特集論文

AreaView2001:WWWからの構造化した領域総覧提示システム

AreaView2001:A System of Presenting Structured Area Knowledge Extracted from WWW

平 博司 Hiroshi Taira
東京大学大学院 新領域創成科学研究科基盤情報学専攻†1
Graduate School of Frontieer Science, the University of Tokyo
taira@miv.t.u-tokyo.ac.jp

福島 伸一 Shinichi Fukushima
株式会社 NTT データ
NTT DATA Co.
fukushimasn@noanet.nttdata.co.jp

大澤 幸生 Yukio Osawa
筑波大学社会工学系研究科
Graduate School of Systems Management, the University of Tsukuba
osawa@gssm.otsuka.tsukuba.ac.jp, http://www.gssm.otsuka.tsukuba.ac.jp/staff/osawa/

伊庭 斉志 Hitoshi Iba
東京大学大学院 新領域創成科学研究科基盤情報学専攻
Graduate School of Frontieer Science, the University of Tokyo
iba@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~iba/

石塚 満 Mitsuru Ishizuka
東京大学大学院 情報理工学系研究科電子情報学専攻
Graduate School of Information Science and Technology, the University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: areaview, information structuring, www, information organization

Summary

Information on the 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 a user can fortunately find useful pages, it is difficult for him/her to acquire all the aspects or a structured knowledge view regarding his/her query.

In this paper, we describe a system called "AreaView2001", which presents an overall structured view of the queried area together with a set of useful Web pages explaining the area and its subareas. The style of the presentation is similar to book style, consisiting of chapters and sections. When extracting important keywords of the area from collected Web pages, the system employs the method of KeyGraph which can extract keywords of the main topics and underlying basis knowledge of a text document.

AreaView2001 is particularly useful for those users that want to know unfamiliar areas, such as academic areas, since the area knowledge available in the WWW will be presented as a collection of useful Web pages sorted out according to the overall structure of the area. Although the area knowledge to be presented by the system is not so well structured as book chapters because of the full-automatic structuring, it can provide a variety of vivid knowledge not available in books. Some experimental evaluations are given to illustrate the effectiveness of the sysytem.

1. ま え が き

WWW(World Wide Web) は Web ページを単位として，関連する情報同士がハイパーリンクで結ばれた巨大な情報空間を形成している．この WWW 情報空間の大きな特徴は，誰もが容易に情報発信できるというオープンな性格にあると言える．

しかしその開かれた性格ゆえに WWW は玉石混淆で，また構造もはっきりしていない非常に不均質な情報空間になっている．このような WWW 情報空間の中から必要とする情報を見出すためにサーチエンジンは不可欠の機能になっており，これはキーワードを含む Web ページのピンポイント的検索には有効である．しかし，今後 WWW に含まれる情報や知識の，その他の活用法も開発していく必要がある [河野 01, 山田 01]．本稿では，不均質な WWW 情報空間を対象として，指定された領域に関する系統だった構造化された知識を抽出して提示するシステム，AreaView2001 について記す．

近年，学術分野をはじめとして「ボーダレス化」,「異種領域の融合」が急速に進んでおり，身近な例としては人

†1 現在，株式会社東芝

工知能とゲノム・金融・芸術との融合などが挙げられる．しかし，自分が専門としない未知の学術領域の知識を得ようとする場合，その全体像を把握していないと，理解できない場合が多い．いきなり専門論文や，あるテーマに特化した Web ページを見ても，理解できないだけでなく「木を見て森を見ず（情報の視野狭窄）」という結果を招く危険がある．

そこで我々が提案するのが AreaView2001 システムである．AreaView2001 の目的は「ある分野を知りたいと望むユーザに対し，簡単な操作で『入門書』を提供し，ユーザがその分野を理解するのを支援する」ことにある．ここでの入門書とは「ある分野の全体像を紹介し，さらに個々の部門についてできるだけ深くその意味を掘り下げていくもの」（先ほどのことわざを用いれば,「まず森を見せ，その後個々の木を見せる」ということになる）という意味である．これによってユーザはまず該当分野の大まかな知識を獲得した上で応用的な知識に進むことが出来，新しい分野の知識獲得をよりスムーズに行えるのではないか，というのが我々の仮説である．

AreaView2001 では，この「入門書」作成を，ユーザが興味を持つ分野に関連のある Web ページを整理・組織化して提供することで実現しようとしている．ここで WWW を対象とするのは，Web のドキュメント量が莫大である上に，Web 上で日夜新たな概念がさまざまな分野で誕生し，しかもそれまでその分野に無関心だった人も巻き込むように発展するグローバルな性格を持っているからである．

本稿の構成は以下のとおりである．まず第 2 章で現在の一般的な WWW 情報検索とその問題点について指摘する．次に第 3 章で AreaView2001 の概要と基本となる考え方について述べ，第 4 章でそれを具体的にどのようにシステム化したかについて説明をする．そして第 5 章で評価実験と考察を行い，第 6 章でまとめる．

2. WWW 情報検索

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，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にある（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」）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．さらに，近年最も注目されている検索システム Google[Brin 98] 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 と呼び，これを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．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もできる．また tf/idf やベクトル空間モデル [Salton 88] などの手法を用いることによる検索精度の向上も行われている．

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，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ユーザ（ex. 「人工知能のことが知りたい」）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並ぶ．しかし,「人工知能のことが知りたい」ユーザにとって，特に人工知能に馴染みの薄い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，人工知能領域の全体像とそのサブエリア（もしくは関連するキーワード）となる「知識表現」や「機械学習」などの構成要素を知ることである．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）記さ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疑問が残る．

そこでわれわれは，ある領域の概観を知りたいと望むユーザに，概論書あるいは入門書のような形式で，素材である関連 Web ページを整理，組織化して提示することが好ましいと考えた．提示形式として「本」を見習うとすると，これは一般に以下のように構成されている．

1) **表題：** その本にある事柄を短い語句で表したもの
2) **見出し：** 表題の内容を分解して，構成要素を短い語句で表したもの
3) **章：** 見出しの内容について具体的に記述されている文章
4) **節：** 章の内容をさらに細かに表した文章

以上の本の構成に見習い，目指すべきシステムを以下のようにした．

> ユーザのクエリーを**表題**とし，その表題の領域の重要構成要素を抽出して**見出し**を作成する．各見出しを説明するのにふさわしい Web ページ群を**章**として，その章の内容を細かく叙述した Web ページ群を**節**としてそれぞれ配置（Indexing）するようなシステム．

このシステムは，一見すると InfoSort[内野 00] のような自動分類技術や，自動ディレクトリ技術 [幡鎌 98, Tsuda 99] の一種に思われるかもしれない．しかし，前者とはクエリーに対応して知識体系を動的に変化させることがで

きるという点で，また後者とはキーワードではなくページそのものを意味的に階層化させることができるという点で異なる．また，Scatter&Gather[Hearst 95] とは，単に頻度の多い語だけでなく頻度が少なくても重要な語を話題にしている点，および階層的クラスタリングである点において異なっている．

3. AreaView2001 の概要

3·1 概　　要

AreaView は,「不均質な WWW 情報空間の（弱）構造化システム」としてわれわれが開発しているもので，本稿の AreaView2001 はその 3 世代目にあたる（第 1 世代 [福島 99] は 1998 年に開発された）．

前章の目標を実現するため，AreaView2001 には次のような機能が必要とされる．

- クエリーとして与えられた領域に関連する多数の Web ページから，その領域をあらわす重要構成要素を見出して，**見出し**（キーワード）を抽出する．
- 見出し語を説明している優良な Web ページをインデクシングして，**章**として配置する．
- 章の内容をさらに詳しく述べているような優良な Web ページをインデクシングし，**節**として配置する．

まず本システムでは，ユーザのクエリーを元に，複数の検索エンジンから検索結果上位となるページ群を収集する．これらのページ群は，ノイズの少ない優良なものではあるが，前章で述べたように内容的には非整理であり，論文・記事・組織の広報などさまざまな種類を含んだものである．次にこれらのページを解析して，ユーザが知りたいと望んでいる分野を説明するのに必要なキーワード群（**見出し**にあたる）を抽出する（このキーワードを**領域キーワード**と呼ぶことにする）．次に各領域キーワードを主題として記しているページ群（**章**にあたる）を抽出し，各領域キーワードと関連付けを行う．そして，領域キーワードによって分けられた各ページ群に，各々の記述の土台となるサブ項目のページ群（**節**にあたる）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**と呼ぶことにする．この一連のプロセスの過程で文書中からその文書が主張（説明）しようとしている主題となるキーワードを抽出する手法であるキーワード抽出システム KeyGraph[大澤 99] を用いる．

ここで，AreaView2001 の特徴について簡単にまとめる．

1) **主な対象：** 未知であったり馴染みが薄いある（学術）領域に触れ，この領域についての知識/情報を領域の全体像とともに得ようとするユーザ
2) **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）
3) **出力情報：** ユーザが求めるクエリー領域のページ群を本のような形式で整理し，階層的構造化したインデックスページ
4) **主な技術的特徴：** 「KeyGraph」「領域キーワード」「階層的構造化」

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する．

3·2 KeyGraph

AreaView2001 において，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[大澤 99] の手法である．以下，KeyGraph の概略について要約して説明する．

「文書から，見出し情報や自然言語解析を用いず，語の出現頻度と同一センテンス内での語の共起関係から，出現語をノードとする共起グラフを形成し，著者の主張を示す語を表すキーワードの自動抽出」を目指したものが 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の表現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 の仮説は以下のようにいえる．

> 建物が立つには，**土台**（説明の基礎となる基本概念）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），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土台に立脚する**柱**（内容の主な展開）がある．

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) で取り出した複数の土台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．熟語についても連続する 2-3 単語のうち文書内での出現回数と長さが極大なものを選択し抽出する．なお屋根キーワード群はより主張を述べていると思われるものから順にランク付けを行っている．

例えば，CNN.com のアメリカ同時多発テロに関する記事 (2001/12/21 付) を KeyGraph システムにかけた場合，以下のような「屋根」「土台」キーワードを得た（上位 5 つを表示した）．

屋根： terrorist, Bush, White House, September 11, United States

土台： terrorist, organization, asset, issue, announce

KeyGraph の大きな特徴は，単一文書のみを扱った（すなわちコーパスや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，高い精度が得られていることにある．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 [大澤 99]．これは，同じクエリーを元に収集し主張が似通っているページ群内での分析を行うため，tf/idf を適用しにくい AreaView2001 にとって大きな利点となる．

3·3 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，クエリーの領域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，「knowledge representation」，「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キーワードを抽出するシステムの例として，Mondou[河野 96] と Lycos[lycos] 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一緒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，頻度順にソートする．
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所定の数，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．

3·4 階層的構造化

AreaView2001 の階層化[*1]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「見出し」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（検索サービスから収集した場合は出力順），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[*2]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群がひとまとまりで「章」となり，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) ある領域キーワードに関連付けられた親ページの「土台」キーワード群を集計・ソートする
(2) 親ページの「土台」キーワード群と他のページの「屋根」キーワード群を比較する
(3) 両キーワード群の重なり合いが大きいものの順に他ページ中から子ページとなるものを選択する

すなわち，各親ページ群の下には，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子ページ群が配置されることになる．この，親と子の関係は「親ページの基礎となっている基本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，「章-節」の関係を導き出す要となっている[*3]．こうして取り出されたページ群が「節」となり，階層的構造化における第 3 階層として配置される．

ここで，入力クエリーから第 2 階層までの構造化と，第 2 階層－第 3 階層間の構造化は性質が異なるものとなっていることに留意していただきたい．前者は大項目から小

*1 ここでの階層的とはキーワードおよびページの完全なディレクトリ化を意味するものではなく，あくまで本のような構成という意味である
*2 ここでの「キーワードの重要度順」とは，各ページにおける複数の屋根キーワードの中において，該当の「領域キーワード」のランクが高いページから順に，という意味である
*3 「節」ページの精度を高めるため，各「章」ページごとではなく「章」ページ群全体に対して「節」ページを設ける

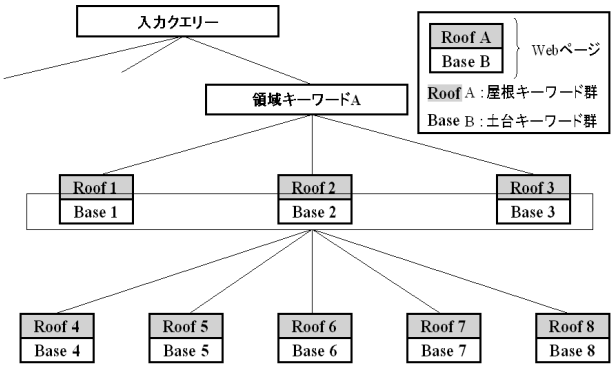


図 1 階層的構造化の様子

項目へと向かう関係となっているのに対し，後者は各項目を詳しく説明していく関係となっているのである．これは，第 1 章で述べたように，アウトプットとして「ある分野の全体像を紹介し（前者），さらに個々の部門についてできるだけ深くその意味を掘り下げていく（後者）」ものを目指しているためである．

こうして出来た階層的構造体の様子を図 2 に示す． Roof は屋根キーワード群，Base は土台キーワード群を表し，この階層化においては，領域キーワード，各 Roof および各 Base に以下のような関係式が成り立つ．

- 領域キーワード A $\subset$ Roof 1$\cdots$ 3
- Base 1$\cdots$ 3 の集計後上位キーワード　　Roof 4$\cdots$ 8

ここでクエリーと第 1，2，3 階層の関係は，本で例えれば「本の表題」－「本の各章の見出し」－「本の各章の具体的な内容」－「各章の基盤となるような内容」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領域知識を概観することが出来る．

4. AreaView2001 システムの具体的な構成

AreaView2001 の処理系は次の 3 フェーズに分かれて構成されている．

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

システムの概略を図 2 に示す． 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利用する形で収集する．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，この作業は不要となる．本システムは，現在のところ英文 Web ページを収集対象にしている．その理由は，日本語ページに比べて格段に量が多く，結果として充実した領域総覧を得ることができるからである．

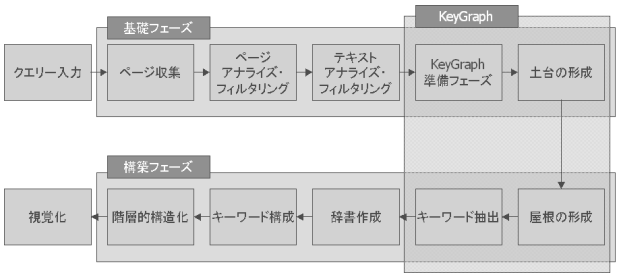


図 2 AreaView2001 システムフロー図

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，あるいは小さなページ」や「インデックス的・リンク集的ページ」といった，主張を持たず，ユーザが自然に読み進めるのを妨げるページを削除する．

具体的には,「150words 以下の文書と 3000words 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[*4]」，これらのページを棄却する．

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 3·2 節で述べた KeyGraph の手法を用いて「屋根」「土台」キーワードの抽出を行う．各キーワードは各ページに対してそれぞれ 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，前章の「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，膨大だが不均質な情報空間の構造化手法であり，この手法を実際に実装した数種のサービスが実現されている．**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．

*4 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 [小野田 97]

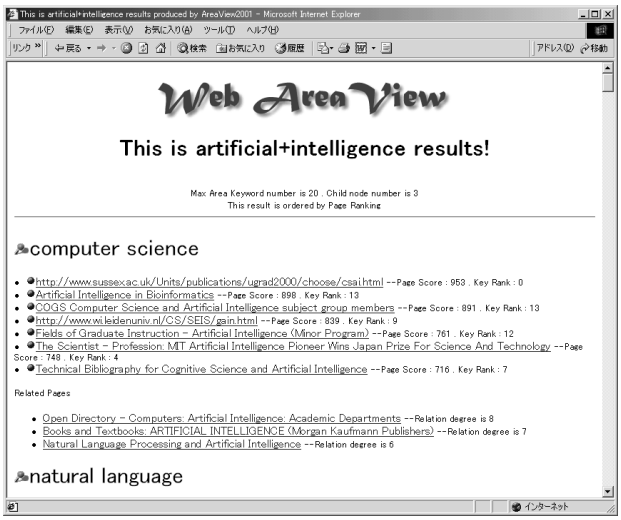


図 3 Web AreaView 動作画面

この AreaView Commander の出力結果をもとに Web サーバのフロントエンドとしてサービス開始しているのが**Web AreaView**[AreaView2001] である．動作画面を図 3 に示す．

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**」[AreaView2001] もある．

そのほかにも，AreaView2001 の構造化データを使って，本の様式で情報を提供する AreaBook システム [坂田 01] がある．

4·5 処 理 時 間

AreaView Commander を例にとり，AreaView2001 の処理時間を示す．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，Web ページ収集のダウンロード時間は通信速度状況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，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．関連 Web ページの収集から処理を行うと，これに 10〜20 分前後の時間がかかることになる．したがって，Web AreaView ではすでに解析，構造化を済ませた領域をリスト化して記録しておき，リストに存在するクエリーの場合はこれを使う，あるいはリストから選択して閲覧できるようにしている．現在，学術領域を中心に約 50 の領域がリスト化され試験的に提供されている [AreaView2001].

5. 評価実験と考察

AreaView Commander の出力結果を元に有効性の評価実験を行った．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）.

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，その精度の善し悪しがシステムの性能を大きく左右する．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 Lycos と比較する実験を行った*5．まず，クエリー「artificial intelligence」を用いて両システムのキーワードを比較してみた結果を表 1〜表 2 に示す.

表 1 Lycos の「artificial intelligence」関連キーワード

・Robotics	・Kerrville
・Tx	・Lake Whitney
・Artificial Intelligence Organization	・Data Mining
・Artificial Intelligence Company	・Robots

表 2 「artificial intelligence」の領域キーワード

・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

表 1〜表 2 を比較してみると，「artificial intelligence」の領域知識を表す語句がかなり良く表現されているなど，AreaView Commander(AreaView2001) のキーワード抽出精度が高いことが定性的にわかる．また，「Chemistry」など 5 つの学術分野をクエリーとして出力を行い，出力結果の領域キーワード (5*20=100 個) をそれぞれの分野を専攻する 5 人の大学院生 (著者グループは含まれていない) に見てもらって正解語をカウントした．この結果 86 語句について「領域キーワードである」との回答を得た.

*5 両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

これに対し Lycos の関連キーワードを用いた同様の実験では，5 つの分野の 28 の関連キーワード*6に対し 9 語句しか実際には関連がないとの回答を得ており，AreaView Commander が別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*7.

なお，本システムの対象外となっている非学術分野のクエリー (ex. PlayStation2, Pokemon) を入力したところ，領域キーワードの抽出精度が悪化した．これは，これらの分野が「Artificial Intelligence」や「Oceanography」などと比べて新規的であり，知識分野が体系だっていないことが主因であると思われる．

5·2 領域理解に関する評価実験

また 10 人の被験者*8を対象に，AreaView2001 を用いてクエリーに関する領域知識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（ユーザにとって馴染みの薄い分野）を用いて関連 Web ページの階層的構造化を行い，その構造化データ（HTML) を元に各ユーザに領域の大まかな概要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ったものである．

普通,「全体像」とたずねられれば，該当の学術分野が何かを端的に述べたあと，その分野のサブドメイン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ろう．本評価実験における多くの回答が以下に例示するようにその形式に近い形で作られたことで，本システムの有効性を示しているといえる．

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*9．これは「oceanography」に関するかなり正確な理解である [酒匂].

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

という形式の回答を，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，6 人が例の 5 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「本の見出し」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様子を把握し，その後に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，AreaView2001 が「まず（サブカテゴリの理解を通して）分野の全体像を把握する」という第 1 ー第 2 階層の目標を、ある程度達成している証例といえる*10. さらに時間的に余裕があった被験者の中には、「海洋生物学 (biological oceanography)」から第 3 階層のページを経て「生物学 (biology)」の分野を読み進めたり,「分析化学（analytical chemistry)」から同じく第 3 階層を経て「分析に関する数学的なページ」へ歩みを進める人もいた．これはサブカテゴリをより深く理解するための「知識の掘り下げ」の作業であり，第 3 階層の目的と合致している．

5·3 考　　察

実験結果を考察すると以下のようになる．

AreaView2001 システムは，Web ページを適切に階層付けてインデクシングすることで，ある分野の「入門書」を作成することを目標としている．この入門書が，ユーザにとって分かりやすく作られているかを測るポイントは 3 つある．すなわち，

(1) いかに適切な見出しが選ばれているか（第 1 階層の精度）
(2) いかに適切に，見出しの内容を理解するのにふさわしいページ群が集められているか（第 2 階層の精度）
(3) いかに適切に，サブカテゴリをより深く理解するためのページ群が集められているか（第 3 階層の精度）

である．このうち第 1 階層と第 2 階層に関しては，それぞれ 5.1，5.2 の各実験によりかなり精度の高い（ユーザが短時間で分野を理解するのに適した）アウトプットを出すことが出来ているのではないかと考えている．

また，第 3 階層に関しても 5.2 の実験により一部のユーザが「知識の掘り下げ」を行うにいたっており，その目標をある程度達成していると思われる．ただ，ノイズページ（サブカテゴリを掘り下げるのにふさわしいとはいえないページ）がある程度含まれているとの指摘も受けた。これは解析対象ページが一部のスコープに限定されている

*6 Lycos では 1 クエリーに対して最高 8 つの関連キーワードまでしか出力されない
*7 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
*8 大学院生 9 人，大学 4 年生 1 人，全員男性で情報学を専攻
*9 もちろん具体的な文言は各個人によって差異がある
*10 なおこれらの学習に使われた時間は，その大部分が 30 分から 1 時間程度である

ことなどが原因と思われ，この第 3 階層の精度を上げていくことが本システムの今後の課題になると考えている．

6. むすび

キーワードを含む Web ページを見出す既存のサーチエンジンとは異なり，クエリー領域を説明するのに有用な Web ページ群を，その領域に関する本のように整理，構造化して提示する AreaView2001 について記した．未知であったり馴染みが薄い（学術）領域を WWW 情報空間からその全体像を把握して，必要な項目を詳細に理解したい場合に特に適している．本のように十分に体系化され整理されたレベルとまではいかないが，Web 上の新鮮で多岐にわたる多くの情報，知識が得られることが利点である．不均質であるが膨大な情報を含む WWW 情報空間の新しい活用法を提供するシステムとなっている．今後，精度の高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 ◇

[AreaView2001] http://www.miv.t.u-tokyo.ac.jp/areaview/index.htm

[米クレバー 99] 米クレバープロジェクト: ハイパーリンクを賢く使う，日経サイエンス，pp.28-35 (1999.9)

[BrightPlanet.com 00] The Deep Web: *Surfacing Hidden Value, White Paper*, http://www.completeplanet.com/Tutorials/DeepWeb/index.asp (2000)

[Brin 98] S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, 7th Int'l WWW Conf. (1998)

[福島 99] 福島伸一，伊庭斉志，石塚満: WWW 情報空間のリンク構造を用いた弱い構造化，信学会，人工知能と知識処理研報告，AI98-93 (1999.3)

[Google] http://www.google.com/

[幡鎌 98] 幡鎌，津田，益岡: ナレッジマネージメントへ向けて－知識検索・整理および基盤技術，人工知能学会誌，Vol.13，No.6，pp.912-919 (1998)

[Hearst 95] Marti A. Hearst, David R. Karger, and Jan O. Pedersen, Scatter/Gather as a Tool for the Navigation of Retrieval Results , Proc. 1995 AAAI Fall Sympo. on Knowledge Navigation (1995)

[河野 96] 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告，96-DBS-108-5 (1996)

[河野 01] 河野浩之: 特集「Web システムにおける情報獲得支援技術」にあたって，人工知能学会誌，Vol.16，No.4 ，p.494 (2001)

[lycos] http://www.lycos.com/

[小野田 97] 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 回情処全大，4Q-11 (1997.9)

[大澤 99] 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)

[Rijsbergen 74] C.J. Rijsbergen: Further Experiments with Hierarchic Clustering in Document Retrieval, Information Storage and Retrieval, Vol.10, pp.1-14 (1974)

[坂田 01] 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 (2001.3)

[酒匂] http://www.nmoc.co.jp/entertainment/lubricants/x-ing/x02/02-03.html

[Salton 88] G. Salton, C. Buckley, Term-weighting Approaches in Automatic Text Retrieval, *Information Proceeding & Management*, Vol.24, No.5, pp.513-523 (1988)

[武田 01] 武田浩一，野美山浩: サイト・アウトライニング-インターネットからの情報収集と可視化技術-，情報処理，Vol.42，No.8，pp.781-786 (2001)

[Tsuda 99] H.Tsuda: WIND:Hyper Keyword Index as a Web Document Directory, IJCAI99 Text Mining Workshop, Stockholm，pp.117-125 (1999)

[内野 00] 内野，宗意，橋本，武智，松井，菊田: ルールベースを用いたテキスト分類サービス-自動分類技術のビジネスへの応用-，INFOSTA シンポジウム (2000)

[山田 01] 山田誠二，村田剛志，北村泰彦: 知的 Web 情報システム，人工知能学会誌，Vol.16，No.4 pp.495-502(2001)

〔担当委員：武田英明〕

2001 年 8 月 30 日　受理

著 者 紹 介

平　博司

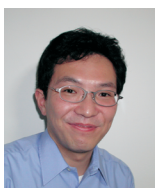
1999 年東京大学工学部電子情報学科卒業．2001 年同大学新領域創成科学研究科基盤情報学専攻修了．同年株式会社東芝に入社．現在同社 SI 技術開発センターに所属．1999 年情報処理学会全国大会奨励賞受賞．

福島　伸一

1997 年東京大学工学部電子情報工学科卒業．同年同大学工学系研究科入学．1999 年同大学院工学系研究科電子情報工学専攻修士課程修了．同年株式会社 NTT データに入社，現在にいたる．ERP パッケージを利用したシステム開発に従事．

大澤　幸生(正会員)

1990 年東京大学工学部卒業．1995 年同大学院博士課程修了，博士（工学）．大阪大学基礎工学部助手を経て 1999 年より筑波大学社会工学系助教授，現在にいたる．2000 年より科学技術振興事業団研究者を兼任，予兆発見研究に従事．情報処理学会，AAAI，IEEE などの会員．1994 年，1999 年人工知能学会全国大会優秀論文賞，1998 年同論文賞受賞．



伊庭　斉志(正会員)

1985 年東京大学理学部情報科学科卒業．1990 年同大学院工学系研究科情報工学専攻博士課程修了，工学博士．同年電子技術総合研究所入所を経て，1998 年より東京大学工学系研究科電子情報工学専攻助教授．2000 年より同新領域創成科学研究科基盤情報学専攻．進化型システムと人工知能基礎の研究に従事．

石塚　満(正会員)

1971 年東京大学工学部電子卒業．1976 年同大学院博士課程修了，工学博士．同年 NTT 横須賀研究所，1978 年東京大学生産技術研究所助教授，同大学工学部電子情報工学科教授を経て，現在，情報理工学系研究科電子情報学専攻教授．研究分野は人工知能，仮説推論，マルチモーダル擬人化インタフェース/コンテンツ，WWW インテリジェンス．IEEE，AAAI，電子情報通信学会，情報処理学会，映像メディア学会，画像電子学会等の会員．