8X-02

AreaView2001: KeyGraphを用いた新しいWWW構造化システム

平 博司* 大澤 幸生† 伊庭 斉志* 石塚 満‡
*東京大学新領域基盤情報学専攻
†筑波大学企業科学専攻
‡東京大学工学系電子工学専攻

1 はじめに

1999年2月に8億ページと推計されたWorld Wide Web(WWW)ページは[Lawrence 99], わずか1.5年後の2000年7月には約33.8億ページと推計されるまでになっており[山名 00], 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, 視覚化システムであるAreaViewの最新版にあたるAreaView2001[平 00]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 AreaView2001の概要

AreaView2001は, 幣研究室 (東京大学石塚研究室) で開発が進められてきたWWW弱構造化システムAreaView[福島 99]の最新版にあたるが, そのシステムの全貌は過去のものと大きく異なっている.

AreaView2001は, **ある知識（特に学術）分野に初めて触れ, 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**を最大のターゲットとしている. このようなユーザにとっては, 知りたいと望むクエリーの事細かな詳細情報や一側面だけを見せるのはユーザの視野を狭めてしまう危険がある. 実際「人工知能」についておおまかに知りたいと望むユーザが, 検索エンジンの検索結果上位にくる「人工生命」や「ロボット」のページ群ばかりを見てこれを人工知能のすべてだと思ってしまい, 比較的順位が後ろの「遺伝的アルゴリズム」や「エキスパートシステム」などのページには触れることさえなかった, などという例はよくある話である.

この問題点を解決するには, 単リストの形で表現される検索エンジンの検索結果をさらに加工して, ユーザが利用しやすい形に変化させる必要がある. 具体的には, 1. 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十分な関連キーワード群を提示し, 2. それを元にページをグループ化し, さらに利用しやすいように構造化する, ことが自動的にできるようなシステムがあれば, ユーザはより幅広く効率的なブラウジングを行うことができるであろう. AreaView2001は, まさにこのようなシステムを目指したものであり, 前述1を「**領域キーワード抽出**」, 2を「**階層的構造化**」と呼称し, この2つをKeyGraph[大澤 99]を用いて実現することで, 「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できるようなシステムを目標としている.

3 KeyGraph

KeyGraphは「主張にこだわるキーワード抽出」として知られ, 「文書は著者独自の考えを主張するために書かれる」という仮設を元にしている. その上で, ノイズを取り除いた文書の中から, その文書の主張となる単語熟語群 (屋根) と, 文書が元にしている基本概念を形成している単語熟語群 (土台) を発見するのがKeyGraphの役割である. AreaView2001では, 「**大多数のWebページ**は著者独自の考えを主張するために書かれている」という仮説をたて, KeyGraphを利用したシステム構築を行っている. なおAreaView2001の処理の過程で, 主張の存在しないリンク集的・インデックス的ページや, 主張が輻輳している規模の大きなページなどは除かれる.

AreaView2001:New WWW Organization System with KeyGraph Technolodgy

Hiroshi Taira, Yukio Osawa, Hitoshi Iba, Mitsuru Ishizuka
Universitiy of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo
taira@miv.t.u-tokyo.ac.jp

4 領域キーワードと階層的構造化

AreaView2001 では, WWW の弱構造化を実現するために, 前述のように「領域キーワード」と「階層的構造化」を用いている. 領域キーワードとは,「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十分な関連キーワード群」を指し, Artificial Intelligence に対する neural networks, expert systems, machine learning などがこれにあたる. AreaView2001 においては, KeyGraph における「屋根」(主張) ワードの重なり合いの度合いの大きいものを領域キーワードとして抽出している.

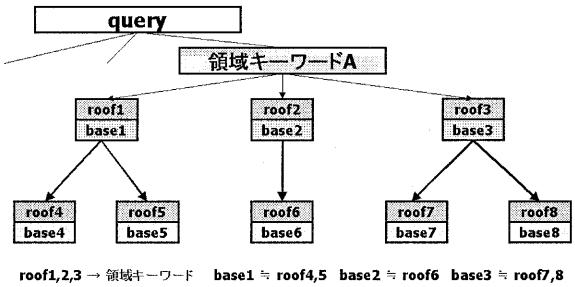

図 1: 階層的構造化

次にこの抽出された各領域キーワードを元に, 収集ページを図 1 のように階層的に構造化する. この構造化は, 以下のルールを用いて行われる.

- 階層的構造化は, KeyGraph によって分析された各ページの屋根 (主張) および土台 (基本概念) キーワードを用いて行われる.
- 領域キーワードを「屋根」(主張) キーワード群の中に持つページ (多くの場合複数) をページのスコアもしくは KeyGraph のスコア順にソートし, 上位のいくつか (ユーザが指定できる) を第 1 段階のページ群として配置する.
- 次にその第 1 段階の各ページの「土台」キーワードを「屋根」として持っているページ群を発見し, これを第 2 段階のページ群として配置する.

この階層化は, ただ似たようなページを単語の相似度を元に構築しようとする従来の手法と違い,「そのページの根幹をなす概念を照会する」という意味合いができ, ユーザの閲覧に大いに有用となる.

5 サービスの全般

AreaView2001 では, 現在ページの収集そのものは Google[Google] を用いて行っている. これについては, 1. 本システムはページのスコアリングや被リンク解析そのものをターゲットにしたものではなく,「構造化」というスコアリングの次の段階をターゲットにしたものである, 2. 経験則的に上位 1000 ページを解析することで非常に広い範囲の領域キーワードを抽出できる, という 2 点から問題とはならない.

AreaView2001 は表現体の部分を分離しており, そのため現在はコマンドラインバージョンの「**AreaView Commander**」, WWW 上のサービスである「**Web AreaView**」, i-mode 用のサービスである「**ぷちえりあ**」を製作しており, 多くのニーズに対応できるようにしている. 2 に Web AreaView の例を示す. また, インタフェースを強化し, より扱いやすいようにした「**AreaBook**」も開発が進められている.

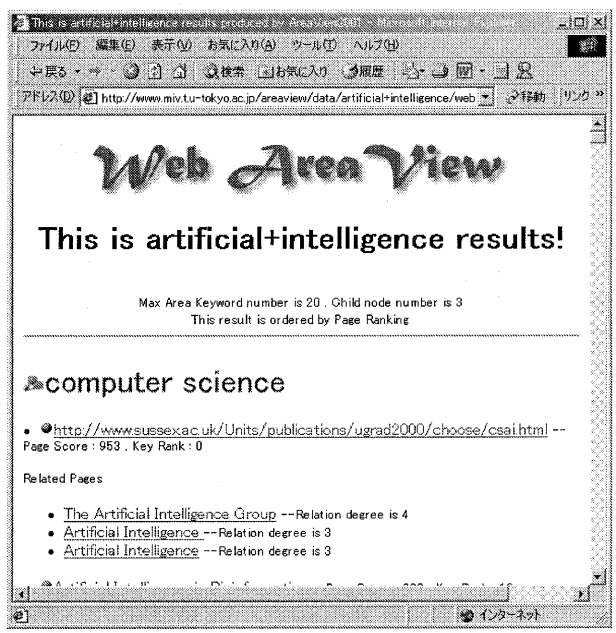

図 2: Web AreaView

これらのサービスは, http://www.miv.t.u-tokyo.ac.jp/areaview/index.htm にて試験的に公開されている.

6 まとめ

本稿では, WWW の弱構造化, 視覚化システムである AreaView2001 について紹介した. 今後は評価実験を重ねて改良を行っていく予定である.

参考文献

[Lawrence 99] Lawrence, Giles: Accessibility of information on the web, Nature, No.400, pp.107-109 (1999)

[山名 00] 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)

[平 00] 平, 大澤, 伊庭, 石塚: KeyGraph を用いた新しい AreaView システム, 第 61 回情処全大, 2P-05 (2000)

[福島 99] 福島, 石塚: 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)

[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)

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