2P-05

KeyGraphを用いた新しいAreaViewシステム

平　博司†　　大澤　幸生‡　　伊庭　斉志†　　石塚　満*

†東京大学基盤情報学専攻 ‡筑波大学企業科学専攻 *東京大学電子工学専攻

1　はじめに

1999年2月に8億ページと推計されたWWWページは[1]、今なお、いや当時よりさらに勢いを増して爆発的なスピードで増えつづけている。これは、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。Web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、様様な言語、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、なかにはまったく意味を持たないページさえある[2]。

そこで弊研究室で開発されたシステムがAreaViewである[3]。AreaViewは、ある分野（たとえば、学術分野の「人工知能」や「分子生物学」など）から利用価値の高いと思われるページ（コアページと呼ぶ）を抽出し、他のページを、これらコアページを中心とするグループに弱構造化して、最後に視覚化して表示するシステムである。

本稿では、このAreaViewシステムを、KeyGraphの考え方を用いて再構築した、新しいAreaViewシステムについて紹介する。

2　旧システムの問題点

これまでのAreaViewシステムでは、ユーザが興味分野の領域総観を行う上での基軸となるページ(コアページ)の選抜を、「被リンク数」とそのリンクの種類のみで行っていた。そのために、実際はその興味分野の基軸とはいいにくいようなページ(ex. 新聞社のページ)が「上位のコアページ」としてカウントされてしまい、領域理解に支障をきたす結果となってしまっていた。このため、新システムでは各ページの「意味・内容」を解析した上で領域総観を行うように、システムを再構築することにした。

3　KeyGraph

ここで、本稿において新しいAreaViewシステムが取り入れた考え方であるKeyGraph[4]について説明を行うことにする。

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。文書全体はその主張を目指してひとつの流れを形成するというわけで、文書を建物にたとえるとKeyGraphの仮説は、「建物がたつには、土台（文書が元にしている基本概念）が必要である。壁（文書の構成に必要な説明部分）、ドアや窓（詳細な記述）、さまざまな装飾（比ゆや例などの付加的な部分）もある。しかし、建物の本質は日射や雨から建物を守る屋根（主張点）であって、屋根を支えるために柱（内容の主な展開）が必要になる。」ということになる。

新AreaViewシステムでは、ページの意味的階層化を図るために、このKeyGraphの考え方が重要となる。

4　新AreaViewシステムの特長

現行の検索エンジンは、ユーザが知りたい情報を具体的に把握している場合（ex. ○○氏の1998年の

New AreaView System with KeyGraph Technolodgy

Hiroshi Taira, Yukio Osawa, Hitoshi Iba, Mitsuru Ishizuka

University of Tokyo,7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

e-mail:taira@miv.t.u-tokyo.ac.jp

本稿に加筆・訂正が生じた場合は、http://www.miv.t.u-tokyo.ac.jp/~taira/に再掲載いたします。ぜひご確認ください。

論文について知りたい) は、ダイレクトにそのページに行き着くことができるため非常に有用である。しかし、もしユーザが興味分野の概観をとりあえず知りたい (細やかなことはとりあえずはおいておいて) というような場合などは、検索結果上位にあがってくるページ群の知識分野に偏在性や冗長性があることから、あまり有効でない場合が多い。

これに対し、新 AreaView システムは次のような特長を有している。

知識 Area の網羅 AreaView では、まずユーザの興味分野のページ群から領域を代表する語群を選び出し、それを確実に網羅するようにページを選び出す。このため、ユーザが知りたい知識 Area を幅広く、もれなくカバーすることが可能である。

ページの意味的階層化 AreaView では、各ページにちりばめられている「キーワード」を、そのページの主張を表す「屋根」と、その屋根を支える「土台」に分けて分析する (これは KeyGraph の考え方である)。その上で、あるページの土台となっているキーワードと別のページの屋根となっているキーワードが一致していた場合、それらを互いにつなげ合わせることで意味的な階層関係を構築していく。これによって、従来の明示的なリンク関係では捉えられなかった「意味的なリンク関係」というべきものを発見することができる。

5 システムの構成

AreaView システムの流れは以下のようになる。

1. 準備フェーズ。まずユーザのクエリーを含むページ群を収集する。そして、tf/idf(term frequency/inverse document frequency) 法と df(document frequency) 値を利用して、 領域全体のキーワード群と df スコアを抽出し、また、収集した各ページのキーワード群とそれにページタイトルを加えたものから各ページの「屋根」「土台」(前述) を発見する。
2. 知識 Area カバーフェーズ。領域全体を代表する語群を漏れなくカバーするようにページ群を選び出す。具体的には、領域全体のキーワード群 (process1 で算出) を「屋根」として 1 語以上含むページを選び出し、その上で被リンク数上位のものから順にキーワードをカバーしていく。そして最終的にすべてのキーワードをカバーできるようになるまでページ選出を続けていく。
3. ページ階層化フェーズ。process2 で選ばれた各ページの「土台」となるキーワード群を、「屋根」として持っているページ (しかも、process2 で選ばれていないもの) を見つけ出し、ページの階下に関連付けしていく (図 1)。2 − 3 段程度これを繰り返す。
4. 視覚化フェーズ。process1-process3 で解析された Area の様子を最終的に視覚化して表示する。

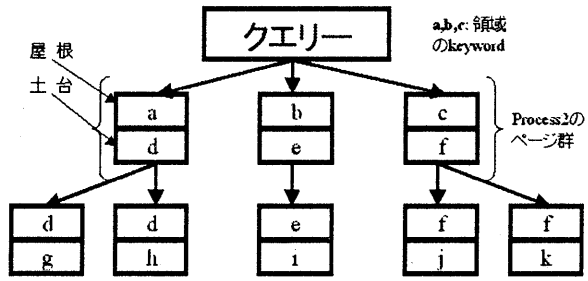

図 1: ページ階層化概要図

6 おわりに

本稿では、AreaView システムを、KeyGraph の考え方を用いて再構築した、新しい AreaView システムについて記した。今後は実際にユーザが簡単な操作で使用可能な、実際のシステムの構築を行おうと考えている。その上で、多くのユーザに利用してもらい評価実験を重ねていきたい。

参考文献

[1] Steve Lawrence, C. Lee Giles: Accessibility of information on the web, Nature, No.400, pp.107-109. 1999

[2] 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

[3] 石塚満: WWW 非均質情報空間のページ意味理解に基づく組織化の研究, 基盤研究 (B)(2) 成果報告書, pp.5. 1999

[4] 大澤幸生, ネルス・E・ベンソン, 谷内田正彦: 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