X):ページ X から出て行くページリンクの数

この PageRank を 13 億といわれる[8] 収集ページ群のすべてに適用することで，Google は従来のサーチエンジンよりもさらに高いページ提示精度を有している．「NTT」というクエリーならば NTT の公式ホームページが，「笑っていいとも」というクエリーならば，フジテレビ HP にある同番組の公式ページがページリストの一番最初に表示されるなど，ページリスト上位の精度は高い．

このように，「ユーザのクエリー入力に対して最も優れていると思われるページを提示する」ことを最大の目的としているサーチエンジンであるが，情報欲求の第3段階にいるユーザには，時に多大な苦労をもたらすことがある．例えば「人工知能」というクエリーを元に google を用いて検索してみると，数々の研究グループや財団，および大学授業のシラバスや新聞記事などがずらりとならぶ．しかし，「人工知能のことが知りたい」ユーザにとって，特に人工知能「初体験」のユーザにとって重要なのは，多くの場合組織や記事のディテールよりは，人工知能とそのサブクラス（もしくは関連するキーワード）となる「ニューラルネットワーク」や「エキスパートシステム」の大まかな概要を知ることである．ところが現行のサーチエンジンを用いてこれらの事柄を知るには，URL リストに付記されているサマリーから文書の性格を見抜き，人工知能の概要についてまとめてあるページに「当たり」をつけて見に行くという職人芸的なことをしなくてはな

らない．しかも，そのような「親切な」ページでさえ，一著者の知りうる範囲で（すなわち偏った見方で）かかれていることが多く，人工知能の関連分野を幅広く網羅できているかどうかは大いに疑問が残る．

これらをまとめると，「サーチエンジンは大量の検索結果が得られ，ユーザが知りたい情報の対象をはっきりと認識している場合には非常に効果的だが，対象をぼんやりとしか理解していない場合には決して使いやすいものとはいえない」ということになる．

以上の観点から，「サーチエンジンが出力した検索結果をさらに分類整理するシステムがあれば」という思いがよぎるのはごく自然な流れであろう．分類整理の方法論としては，あらかじめ定められたカテゴリ体系へページ群を割り振る「カテゴリ分類」と，互いに類似した内容を動的にグループ化する「クラスタリング」の2つがあり[5]，Xerox PARC における研究[9]ではインタラクティブな情報検索の観点からクラスタリングの適用を主張している．クラスタリングの検索エンジンへの適用例としては NTT の TITAN などがあげられる☆．

これらのクラスタリングはそのほとんどが WWW のサービスとして行われており，そのページデータはもちろんサービスの提供元が収集したものを使用している．しかし，時には自前で収集したデータをクラスタリングし整理したいと思うことがあるだろう．WWW9 国際会議のページを一括して収集しキートピックを元に再整理したい，自動巡回ソフトが集めてきたデータを整理して表示させたい，などが例としてあげられる．通信コストが下がり，検索 Ninja[10] など市販のダウンロードツールが高機能になり，複数のサーチエンジンが提示した良質なページ群を一度に収集できるようになった現在，混ぜん雑多としたページ群を整理したり，それらのページ群から構成されるクエリー知識分野の全体像をつまびらかにしたりするためのツールがますます必要になってくるように思われる．

3. AreaView2001 の概要

そこでわれわれが開発されたのが AreaView2001 である．AreaView は，「WWW の構造化システム」としてわれわれのグループがシリーズで開発しているもので，AreaView2001 はその3世代目にあたる（初代[11]は 1997 年に開発された）．AreaView2001 は収集されたページをユーザが読みやすい形にクラスタリング（このクラスタリングは文書の相似性を用いた既存のものとは異なる手法を用いている）し，各クラスターにラベルをつけて表示する（このラベルを**領域キーワード**と呼称する）．また，クラスタリングによって分けられた各ページに，「子供」にあたるページ群を新たに関連付けることでユーザの各ページの理解を助ける仕組みになっている（この一連のクラスタリングを**階層的構造化**と呼称する）．そしてこのクラスタリングの過程で「主張にこだわる」キーワード抽出システム，KeyGraph[2] を用いている．

ここで，AreaView2001 の特徴について簡単にまとめる．

1) **正式名**： AreaView Ver 3.x
2) **主な対象**： ある知識分野（特に学術分野）にはじめて触れ，この知識分野についての概観となる情報の獲得を目指すユーザ
3) **入力情報**： ユーザのクエリーと，それをもとに Google から収集したページ群（もしくはユーザがすでに収集済みのページ）
4) **出力情報**： ユーザが求めるクエリー分野のページ群を階層的構造化したデータ
5) **主な技術的特長**： 「KeyGraph」「領域キーワード」「階層的構造化」
6) **サービス**： HTML 形式の出力を行う単体のソフトウエア（AreaView Commander）のほか，WWW 上のサービス（Web AreaView），i-mode 用のサービス（ぷちえりあ）として提供．他のサービスへもデータを供給．

以下，KeyGraph，領域キーワード，階層的構造化についてそれぞれ説明を行う．

3.1 KeyGraph

AreaView2001 において，非常に重要な役割を果たしているのがキーワード抽出システム KeyGraph[2] である．以下，KeyGraph の概略について要約して説明する．

「文書の見出し情報や自然言語解析を用いず，単純な頻度だけで重要度を比較しないが，著者の主張を表す語を抜き出すことのできるキーワードの自動抽出」をめざしたものが KeyGraph である．KeyGraph は，**文書は著者独自の考えを主張するために書かれる**という仮説を基にしている．文書全体はその主張を目指して一つの流れを形成するという訳で，文書を建物に喩えると KeyGraph の仮説は

> 建物が立つには，**土台**（文書が基にしている基本概念）が必要である．壁（文章の構成に必要な説明部分），ドアや窓（詳細な記述），

☆ 試験的なものであり現在は公開サービスを行っていない

様々な装飾（比喩や例など，付加的な記述）
もある．しかし，建物の本質は日射や雨から
住人を守る**屋根**（主張点）であって，屋根を
支えるために**柱**（内容の主な展開）がある．

ということになる．例えば学術論文には，冒頭に要点が密集している新聞記事とは異なり，文章が論理的な鎖状に構成されているものも多い．その中には数式やその説明，例証などのまとまりもあるが，その中で繰り返される頻出語（例えば高速アルゴリズムの論文であれば『アルゴリズム』）は要点とは別の，いわばその文章が書かれる上で当然のように前提とされる「土台」の概念を表すことが多い．これらの土台の上に立つ「柱」に支えられて文書全体を束ね方向付けるのが主張（「屋根」）である．この土台・屋根・柱を頼りにキーワードとして取り出すのがKeyGraphの基本戦略である．KeyGraphのアルゴリズムは，次の3フェーズからなる．

1) **土台の形成**： 文書形成の準備あるいは前提となる基本概念（具体的には，後述の語の共起グラフにおいて強く連結しあう語の集まり）を**土台**とする．
2) **屋根の形成**： 1) で取り出した土台たちに強い力で支えられて文章を統合する語を**屋根**とする．
3) **キーワードの抽出**： 土台と屋根を結ぶ強い柱が多く集まった語をキーワードとする．

KeyGraphの大きな特徴は，単一文書のみを扱った（すなわち他の文書との比較を行わない）キーワード抽出にもかかわらず，非常に高い精度を誇っていることにある．実際，KeyGraphで得られたキーワードを，同一ドキュメントを用いてtf/idfで得られたものと比較したところ，再現率・適合率ともにtf/idfを上回っていたという実験結果が出ている[2]．これは，性質の似たページ群の分析を行うためtf/idfを適用しにくいAreaView2001にとって大きな支えとなる．☆

3.2 領域キーワード

AreaView2001がWWW構造化の過程でまず行うのは**領域キーワードの抽出**である．領域キーワードの定義は以下の通りである．

ユーザのクエリーと関連があり，クエリー知
識を理解するのに必要なキーワード群

例えば，ユーザの入力クエリーが「Artificial Intelligence」だとすれば，「Neural Networks」，「Machine Learning」などのキーワード群がこれにあたる．

同様にクエリーに関連するタームを抽出するシステムの例として，Mondou[12]とLycos[13]がある．Mondouは，WWWロボットにより収集したデータをデータベースに蓄積し，ユーザの提示したキーワード集合から，重み付き相関ルールによってキーワード集合を導出するサーチエンジンで検索結果上部にクエリーと関連のあるキーワードが出現する．Lycosでは，過去に人々がどのようなクエリーで検索したかの情報を蓄積しており，ユーザが入力したクエリーが他にどのようなキーワードといっしょに検索されたかを調べて，その「相方」のキーワードを関連のあるキーワードとして提示する．

これに対し，AreaView2001では，先のKeyGraphの結果を用いて領域キーワードの抽出を行う．そのアルゴリズムを以下に示す．

1) **「屋根」キーワードの抽出**： KeyGraphが各WWWページを分析して「屋根」（主張）キーワード群を抽出する．
2) **集計とソート**： 1) で取り出した「屋根」キーワード群を全ページ集計し，頻度ごとにソートする．
3) **領域キーワードの抽出**： 熟語を優先させる形で，頻度上位のものから任意の数，領域キーワード群として抽出する．

「屋根」キーワードはKeyGraphシステムにおいて，そのページの「主張」と判断されたキーワードであり，このアルゴリズムは「あるクエリーに基づいて集められたページ群において，主張の重なり合いの強いものから順にそのクエリーの領域キーワードとする」ことを意味している．このため，抽出される領域キーワードの質は非常に高い．「artificial intelligence」を例にとったときの領域キーワード上位20個を表1に示す．

表1 artificial intelligence 領域キーワード
Table 1 Area Keyword of artificial intelligence

・artificial intelligence	・artificial neural networks
・computer science	・logic programming
・natural language	・international conference
・machine lerning	・multi-agent systems
・cognitive science	・soft computing
・knowledge representation	・common lisp
・artificial life	・research group
・expert system	・distributed artificial intelligence
・fuzzy logic	・common sense
・neural network	・data mining

これらは人工知能の諸本に章の名前として登場してくるものばかりであり，領域キーワード抽出精度の高さを裏付けている．なお実際に行った比較実験の結果は，第5章で述べる．

☆ ただしKeyGraphの現バージョンは，英語の文書のみが対象である

3.3 階層的構造化

AreaView2001 の構造化は，「文書集合のグループ化」という意味においては広義のクラスタリングになるだろうが，その手法は，従来の文書間距離を用いたもの [14] とは大きく異なる．階層化は大きく以下の 3 つのプロセスを経て行われる．

(1) 領域キーワードの構成
(2) 領域キーワードに対するページの構造化
(3) 上位ページに対する下位ページの構造化

[領域キーワードの構成] 領域キーワード抽出プロセスで抽出された各キーワードをユーザの指定に応じた数だけ（デフォルトでは 20 個）取り出す．このキーワード群が各ページクラスタのラベルとなり，階層的構造化における第 1 階層として配置される．

[領域キーワードに対するページの構造化] 領域キーワードを構成したあと，各キーワードを「屋根」（主張）キーワードとするページ群がユーザの指定に応じた数だけ（デフォルトでは 8 個）選ばれる．ページは，ページの重要度順，もしくはページ内における該当領域キーワードの重要度順のいずれかの順序（ユーザが任意に指定できる）でソートされ抽出される．なお，この際 1 つのページが複数の領域キーワードに属してもかまわない．こうして取り出されたページは，階層的構造化における第 2 階層として配置される．

[上位ページに対する下位ページの構造化] 最後に前プロセスで配置されたページ群の下に「子」ともいうべきページ群をユーザの指定に応じた数だけ（デフォルトでは 3 個）配置する．「子」となるページは，以下のアルゴリズムで選択される．

(1) 親ページの「土台」キーワード群と子ページの「屋根」キーワード群を比較する
(2) 両キーワード群の重なり合いが大きいものの順に選択する

すなわち，各親ページの下には，その親ページの「土台」キーワード群と相似度の高い「屋根」キーワード群を持つページ群が配置されることになる．この，親と子の関係は「親ページの基礎概念を子ページで読み深める」という意味合いを持つものであり，新規性の高い構造化となっている．こうして取り出されたページは，階層的構造化における第 3 階層として配置される．

こうして出来た階層的構造体の全体像を図 1 に示す．ここで，クエリーと第 1，2 階層の関係は，本で例えれば「本のタイトル」－「本の各章の見出し」－「本の各章の具体的な内容」ということになり，AreaView2001 の目的と合致している．また，領域キーワードを追っていくだけでもクエリーに関する知識分野の「外延」に触れることが出来るようになっている．

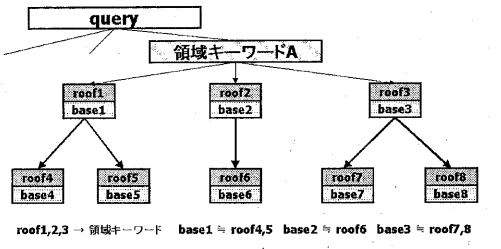

図 1 階層的構造化
Fig. 1 a Hierarchical Organization

4. AreaView2001 システムの構成

AreaView2001 の処理系は全部で 3 つのパートに分かれている．

1) **基礎フェーズ**： ユーザのクエリーを受け取り，KeyGraph にデータを送るまでの前処理を行う
2) **KeyGraph フェーズ**： KeyGraph で処理を行う
3) **構築フェーズ**： 領域キーワードの抽出や階層的構造化を行うフェーズ

以下，それぞれのフェーズについて説明する．

4.1 基礎フェーズ

基礎フェーズでは，まずユーザのクエリーを受け取って**ページの収集**を開始する．ページは商用の検索サービスを介する形で収集する（デフォルトは Google）．ただし，すでにページデータが存在している場合（ダウンロードソフトで複数の検索サイトからダウンロードしてきている場合など）は，この作業は割愛する．

次に**ページの解析とフィルタリング**を行う．まず，収集しているページの URL とタイトルを解析して，データテーブルとしてファイルに保存する．その後，「過度に大きな（小さな）ページ」および「インデックス的・リンク集的ページ」を削除する．このようなページをフィルタリングする理由は 2 つある．(1) KeyGraph は，そのページの主張となるようなキーワード群を抽出するシステムであり，主張がないようなページに適用する意味はまったくない．過度に小さなページや，インデックス的・リンク集的ページは，基本的に単語の羅列に終始しているものが多く主張を取り出すのは非常に難しいため削除される．(2) ユーザが自然にページ群を読み進めていくことを考えると，突然リンク集が出てきたり過度に大きなページが出現することは好ましいこととはいえない．このため，これらのページはふるい落とされることになる．

具体的には，「150word 以下の文書と 3000word 以上の文書」および「全単語中の 30%以上をアンカーテキストが占めている場合☆」，これらのページ群を棄却することにする．

その後，収集した WWW ページからタグや HTML 特有の表現，および冠詞や接続詞などの stop word を取り除いて KeyGraph フェーズに処理を渡す．

4.2 KeyGraph フェーズ

KeyGraph フェーズでは，第 3.1 章で述べた手法を用いて「屋根」「土台」キーワードの抽出を行う．各キーワードはそれぞれ 30 個ずつ取り出されデータベースに保存される．

4.3 構築フェーズ

構築フェーズでは，まず取り出された全ページの「屋根」「土台」キーワード群から単語辞書を作成する．その後，前述の「領域キーワード抽出」および「階層的構造化」を行う．

4.4 各種サービス

AreaView2001 システムは，構造化の「手法」であり，この手法を実際に実装した様々なサービスがすでに展開されている．**AreaView Commander** は，AreaView2001 の機能を忠実に実現した perl のコマンドラインスクリプトであり，HTML の形式でユーザに構造化結果の提供を行う．Perl5 と必要なモジュールが動く環境であれば，どの計算機でも動作可能である．この AreaView Commander の出力結果をもとに WWW でサービス開始しているのが**Web AreaView** である．動作画面を図 2 に示す．

現在は，残念ながら出来合いのクエリーで動くようにしか作られていないが，将来的にはクローラー走らせてページを収集し，リアルタイムに任意のクエリーに対して構造化を行うシステムを製作したいと考えている．また，Web AreaView の姉妹版で i-mode で見やすいようにページ構成しなおしたサービス「**ぷちえりあ**」もある．

そのほかにも，AreaView2001 の構造化データを使った例として AreaBook システム [16]，CCR システム [17] などがある．

5. 評価実験と考察

ここで，AreaView Commander Ver1.10 の出力結果を元に評価実験を行った．パラメータは，領域キーワード数 20，第 1 階層ページ 8 個，第 2 階層ページ 3 個とした（すべてデフォルト）．

☆ WWW においてリンク集的・インデックス的ページはアンカー比率を用いることで高い確率で判別することが出来ると予想されている [15]

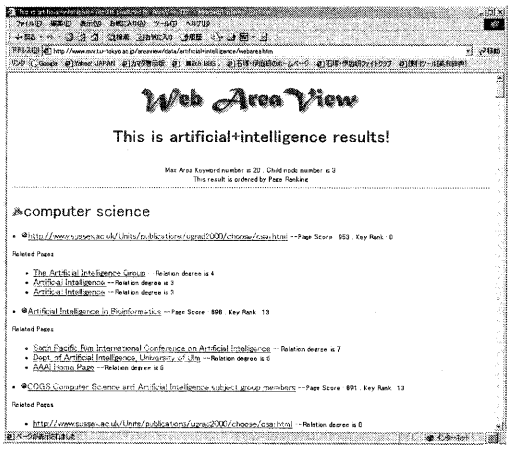

図 2 Web AreaView 動作画面
Fig. 2 the interface of Web AreaView

5.1 領域キーワード比較実験

AreaView2001 の構造化において，領域キーワードは最も重要なファクターであり，その精度のよしあしがシステムの性能を大きく左右するといっても過言ではない．そこで，出力された領域キーワードを，同様に関連キーワードを表示するシステムである Mondou, Lycos と比較する実験を行った．まず，クエリー「physics」を用いて 3 システムのキーワードを比較してみた結果☆☆を表 2〜4 に示す．

表 2 Mondou の関連キーワード
Table 2 Related keyword produced by Mondou

・物理学	・ap
・物理	・web
・server	・brown
・department	

表 3 Lycos の関連キーワード
Table 3 Related keyword produced by Lycos

・science	・math
・biology	・phisics
・chemistry	・gravity
・astronomy	・light

表 2〜4 を比較してみると，「物理」のサブクラスがきちんと表現されているなど AreaView Commander(AreaView2001) のキーワード抽出精度が格段に

☆☆ 3 システムは目的および使用可能言語範囲が完全に同一ではないため参考比較という形になる

表 4　AreaView Commander の領域キーワード
Table 4　Area Keyword produced by AreaView2001

・particle physics	・physics astronomy
・energy physics	・matter physics
・space physics	・plasma physics
・nuclear physics	・computational physics
・phisics society	・medical physics
・physics education	・condensed matter physics
・quantum mechanics	・american physics society
・condensed matter	・theoretical physics
・national laboratory	・american institute
・physics news	・physics university

高いことが定性的にわかる．

また，学術分野を中心に選んだ 20 クエリーを元に上記の 3 システムでキーワード抽出を行い，これを筆者とチェックのための被験者 1 名により 3 段階評価したところ表 5 のような結果を得た．

表 5　20 クエリーによる関連キーワード比較
Table 5　Comparison Related Keywords

	all	ave	Lev.3	Lev.2	Lev.1	score
Mondou	144	7.2	15	17	112	10.4
Lycos	112	5.6	52	30	40	46.4
AreaView	400	20.0	274	59	67	68.5

ここで all は全関連キーワード数，ave は 1 クエリー当たりの平均関連キーワード数，Lev.3〜1 はキーワードの精度を被験者が評価したもの（数値が高いほうが，よりクエリーに関連のあるキーワードである），score は Lev.3 のキーワードの全キーワード数に対する割合を示す．表 5 を見ると，AreaView Commander が他のシステムに比べて多くの，かつ精度の高い関連（領域）キーワードを抽出していることがわかる☆．これはすなわち，あるクエリーに関する知識領域を広く正確にカバーしていることを意味している．この実験により，AreaView2001 のキーワード抽出は実用的なレベルの精度を有していることがわかった．

5.2　領域理解に関する評価実験

また 10 人の被験者☆☆を対象に，AreaView2001 を用いてクエリーに関する知識領域をどれくらい広く正確に学ぶことが出来るかについての実験を行った．実験内容は，4 つの学術的クエリー（ユーザにとってなじみの薄い分野）を用いて階層的構造化を行い，その構造化データ（HTML）を元に各ユーザに大まかなことがわかるまで学習してもらい，各分野の全体像と学習にかかった時間をアンケート形式で答えてもらうものである．

普通，「全体像」とたずねられれば，「クエリー」とは何かを端的に述べたあと，そのクエリーのサブクラスや関連分野について名前を挙げ，それぞれについてコメントするという形式が一般的であり，実際多くの解答がそのフォームに近い形で作られている．ここで，本システムの有効性を示すいくつかの回答がある．

例えば，「oceanography」に関する全体像であるが，10 人の被験者中 7 人が，

> 「oceanography」は「海洋上の物質や生物，気象や現象などを扱う学問」である．そしてそのサブクラスには「Physical Oceanography」「Biological Oceanography」「Chemical Oceanography」「Geological Oceanography」がある．

と記述し，各サブクラスの具体像についてコメントしている☆☆☆．これは「Oceanography」に関するきわめて正確な分析である [18]．

また，「chemistry」に関する全体像においても，

> 「chemistry」は化学物質を扱う分野であり，そのサブクラスの例として「organic chemistry」「inorganic chemistry」「physical chemistry」「biological chemistry」「analytical chemistry」「environmental chemistry」「polymor chemistry」がある

というスタイルのコメントを，多くの被験者が行っている（2 人がサブクラス例のすべてを，8 人が例の 3 つ以上をあげていた）．

実は，このサブクラス名は，そのほとんどが AreaView2001 の領域キーワードとして抽出されていたものであり，ユーザはまず「本の章立て」ともいうべき各分野の領域キーワードを眺めてその分野の全体的な雰囲気を把握し，それから具体的に各ページを読み始めていったと思われる（実験後の被験者へのヒアリングで同様の回答を得ている）．これは，われわれが専門書や入門書を読むのと同じスタンスであり，AreaView2001 システムが当初の目標をある程度達成している証例といえる．なおこれらの学習にかかった時間は，その大部分が 30 分から 1 時間であり，コンパクトな時間で効率的な学習を達成している事がわかる．

また，同じ 10 人の被験者に対して，使用後感として本を用いた学習，および既存の検索システムを用いた学習と比較しての長所と短所をアンケート形式で答えてもらった．その結果をまとめると，以下のようになる．

☆ 前述のようにシステムそのものの優劣を競ったものではもちろんない

☆☆ 大学院生 9 人，大学 4 年生 1 人，全員男性で情報学を専攻

☆☆☆ もちろん具体的な文言は各個人によって多少の差異がある

AreaView2001 は，書籍ほどではないが既存の検索システムに比べてある知識分野の全体像を把握するのに適しており，それによって効率的なネットサーフィンを行うことが出来る．また WWW 上の情報であるため，書籍と違って情報は新鮮であり，本の購入コストや探索時間コストもかからない．しかし，自前でページを持っているときだけでなくただあるクエリーについてまとめてほしいというときにもその都度ページをダウンロードしてこなければならず総合的にはかなりの時間かかってしまう．また構造化に一部ミスがある．

この後半部分に挙げられている問題点を解決する手段としては，「サーチエンジンと連携することで（サーバ側にページデータを収集しておくことで）ユーザのページ収集負担をなくす」「HTML タグを加味したり，被リンク解析を多少加えることで構造化の精度を改善させる」などの方法が考えられ，現在改良中である．

5.3 AreaView2001 の処理時間

AreaView Commander を例にとり，AreaView2001 の処理時間を考察してみる．実験に使用したマシンは OS が Vine Linux 2.0，CPU が Athlon 1GHz，メモリーが 384MB である．なお，ダウンロード部分は通信速度によって結果が大きく異なるのでここでは含めない．

10 クエリーについて構造化を行った結果（平均ページ数 780 ページ），かかった平均時間は約 1 分 6 秒であった．処理時間のボトルネックとなっているのは KeyGraph フェーズと階層的構造化の処理過程であり，この 2 つで全体の約 75.8 %を占めている．これは，今後データ構造の最適化などを施すことで，さらに短縮することが可能と思われる．なお，AreaView Commander では，処理途中に「KeyGraph process...」などのメッセージを出力することで，ユーザのストレス軽減を図っており，待ち時間をさほど長く感じないようになっている．

6. おわりに

本稿では，ユーザが自分が知りたいと思うクエリー分野に対して，「あたかも本を読むように」自然にページを読み進められ，その分野に関する大まかな知識を獲得することが出来るようになることを目標とした WWW 構造化システム AreaView2001 について説明を行った．今後は問題点を是正し，よりユーザが使いやすいシステムを目指して改良を続けていく予定である．

参考文献

1) 山名早人: 検索エンジンと高速ページ収集技術, bit, Vol.32, No.12, pp.72-79 (2000)
2) 大澤, Benson, 谷内田: KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出, 電情 D-1, pp.391-400 (1999)
3) R. Armstrong, D. Freitag, T. Joachims, WebWatcher: A learning Apprentics for the World Wide Web, the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, pp.13-18 (1995)
4) R.S. Taylor: Question-negotiation and information seeking in libraries, College & Research Libraries, pp.178-194 (1968)
5) 林良彦，小林喜嗣: WWW 上の検索サービスの技術動向，IPSJ Magazine, Vol.39, No.9 (1998)
6) G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Proceeding & Management*, Vol.24, No.5, pp.513-523 (1988)
7) S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, 7th International WWW Conference (1998)
8) http://www.google.com/
9) M. Hearst: Interfaces for Searching the Web, Scientific American, pp.68-72 (1997)
10) http://www.ifour.co.jp/ninja/kensaku/index.htm
11) 福島伸一，石塚満: WWW 情報空間の弱い構造化とエリアビュー機能，情処第 58 回全大，3P-06 (1999)
12) 河野，長谷川: WWW データ資源検索におけるデータマイニング手法, 情処データベース研報, 第 96-DBS-108-5 巻 (1996)
13) http://www.lycos.com/
14) C.J. Rijsbergen: Further experiments with hierarchic clustering in document retrieval, Information Storage and Retrieval, Vol.10, pp.1-14
15) 小野田，土肥，石塚: ハイパーリンクの意味理解と意味ネットワーク形状への組織化，第 55 回情報処理全国大会, Vol.3, pp.224-225 (1997)
16) 坂田，平，大澤，伊庭，石塚: WWW 情報空間の AreaView におけるオンラインブックの構築，第 62 回情処全大，8X-01 (2000)(発表予定)
17) ギャン，友部，伊庭，石塚: Meta-Structuring of Concept Chemical Representation(CCR), Natural-Language-Like knowledge representation, 第 62 回情処全大, 6M-09 (2000)(発表予定)
18) http://www.nmoc.co.jp/entertainment/lubricants/x-ing/x02/02-03.html