JAIST Repository
https://dspace.jaist.ac.jp/

Title	Webページ閲覧時の気付きを支援する手法に関する研究
Author(s)	宮田，諭
Citation	
Issue Date	2007-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/3534
Rights	
Description	Supervisor:吉田　武稔，知識科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修　士　論　文

Web ページ閲覧時の気付きを支援する手法に関する研究

指導教官　　吉田武稔　教授

北陸先端科学技術大学院大学
知識科学研究科知識社会システム学専攻

550068　宮田 諭

審査委員：　吉田　武稔　教授（主査）
杉山　公造　教授
佐藤　賢二　　助教授
由井薗　隆也　助教授

2007 年 2 月

Copyright © 2007 by Satoshi Miyata

目　次

第1章　はじめに　1
第2章　閲覧 Web ページからの知識発見　3
2．1　関連研究・・・・・・・・・・・・・・・・・・・・・・・・・・・・3
2．2　Web ページの収集・・・・・・・・・・・・・・・・・・・・・・・5
2．3　気付き・・・・・・・・・・・・・・・・・・・・・・・・・・・・・7
2．4　情報共有による気付き・・・・・・・・・・・・・・・・・・・・10

第3章　システムの実装　11
3．1　システムの構成・・・・・・・・・・・・・・・・・・・・・・・・11
3．2　閲覧 Web ページ収集部 ・・・・・・・・・・・・・・・・・・・・13
3．3　キーワード抽出部および重要文抽出部 ・・・・・・・・・・・・・13
3．3．1　GUI 画面からの形態素解析辞書の用語登録機能・・・・・・・・・14
3．3．2　キーワードの選定・・・・・・・・・・・・・・・・・・・・・・16
3．3．3　絶対頻度と相対頻度・・・・・・・・・・・・・・・・・・・・・19
3．3．4　データの解析間隔・・・・・・・・・・・・・・・・・・・・・・19
3．3．5　重要文抽出・・・・・・・・・・・・・・・・・・・・・・・・・19
3．4　コミュニケーション支援機能　21
3．4．1　キーワードのマッチング機能 ・・・・・・・・・・・・・・・・21
3．4．2　関心キーワード追加機能・・・・・・・・・・・・・・・・・・・21
3．4．3　テキストデータの追加機能 ・・・・・・・・・・・・・・・・・・22
3．4．4　原文抽出機能 ・・・・・・・・・・・・・・・・・・・・・・・・22

第 4 章　評価実験　23
4．1　実験 1　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・23
4．2　実験概要 1　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・24
4．3　評価結果および考察　・・・・・・・・・・・・・・・・・・・・・・・・・24
4．4　実験 2　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・25
4．5　実験概要 2　・・・・・・・・・・・・・・・・・・・・・・・・・・・・・26
4．3　評価結果および考察　・・・・・・・・・・・・・・・・・・・・・・・・・26

第 5 章　まとめと今後の課題　28

謝　辞　29

参考文献　30

図　目　次

図 1　気付きモデル ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・8
図 2　システム環境および構成図・・・・・・・・・・・・・・・・・・・・・・・・・・・・・12
図 3　キーワード抽出部および重要文抽出部の構成図・・・・・・・・・・・・・・・・・・14
図 4　未知語の可視化画面 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・15
図 5　重要キーワードの可視化画面 1・・・・・・・・・・・・・・・・・・・・・・・・・・17
図 6　重要キーワードの可視化画面 2・・・・・・・・・・・・・・・・・・・・・・・・・・18
図 7　重要文の可視化画面 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・20
図 8　関心キーワードの可視化画面 ・・・・・・・・・・・・・・・・・・・・・・・・・・22

表　目　次

表１　システム各部・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・11
表２　評価結果（個人の解析結果を利用）・・・・・・・・・・・・・・・・・・・・・・・・・24
表３　評価結果（他者の解析結果を利用）・・・・・・・・・・・・・・・・・・・・・・・・・26

第 1 章
は じ め に

我々は，インターネットの発展により検索エンジンを用いて Web ページを検索することで，容易に興味のある情報を得ることができるようになった．通常，我々は，興味に沿ったキーワードから検索エンジンが提示した Web ページを閲覧する．そして，閲覧した Web ページ（以下，閲覧 Web ページと呼ぶ）に興味のある情報があれば，その箇所を記憶したり利用したりする．しかし，インターネット上には，類似した膨大な数の Web ページが存在し内容についても玉石混淆である．そのため，閲覧者は，検索エンジンによって提示された Web ページが自分にとって有用な情報であるかどうかを一瞥し，深く閲覧するかどうかを感覚的に判断する．

但し，感覚的に有用な情報は載っていないと判断し，深く閲覧しなかった Web ページにも意識しなかっただけで有用な情報が存在する可能性がある．また，共通する興味や関心に沿って閲覧した Web ページであっても，各 Web ページで取り上げられている意見や主張は各様であるため，それらを蓄積し，解析したデータを比較することで有用な情報を発見することが期待できる．そこで，本研究では，興味を持って閲覧した Web ページの情報に着目し，閲覧 Web ページを再利用することでなんらかの気付きを支援する手法について提案する．

また，本手法を共通の関心や問題意識を持ったメンバー間で利用し，解析結果を共有することで，より有益な情報を得ることができると考えた．なぜなら，関心のあるキーワードが共通であったとしても同様の Web ページを閲覧するとは限らず，異なる観点からそのキーワードに関する Web ページを閲覧しているメンバーが存在すれば，１つのキーワードに関して異なる観点で情報を得ることができ，今まで気付かなかったような知見の発見が期待できるためである．そこで，本研究では，本手法を共

通の関心や問題意識を持ったメンバー間で利用する仕組みについても提案を行った．

堀井ら[1]は, ネットワークに流れる医療情報を解析し医学的特徴を表現することで, 医療ネットワークの監視に役立てようという研究を行っている. 堀井[1]らの研究では, 医療情報の収集方法として，インターネット上に存在する情報を機械的に収集するのではなく，ネットワークの利用者が使用した情報のみをパケットとして取り出し，蓄積する方法を用いている．本研究においても，閲覧された Web ページのみを収集し解析することで，その特徴を抽出し，利用者にとって有益な情報を提示するシステムの開発を行った. 以下, 2 章では関連研究および閲覧 Web ページの情報をどのような方針で利用するかについて述べる．3 章でシステムについて説明を行い，4 章では提案手法を実装したシステムの評価実験および実験結果を示す．最後に，5 章でまとめと今後の課題について述べる.

第　2　章
閲覧 Web ページからの気付き

2.1　関連研究

Web ページには，世界中の人々が持つ知識や意見が埋まっており，これらの情報を利用することは大変有用である．しかし，インターネット上には，膨大な数の Web ページが存在し内容についても玉石混淆である．そのため，Web ページの閲覧履歴から利用者の嗜好を学習し，その嗜好に沿って Web ページを提示することで，利用者の効率的な Web ページの閲覧を支援するシステムの開発[2]や利用者の閲覧 Web ページから取得したキーワードをハイライトすることで Web ページの閲覧を支援する研究[3]など，インターネットに散在する膨大な情報から利用者の閲覧履歴を用いて興味のある情報を効率よく提示する手法についての研究が盛んに行われている[2-5]．本研究においても，閲覧された Web ページを基に利用者にとって有用な情報を提示するという点ではこれらの研究との関連は深い．但し，これらの研究[2,4,5]は，Web ページの閲覧履歴や視聴履歴を基に，閲覧者の好みに沿った情報をインターネットから収集し閲覧者に提示する研究である．本研究では，これらの手法[2,4,5]よって得られる，閲覧者にとって関心の高い情報から，閲覧者が関心を持っているキーワードを可視化することにより，なんらかの気付きを支援することを目指している．

閲覧 Web ページの情報に基づいて可視化を行う研究として，戸川ら[6]は，管理しているネットワークの利用者が閲覧した Web ページから，抽出した名詞語句を各カテゴリに分類し，分類されたカテゴリを可視化することでネットワークの利用動向を把握するためのシステム（WAVISABI: Web-browsing-Activity Visualization System

for Administrator assistance using users Browsing Information）の開発を行っている．但し，WAVISABI は，ネットワークの管理支援を目的としているため，プライバシーの観点からも可視化する情報をカテゴリに限定し，閲覧された Web ページの詳細な情報を可視化する仕組みは取り入れていない．また，WAVISABI は，蓄積された利用者の閲覧 Web ページを解析し，管理しているネットワークではどのような Web ページが閲覧されているのかを把握しようという研究であり，閲覧者にとって有用な情報を提示することを目的としているわけではない点で本研究とは異なる．

インターネット上の情報から気付きを見出すという点で関連する研究として，オンラインコミュニティのテキスト情報を分析することにより，チャンスを発見しようという試み[7,8]がある．松村ら[7,8]は，テキストによるコミュニケーションにおける影響の普及モデル（IDM：Influence Diffusion Model）を提案し，そのモデルを用いて電子掲示板で盛り上がる話題や話題の提供者，語などを見つけることに成功している．この研究は，蓄積されたオンラインコミュニティのテキスト情報における共通の話題から抽出した特徴によって，「盛り上る～」といった限定された内容に関する気付きではあるが，新たな知見をみつけ出すという点において本研究との関連は深い．松村ら[7,8]の研究で用いられている影響の普及モデル IDM は，テキスト情報におけるコメントチェーン（コメントの返信関係の連なり）から 盛り上がる話題や話題の提供者，語の特徴を把握するという大変興味深い手法である．しかし，IDM は，コメントチェーンのあるテキスト情報を対象としており，閲覧 Web ページという漠然とした情報から特徴を把握するのには適さない．

本研究では，閲覧 Web ページに高頻度で出現する語を基に重要キーワードを抽出することで，閲覧 Web ページに共通で出現するキーワードや閲覧時には意識しなかったキーワードから，利用者が閲覧した Web ページの特徴を把握する．そして，重要キーワードを基に重要文抽出を行い，関心を持って閲覧した Web ページで取り上げられる共通の話題や新たな知見など，なんらかの気付きを促すことを支援する．また，本研究では，本手法によって得られたキーワードや情報を共通の関心や問題意識を持ったメンバー間で共有することで，より効果的な気付きの支援やメンバー間のコミュニケーションの活性化，知識共有の促進を目指している．

ウェブコミュニティやグループにおいて知識や情報の共有を支援するシステムの研究は，ナレッジマネジメントや発想支援の観点から盛んに行われており[9-11]，グ

ループのメンバー間で関心を共有することにより，知識の共有を促進させようという基本的な考え方は戸川らの研究[10]に近い．戸川らの研究では，発話から抽出した語と，その語についての共起関係を用いて生成した関心プロファイル[10]と呼ばれるグラフ構造により関心を表現している．また，その関心プロファイルに基づいてメンバー間で関心の類似性が高いメンバーを特定し，会話支援エージェントがそのメンバーの名前や関心に関する話題を提示することで知識共有のきっかけを提供している．本研究との違いは，戸川らの手法[10]が発話から抽出した語に基づいて関心を表現するのに対して，本手法では，関心を持って閲覧した Web ページに出現する語を閲覧者の関心として表現する点である．

また，本研究では，閲覧された Web ページのみを収集し解析することで，その特徴を抽出し，利用者にとって有益な情報を提示する手法の提案を行っているが，解析する情報を閲覧された Web ページに限定するという考え方は，堀井ら[1]の研究に着想を得たものである．堀井[1]らは，ネットワーク通信に含まれる医学的特徴を体系的に把握するための監視結果検証手法の提案を行っており，医療情報の収集方法として，インターネット上に存在する情報を機械的に収集するのではなく，ネットワークで利用されている情報のみを収集し蓄積する方法の提案を行っている．

2.2　Web ページの収集

研究や統計資料の作成を目的として Web ページの収集を行う手段には，自動的に Web ページを収集する方法と手作業により収集する方法がある．手作業によって Web ページの収集を行えば，収集の目的に対して質の高い情報を得ることができるが時間と手間がかかる．また，自動的に Web ページを収集する方法では，手作業により集める情報ほど質の高い情報を得ることは難しい．自動的に Web ページを収集する手段として一般的に利用される技術には，クローラーや閲覧履歴を利用した方法がある[3-6,12]．

クローラーとは，HTML に記載されているリンクを辿りながら Web ページを自動的に収集するソフトウェアである．クローラーによって機械的に Web ページを収集すれば，短時間で大量の情報を集めることができる．また，Google の検索エンジン

が利用していることで知られる PageRank 手法のランキングをクローラーの Web ページの収集基準として用いることで, 重要な Web ページを効率的に収集する研究[12]も行われており，一定の質を保った大量の Web ページを収集する方法としては有効である．つまり，膨大な情報からなんらかの傾向や相関の発見を目指すテキストマイニングであれば，クローラーのような機械的に大量の情報を収集する手法を用いることができる．但し，クローラーは，機械的に Web ページを収集するため，インターネットを用いた情報検索に関する人間の思考や行動を推測する手掛かりとして利用するのには適さない. また, クローラーにより機械的に収集した Web ページの質は, 検索エンジンによって提示された情報から人間が選択して閲覧する Web ページほどの質は期待できない.

例えば，石川県にある温泉を調べるために，Google の検索エンジンを用いて「石川」,「温泉」という 2 つの検索キーワードで Web ページを検索すると，検索結果の上位 10 ページ中 2 ページは，信州にある温泉宿のホームページであった．また，その際に提示されたタイトルのひとつは「信州 松本／浅間温泉 西石川旅館 <長野県松本市> 源泉掛け流しの宿」となっており，容易に石川県にある温泉宿でないと判断できるものであった．つまり，我々は，検索エンジンによって提示された Web ページの中からタイトルや要約文を参考に Web ページを選択するため，閲覧目的に関係のない Web ページを閲覧することはまれであり，クローラーによって収集される Web ページの情報よりも，人が検索エンジンを用いて情報を収集する方が閲覧目的に関連している情報を得られる可能性が高いといえる.

一方，閲覧履歴を基に Web ページを収集する方法には，プロキシサーバのログやキャッシュデータを用いる方法があり，戸川ら[6]は，ネットワークの利用動向の把握を目的として，閲覧 Web ページの収集方法にプロキシサーバのログを用いている．プロキシサーバとは, 内部ネットワークからインターネットへのアクセスの制御を行う機能や内部ネットワークからのアクセス要求があった Web ページをキャッシュする機能などを持つサーバである．一般的に，プロキシサーバとは HTTP プロキシサーバを指し，セキュリティやトラフィックの軽減を目的として利用される．プロキシサーバのログには，内部ネットワークからアクセス要求のあった Web ページの URL やアクセスの要求を出した利用者の IP アドレスが記録されている．つまり，プロキシサーバのログに記録されているアドレスを辿ることで，利用者を特定したり利用者

が閲覧した Web ページを収集したりすることができる．そのため，ネットワークの利用者が目的を持って Web ページを閲覧していると仮定するならば，プロキシサーバにより Web ページを収集することで，閲覧目的に対して質の高い情報を取得することができるのではないかと考えた．そこで，本手法では，クローラーによって Web ページを収集する方法ではなく，プロキシサーバのログを利用することにより，目的に沿って閲覧された Web ページのみを収集する方法を用いることにした．

2.3 気付き

本研究では，閲覧された Web ページを収集し，解析した結果を利用者に提示することで気付きを支援する手法を提案する．通常，我々は，インターネットを用いて関心のある情報を探し出す際，Google 等の検索エンジンを利用する．この際に利用される一般的な検索エンジンの多くは， PageRank 等の手法によって得られた Web ページの重要度や検索キーワードなどを基に，検索者にとって有用な情報が載っているであろう Web ページから順に検索者へ提示する．

このような Web 検索の技術は年々高まっており，検索エンジンによってランキングの上位に提示される情報は，検索キーワードに深く関連した情報である可能性が高く，的確な検索キーワードさえ用いれば，検索者の要求する情報を見つけ出すことも容易となってきている．また，我々は，検索エンジンによって提示された Web ページの上位から順に Web ページを閲覧していくのではなく，提示されたタイトルや要約文を参考として，自分にとって有用である情報が載っていそうな Web ページを選択し閲覧するので閲覧目的に対してノイズが少ない．そのため，検索エンジンを介して共通の関心に沿って閲覧された Web ページの情報を収集すれば，閲覧者にとって関心のある有用な情報が含まれている可能性が高い．そして，閲覧 Web ページに高い頻度で出現するキーワードは，閲覧者の関心や興味に関連している可能性も高いはずである．また，閲覧者にとって興味深い情報やキーワードが閲覧 Web ページに含まれるならば，閲覧時には意識しなかった有用な情報や高頻度で出現するにもかかわらず気付かなかった意外性を持ったキーワードの発見など，なんらかの気付きが期待できるのではないかと考えた．

そこで，本研究では，閲覧 Web ページに出現する語句の頻度に基づいて重要キーワードを抽出し可視化することで, Web ページの閲覧者にとって有用な情報を提示し，気付きを支援するシステムの開発を行った．本システムは，重要キーワードを用いて重要文抽出を行うことで，そのキーワードに関する内容を確認できる仕組みを取り入れている．この仕組みにより，閲覧はしたが読み飛ばしていた有用な情報の発見や意外なキーワードの内容の詳細を確認することが可能となる．また，１つの関心事に沿って閲覧した Web ページであっても，各 Web ページで取り上げられている意見や主張は各様であり，そのキーワードに関する様々な文章を提示すれば，利用者はそれらの情報を基に広い視野で物事を捉えることができ，新たな発想を得るきっかけを提供できるのではないかと考えた．

本手法によって想定する気付きの例について，図１のモデルを用いて説明を行う．図１における円 A，B，C は，Web ページから得られた一定の出現頻度を超えるキーワード群とそのキーワードを含む情報を示し，斜線部分はその情報の内容やキーワードが共通で含まれていることを示しているとする．円 A，B，C の情報が共通の関心事に対してそれぞれ異なる複数の検索キーワードによって閲覧された Web ページで

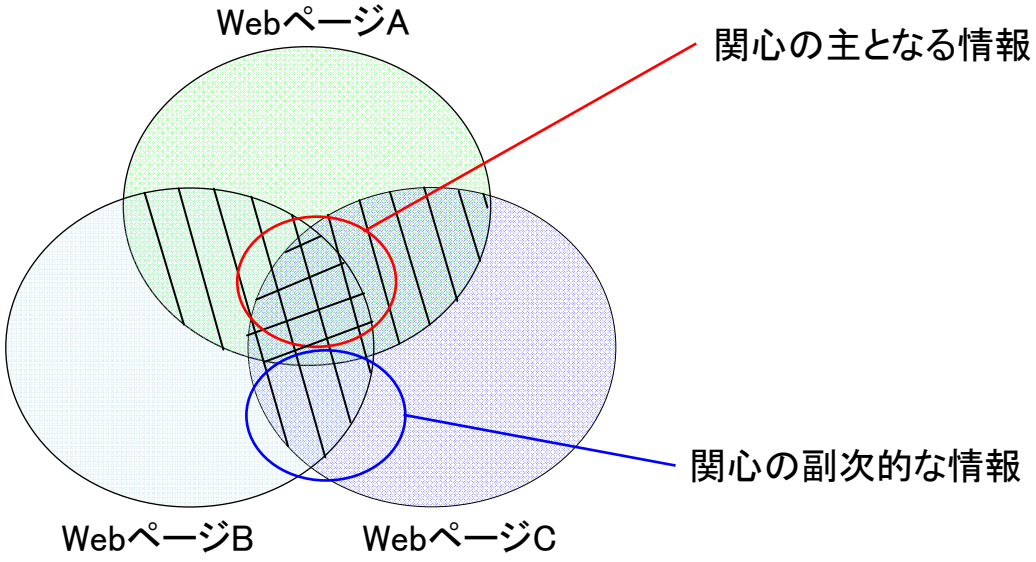

図1．気付きモデル

あるとすれば，3 つの円が重なっているキーワード群には関心の主となるキーワードが含まれる可能性が高く，2 つの円が重なっている部分には，関心事について副次的なキーワードが含まれている可能性が高いことが想像できる．ここで，関心の主となるキーワードとは，関心の対象そのものや関心の対象に深く関わるキーワードを指し，関心事について副次的なキーワードとは，その対象から容易に連想できそうなキーワードや関心の対象を修飾するようなキーワードを示す．

通常，我々は，ある関心事について検索エンジンを用いて調べたい場合，関心の主となるキーワードを検索キーワードとし，必要に応じて副次的なキーワードを組み合わせて情報の検索を行う．関心はあるが欲しい情報が明確でない場合，副次的なキーワードは，その関心事についてどのような情報が欲しいのかを見つけるための大きな手掛かりとなる．つまり，図 1 における 2 つの円が重なる部分に含まれる関心事の副次的なキーワード群を含む情報を閲覧者に提示すれば，それらの情報から閲覧者の要求する情報を明確にするヒントを発見することが期待できる．また，関心の主となるキーワードすら漠然とした関心事について情報を集めているのならば，円 A，B，C のすべてが重なる関心の主となるキーワード群を含む情報から，その関心事について主となるキーワードに気付くことや漠然とした関心事を明確にするための手掛かりを発見することが期待できるのではないかと考えた．

なお，検索エンジンに用いた検索キーワードは，検索者の大きな関心を示しているといえるが，検索者がそのキーワード自体からなんらかの気付きを促されるとは考えられないため，閲覧された Web ページに出現した他のキーワードと同様，出現頻度に基づいて重要度を設定する．本システムには，各メンバーが自分の関心事や関心のあるキーワードについて登録し，そのキーワードを基に重要文抽出を行ったり，関心事をグループのメンバー間で共有したりできる仕組みを取り入れているが，本システムをグループで利用した際，検索キーワードは他のメンバーの関心について知る手掛かりとなるため，検索キーワードの取り扱いについては検討が必要である．

2.4　情報共有による気付き

ナレッジマネジメントや発想支援の観点から，ウェブコミュニティやグループにおいて知識や情報の共有を支援するシステムの研究が盛んに行われている[9-11]．本手法においても，共通の関心や問題意識を持ったメンバー間で各メンバーの閲覧 Web ページの解析結果を共有することでより有益な情報を得ることができると考えた．なぜなら，関心のあるキーワードが共通であっても同様の Web ページを閲覧するとは限らず，異なる観点でそのキーワードに関する Web ページを閲覧しているメンバーが存在すれば，１つのキーワードに関して異なる観点の情報を得ることができ，今まで気付かなかったような知見の発見が期待できるためである．

さらに，関心のある情報を収集するために Web ページを閲覧するといった，日常的な行動から閲覧者の関心を推定するため，関心を共有する際の負担は少なく，本手法が活発に利用されることでメンバー間のコミュニケーションのきっかけが増えることが期待できる．また，本手法をシステムに実装する際，コミュニケーションの活性化を促すため，各メンバーが共有しているキーワードをマッチングし，マッチしたキーワードがあればそのキーワードと利用者および閲覧した時間を提示する機能など，コミュニケーションを支援するための機能を取り入れている．

第　3　章
システムの実装

3.1　システムの構成

閲覧 Web ページからの気付きについて提案手法の有効性を検証するため，本手法を Web アプリケーションとして実装した．本システムは Java（JDK1.5.0_07）により開発を行った．また，本システムを設置するアプリケーションサーバには Apache Tomcat（Ver.5.5.17）[13] を用い，本システムで利用するデータベースには PostgreSQL（Ver.8.1.4）[14]を採用している．システム環境およびシステム構成を図 2 に，システムを構成する各部の機能を表 1 に示す．

表 1．システム各部

閲覧 Web ページ収集部	利用者が閲覧した Web ページを収集する．
キーワード抽出部	閲覧 Web ページに出現する語句からキーワードを抽出する．
重要文抽出部	キーワードに基づいて重要文の抽出を行う．
可視化部	重要キーワード，関心キーワード，重要文，原文および未知語の 5 つを可視化する．

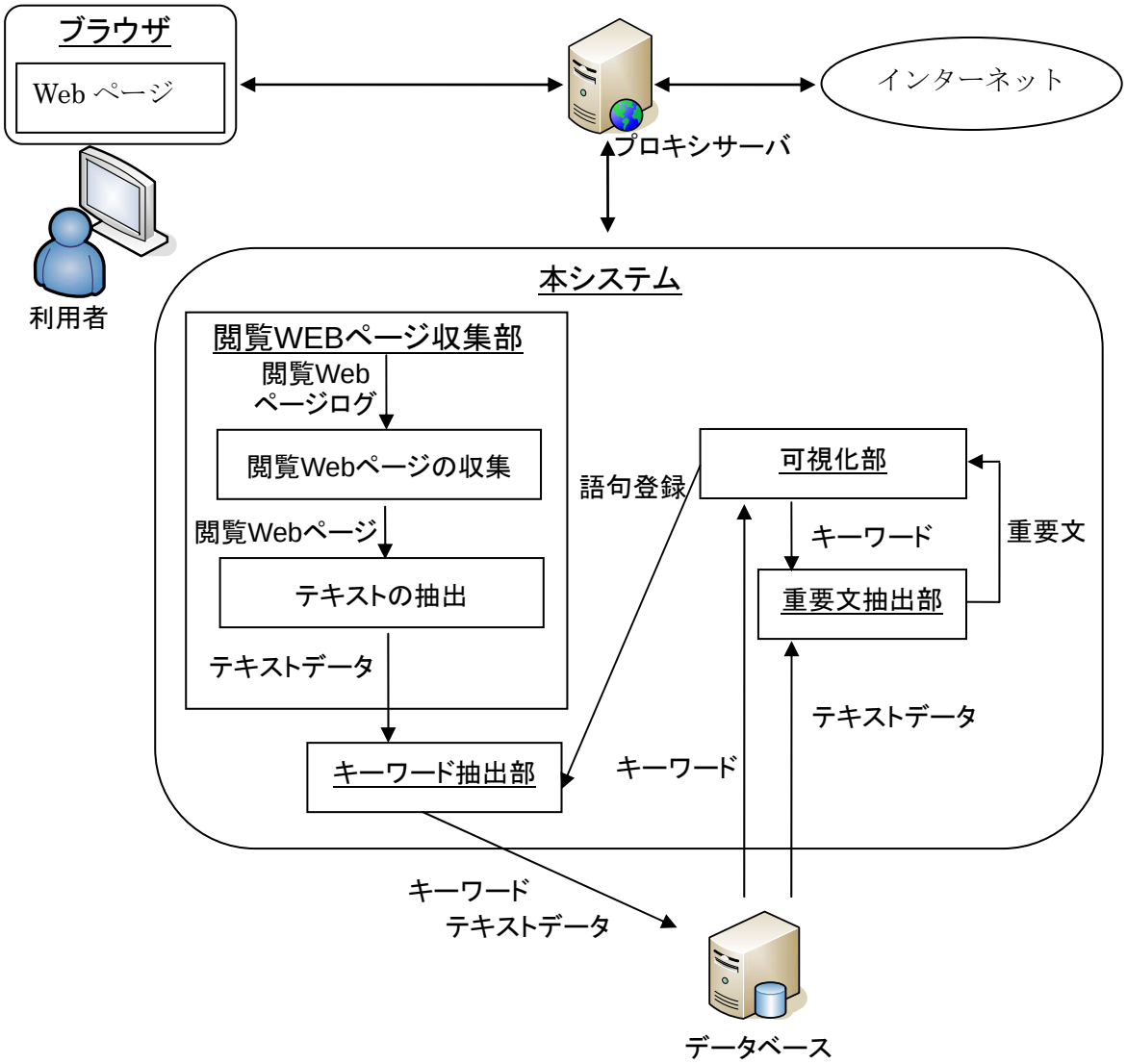

図 2．システム環境および構成図

図 2 を用いて本システムの利用プロセスについての説明を以下に述べる．

1．システムの利用者は，プロキシサーバを介して興味のある Web ページを閲覧する．

2．閲覧 Web ページ収集部によって収集された閲覧 Web ページからテキストデータを抽出し，そのデータはキーワード抽出部によって解析され，解析結果がデータベースへ保存される．

３．システムの利用者は，システムへアクセスし可視化部を通して閲覧 Web ページの解析結果について分析を行う．

3.2　閲覧 Web ページ収集部

閲覧 Web ページ収集部の構成を図 2 の閲覧 Web ページ収集部に示す．システムの利用者はプロキシサーバを介して Web ページを閲覧することを前提としており，閲覧 Web ページ収集部は，プロキシサーバのアクセスログを基に利用者が閲覧した Web ページを収集しテキストの抽出を行う．図 2 の閲覧 Web ページ収集部は，そのプロセスを示したものである．なお，本システムでは，プロキシサーバとして Squid（Ver.2.6）[15]，Web ページからのテキスト抽出には HTML Parser（Ver.1.6）[16]を用いている．

3.3　キーワード抽出部および重要文抽出部

本システムでは，古くから重要文抽出の手法として広く利用されている Luhn らの手法[17,18]を用いた．Luhn らの手法[17]は，文章に高い頻度で出現する語をキーワードとし，ひとつの文にそのキーワードがいくつ含まれるかによって文の重要度を設定する手法である．Zhang[18]らは，Luhn らの手法[17]を Web ページの要約手法の一部に取り入れることで，Web ページ要約システムの開発を実現している．本システムでは Zhang[18]らの Web ページ要約手法を基に重要キーワード抽出部および重要文抽出部の開発を行った．キーワード抽出部および重要文抽出部の構成を図 3 に示す．図 3 は，キーワードの抽出プロセスおよび可視化部を介した形態素解析辞書への用語登録，重要文抽出のプロセスについて示している．

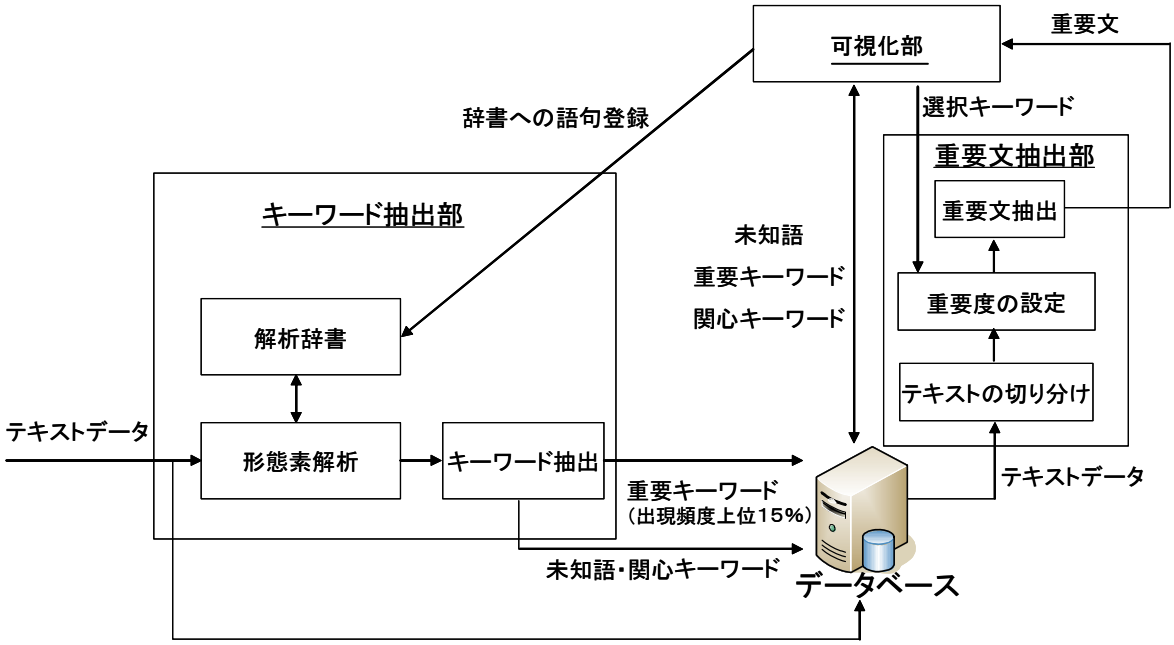

図 3. キーワード抽出部および重要文抽出部の構成図

3.3.1 GUI 画面からの形態素解析辞書の用語登録機能

本システムでは，Web ページから取得した文章を形態素解析システム Sen[19]により解析し，名詞を重要キーワードの候補として抽出する．しかし，形態素解析では，あらかじめ与えられた辞書に基づいて文章から名詞や動詞などの各品詞を特定するため，辞書に登録されていない品詞を特定し抽出することはできない．もちろん，関心のある分野が特定していれば，その分野の専門用語辞書から形態素解析システムのフォーマットに合わせて辞書を生成することはできる．

しかし，Web 上では毎日のように新たな用語が生み出され，短期間のうちにそれらの用語が一般的に使用されることも珍しくない．そのため，Web 上の情報を形態素解析システムによって解析し利用するには，形態素解析で用いる辞書に新たな用語を動的に追加する仕組みが必要である．そこで，本システムには，形態素解析システムによって未知語と判断された用語とその用語の出現頻度情報を利用者へ提示し，GUI の画面上から，利用者の判断で新たな用語を形態素解析の辞書へ登録できる機能を追加した．

本システムでは，形態素解析の辞書へ新たな用語を追加するかどうかを利用者が判断する必要があるが，Web 文書から専門用語を機械的に獲得するための手法[20]についても研究が行われており，用語登録機能については今後改善の余地が残されている．但し，新たな用語を機械的に形態素解析辞書へ追加するのではなく，新たな用語を登録するかどうかを利用者が判断する本システムの機能を共通の関心や問題意識を持ったグループで共有すれば，形態素解析辞書からグループの特性などなんらかの気付きを導き出すことができるかもしれない．図 4 に GUI 画面からの形態素解析辞書の用語登録機能による未知語の可視化画面を示す．図 4 において円で囲まれた部分の左部の語が形態素解析によって未知語と判断された語であり，右部の数値がその未知語が出現した頻度である．

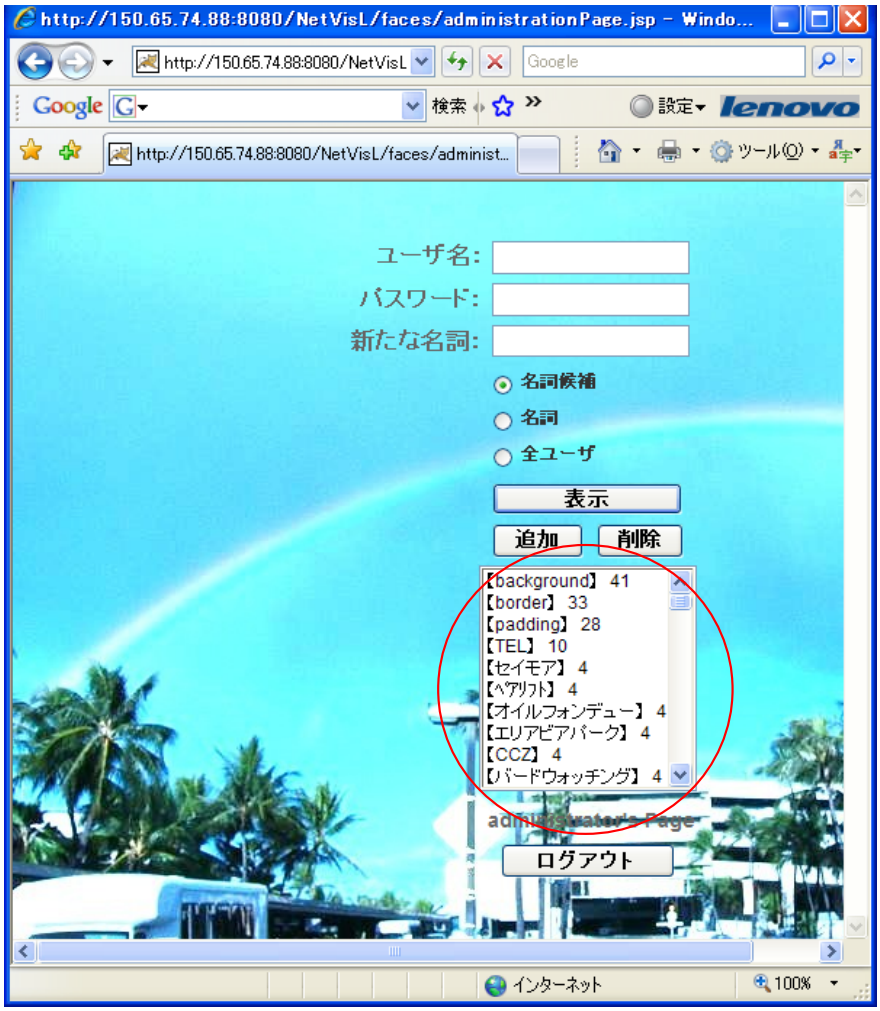

図 4.　未知語の可視化画面

3.3.2　キーワードの選定

本システムによる気付きへの初期のプロセスは，利用者が閲覧した Web ページがどのような特徴を持つのかを利用者にわかりやすく提示することである. Web ページから取得した膨大なテキストデータから，形態素解析によって抽出したすべての品詞を出現頻度に基づいて単にランキングして提示するだけでは，提示された品詞に一貫性がなく，Web ページの特徴を掴むのは困難である．また，本システムでは，重要文抽出を行う際に文章の抽出要素の 1 つとして修飾語句を設定できる仕組みとなっているため，名詞情報を取得できればその名詞を基に利用者の興味に沿った情報を取得することができる．そこで，本システムでは，Web ページの特徴を表す品詞として名詞のみを抽出し，重要キーワードの候補とする．

キーワードの選定プロセスを以下に記す.

Step 1：閲覧 Web ページ収集部によって収集された閲覧 Web ページのテキストデータを 1 ページごとに形態素解析システムによって解析し，各品詞に切り分ける．

Step 2：切り分けられた品詞群から，名詞および未知語と判断された語句を抽出する．

Step 3：各未知語とその未知語の出現頻度をデータベースに保存する．

Step 4：各名詞の出現頻度が高いものから上位 15%を抽出し重要キーワードとする．この際，重要キーワードの重要度には出現頻度そのものと各名詞の出現頻度数の高い上位 15%の出現頻度数の合計値を 100%とした際の割合を用いる．本研究では前者を絶対頻度，後者を相対頻度と定義し，Web ページの閲覧者名および重要キーワードと伴にデータベースへ保存する．

重要キーワードの可視化画面を図 5，6 に示す．図 5，6 はともに石川県の観光地について，インターネットを利用して情報を収集する際に本システムの各利用者が 1 時間の間隔で閲覧した Web ページ全体に含まれるキーワード群を可視化した様子である．図 5 におけるプロットチャートは，その各キーワード群のなかで最も高頻度で出現したキーワードをプロットしている様子である．プロットチャートの縦軸はキーワードが出現した頻度を示し，横軸は時間を示す．そして，図 5 におけるパイチャートは，プロットチャートにプロットされている点をクリックすることで，本システムの

各利用者が 1 時間の間隔で閲覧した Web ページ全体にどのようなキーワードが高い頻度で出現しているかを示している様子である．パイチャートを構成する各フィールドの名前は重要キーワードを示し，その隣の数値は重要度における絶対頻度である．

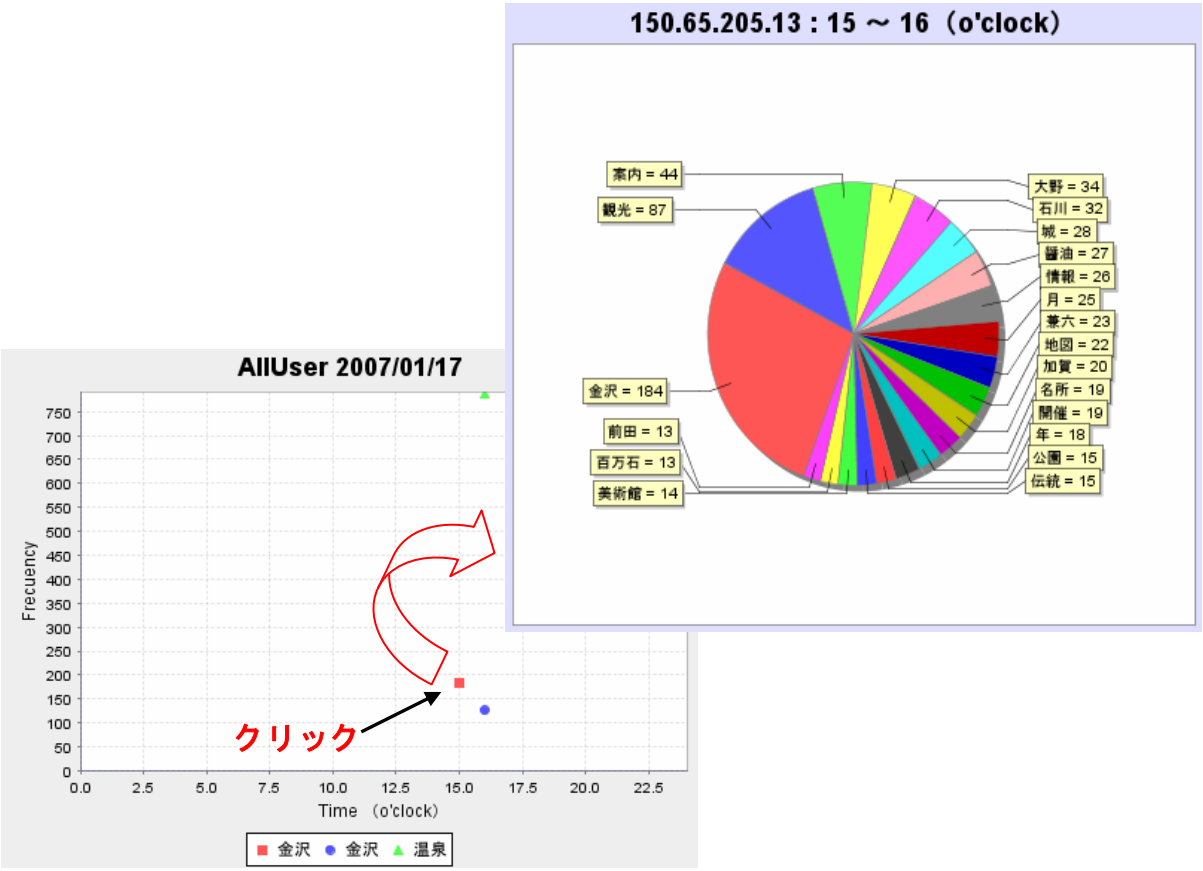

図 5．重要キーワードの可視化画面 1

また，図 6 は重要キーワード，Web ページの閲覧者名，閲覧された時間，キーワードの絶対頻度および相対頻度を提示している様子である．

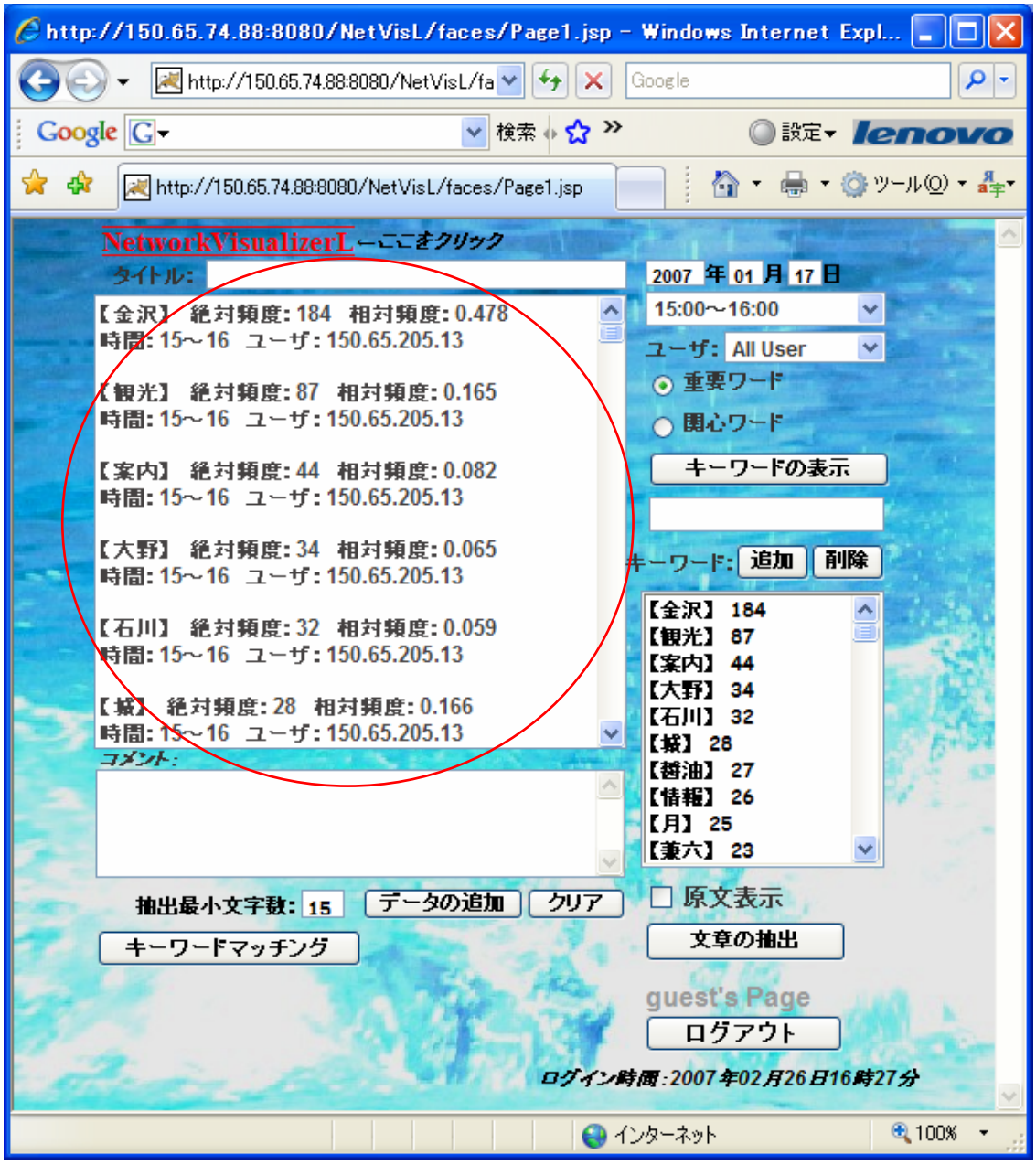

図 6. 重要キーワードの可視化画面 2

3.3.3　絶対頻度と相対頻度

本システムでは，重要キーワードの重要度を閲覧 Web ページのテキストデータ 1 ページごとに出現する頻度そのものと出現頻度の高い上位 15%の出現頻度数の合計値を 100%とした際の割合により設定する．本システムは 1 時間の間隔で閲覧された Web ページ全体に含まれる重要キーワードおよび重要度を利用者に提示する．この際，複数の Web ページに同じ重要キーワードが出現する場合は，それらの重要度の合計値を利用者に提示するため，利用者は絶対頻度および相対頻度がともに高いキーワードは閲覧者が関心を持って閲覧した Web ページに共通して出現するキーワードであると判断することができる．

3.3.4　データの解析間隔

本システムでは，閲覧 Web ページ収集部によって収集されるテキスト情報を解析するタイミングを一定の時間で区切り，閲覧 Web ページの解析を行っている．本システムでは，単純に時間という区切りでデータの解析を行っているが，データの解析をどのようなタイミングで行うかは大変重要なポイントであり，この点については一般的な Web 閲覧者のブラウジング特性や閲覧時間など様々な観点から検討する必要がある．

3.3.5　重要文抽出

以下に重要文抽出のプロセスを記す．

Step 1 : キーワード抽出部によって抽出されたキーワードを基に利用者が興味を持つキーワードを選択する．なお，キーワードは複数選択するこができる．

Step 2 : 閲覧 Web ページから得られたテキストデータを 1 文ごとに切り分ける．

Step 3 : 選択されたキーワードが 1 文に含まれる頻度によって文の重要度を決定する．

Step 4 : 重要度の高い文から降順に抽出を行う．

重要文の可視化画面を図 7 に示す．図 7 は，石川県の観光地について，インターネットを利用して情報を収集する際に閲覧した Web ページに含まれるテキストデータ

から重要文抽出を行い，重要文抽出に用いた選択キーワード，Web ページの閲覧者名，重要文の重要度および重要文を提示している様子である．

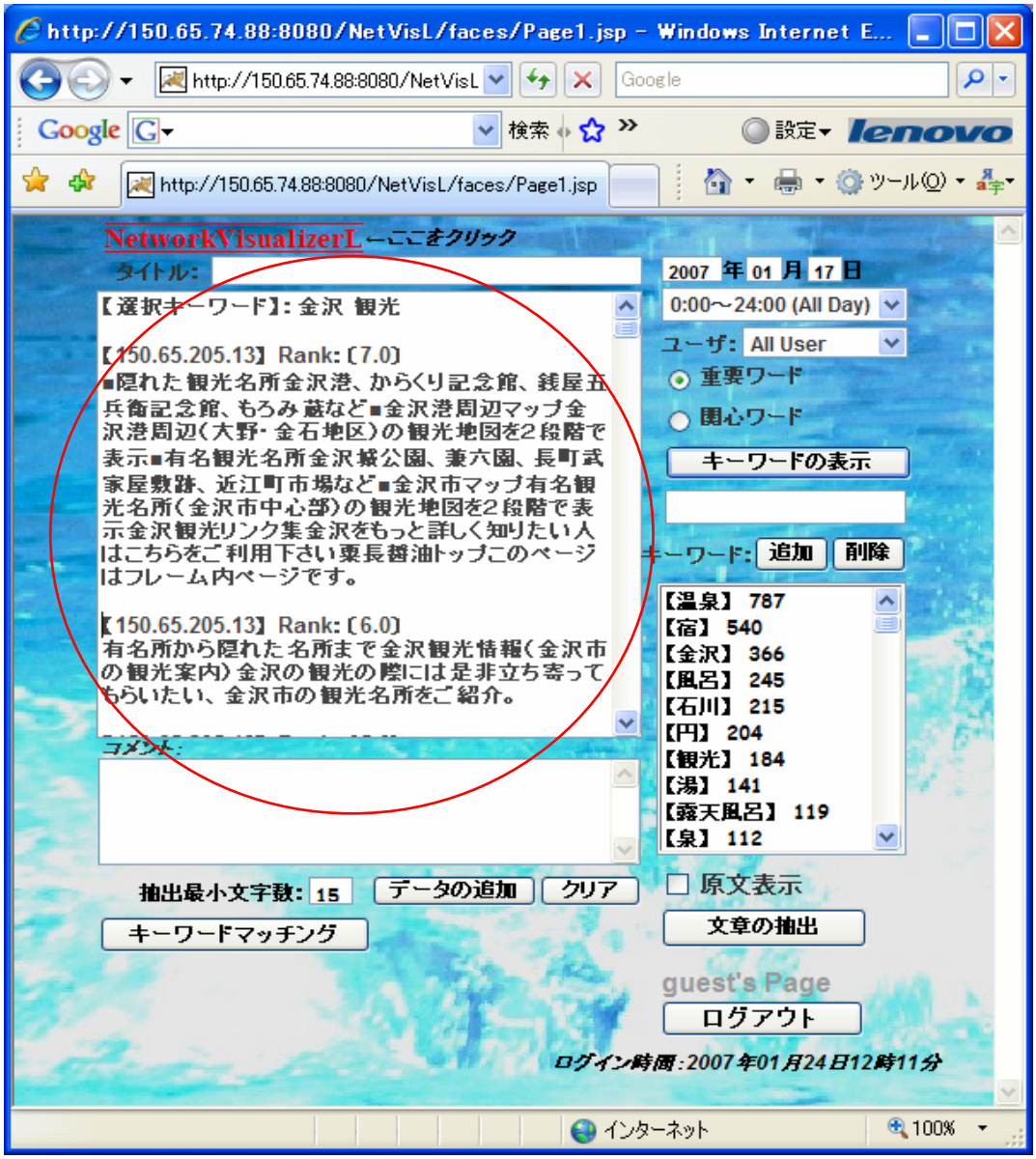

図 7．重要文の可視化画面

3.4　コミュニケーション支援機能

本項では，本システムを共通の関心や問題意識を持ったメンバー間で利用する際のコミュニケーションを支援する機能について説明する．

3.4.1　キーワードのマッチング機能

キーワードのマッチング機能とは，共通の関心や問題意識を持ったメンバー間で共有している閲覧 Web ページの解析データから，マッチング機能の利用者が閲覧した Web ページに含まれる重要キーワードもしくは関心キーワードをマッチングし，マッチしたキーワードがあれば，そのキーワードとメンバー名および閲覧した時間を提示する機能である．

3.4.2　関心キーワード追加機能

関心キーワード追加機能とは，本システムの利用者が関心を持つキーワードを登録する機能である．関心キーワードを登録しておくことで，キーワードマッチング機能により，容易に自分の関心にマッチするキーワードを発見することができる．また，メンバー間で関心キーワードを共有していれば，各メンバーの最新の関心事を知ることができ，コミュニケーションのきっかけとなる．

なお，3．3．2 節における，キーワードの選定プロセスの Step 4 において，最も出現頻度が高かった重要キーワードを閲覧者の関心事であると推測し，関心キーワードとして関心度を 1 に設定し，Web ページの閲覧者名とともにデータベースへ保存する．もし，データベースへ既に関心キーワードが登録されていれば，その閲覧者の関心キーワードの関心度をインクリメントする．関心キーワードの可視化画面を図 8 に示す．図 8 は，石川県の観光地について，インターネットを利用して情報を収集した際に抽出された関心キーワード群を提示した様子である．

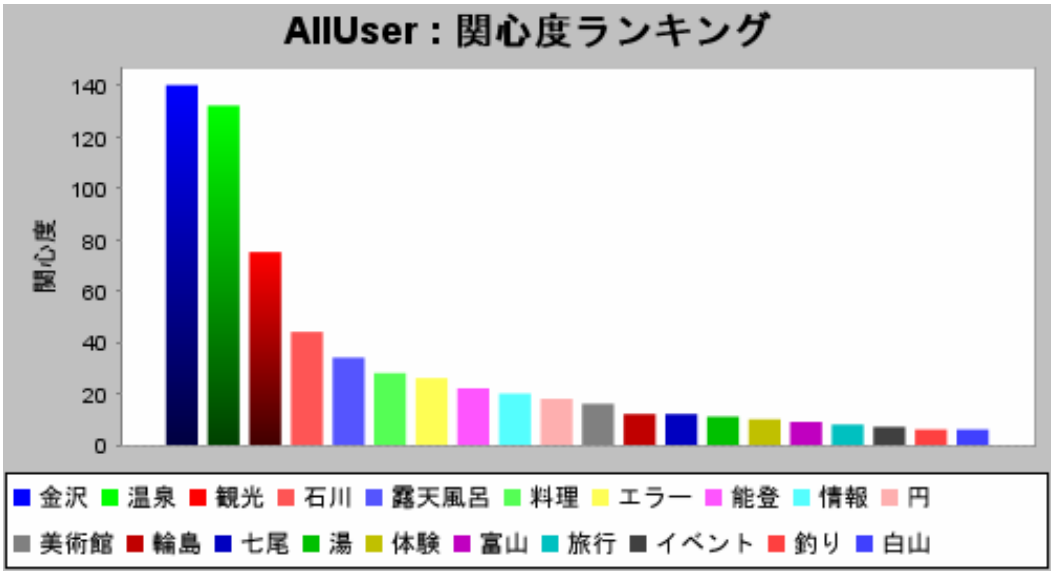

図 8. 関心キーワードの可視化画面

3.4.3 テキストデータの追加機能

テキストデータの追加機能とは，閲覧 Web ページ以外にも自分の興味や関心を持つ情報をテキスト形式で追加することで，その解析結果およびテキスト情報を共有するための機能である．テキストデータを追加する際には，その情報のタイトルやコメントも同時に追加し共有することができる．

3.4.4 原文抽出機能

原文抽出機能とは，Web ページ収集部によって収集された閲覧 Web ページおよびテキストデータ追加機能によって追加されたテキストデータの原文を抽出する機能である．

第　4　章
評価実験

4.1　実験 1

本手法は，インターネットを利用して閲覧した Web ページの解析を行い，利用者にとって有用な情報を提示することを目的としている．本手法では，利用者は閲覧 Web ページの解析によって得られた重要キーワードのランキングによって，自分もしくは他の利用者がどのような内容の Web ページを閲覧したのかを把握し，その情報を基になんらかの気付きを支援する．そのため，ランキングされた重要キーワードから閲覧者がどのような内容に興味を持っているかを推測できる必要がある．また，その重要キーワードや重要キーワードを基に重要文抽出を行い得られた結果が，利用者とって有用な情報である必要がある．そこで，本実験では，本手法を個人で利用した際の評価指標として以下の 3 つの観点から評価する．

A) 閲覧 Web ページの解析結果によって得られた重要キーワードのランキングが閲覧 Web ページの特徴を表しているか.
B) ランキングされたキーワードから閲覧目的に対して新たな視点が見出されたか.
C) ランキングされたキーワードを基にした重要文抽出により，閲覧目的に対して新たな知見やアイデアが発見できたか.

4.2　実験概要 1

本実験では，石川県に住む大学院生 12 人に，「石川県に知人を招待するならどのような場所へ案内したいか」についてインターネットを使って検索してもらい，案内したい場所とその理由を記述してもらった. なお, 案内したい場所が複数ある場合には, そのすべての場所を記述してもらった. また, 人によって検索スピードが異なるため, 検索時間は 15 分間を基準とし要望に応じて±5 分の差異を認めた．その後，各被験者が閲覧した Web ページの解析結果を用いて本システムを利用してもらい，本手法を個人で利用した際の評価指標 3 点について 5 段階（5 : すごくそう思う，4 : そう思う，3 : どちらともいえない，2 : そう思わない，1 : まったくそう思わない）の評価アンケートを取り，案内したい場所に変化があればその場所およびその理由について記述してもらった.

4.3　評価結果および考察

被験者自身が閲覧した Web ページの解析結果を用いて本システムを利用してもらった際の評価結果を表 2 に示す．表 2 における A，B，C は 4.1 節に記した個人で本システムを利用した際の評価指標 A，B，C をそれぞれ示している.

表 2．評価結果（個人での利用）

個人における本システムの利用に関するアンケート結果														
	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	平均	標準偏差
A	5	4	4	4	5	4	4	5	4	4	4	4	4.25	0.45
B	4	2	3	4	2	3	2	5	5	3	4	3	3.33	1.07
C	4	3	4	4	4	3	4	5	5	5	3	4	4.00	0.74

本システムを利用した際の評価アンケートの評価指標 A では， 12 人すべての被験者が 5 段階の評価で 4 以上の評価を与えており, 本手法によって抽出された重要キー

ワードの多くが閲覧目的に関連したキーワードであったことを示している．また，評価指標 B において 4 以上の評価を与えた被験者は 12 人中 5 であった．これは，本手法では，重要キーワードの選定に出現頻度のみを用いているため，興味を持って閲覧した Web ページに高い頻度で出現したキーワード群が閲覧者にとって新鮮なキーワードであることがまれであったためだと考えられる．一方，評価指標 B において 12 人中 9 人の被験者が 4 以上の評価を与えている．これは，個人で本システムを利用した際に得られる重要キーワード群は，閲覧 Web ページの特徴を捉えているものの，利用者が有用な情報を得るには十分でなく，重要文抽出によって得られる情報が必要であったと考えられる．また，本システムを利用することで新たに案内したい場所がみつかった被験者は， 12 人中 8 人おり，本システムを利用することで，新たな知見の発見を支援できる可能性があることがわかった．

4.4 実験 2

本実験では，共通の関心や問題意識を持った他者が閲覧した Web ページの解析結果を共有した際に本システムが有用であるかの検証を行った．共通の関心や問題意識を持った他者が閲覧した Web ページを本システムで利用した際の有効性の検証には以下の 3 つの評価指標を用いた．

A) 他者の閲覧 Web ページの解析結果によって得られた重要キーワードランキングは与えられた課題に関連したものであるか.
B) 他者の重要キーワードのランキングから課題の答えを導き出すための新たな視点を見出せたか.
C) 他者のランキングを基にした重要文抽出により閲覧目的に対して新たな知見やアイデアが発見できたか.

4.5　実験概要 2

本実験は，実験 1 の後に本システムを使用してもらうことにより行った．但し，閲覧 Web ページの解析結果として，筆者が事前に【温泉】,【料理】,【金沢】の 3 つの観点から「石川県に知人を招待するならどのような場所へ案内したいか」についてインターネットを用いて検索した際に閲覧した Web ページの解析結果を使用した．本システムを利用してもらった後に，他者が閲覧した Web ページを本システムで利用した際の評価指標 3 点について 5 段階の評価アンケートを取り，案内したい場所に変化があればその場所およびその理由について記述してもらった．最後に，本システムについての感想や意見を自由に記述してもらい実験を終了した．

4.6　評価結果および考察

他者が閲覧した Web ページの解析結果を用いて本システムを利用してもらった際の評価アンケートの結果を表 3 に示す．表 3 における A，B，C は 4.4 節に記した他者が閲覧した Web ページを本システムで利用した際の評価指標 A，B，C を示している．

表 3．評価結果（他者の解析結果を利用）

他者の解析結果の利用に関する本システムのアンケート結果														
	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	平均	標準偏差
A	4	5	4	4	5	4	4	4	4	4	4	4	4.17	0.39
B	5	5	4	5	4	2	5	4	4	3	4	2	3.92	1.08
C	4	2	4	3	4	4	4	5	4	5	4	2	3.75	0.97

他者が閲覧した Web ページの解析結果を用いて本システムを利用した際の評価アンケートの評価指標 A では，12 人すべての被験者が 5 段階の評価で 4 以上の評価を与えており，本手法によって抽出された重要キーワードの多くが閲覧目的に関連した

キーワードであったことがわかる．また，評価指標 B において 12 人中 9 人の被験者が 4 以上の評価を与えている．やはり，共通の関心に沿って閲覧した Web ページであっても，各被験者は異なる観点から情報を検索しており，他の被験者が収集した情報には，多くの有用なキーワードが含まれていたことが示唆される．また，被験者自らが閲覧したわけではないが，知っていたものの検索し忘れていたキーワードを発見できたという捉え方もできる．実際に，アンケートの自由記述欄には，「知っていたが意識していなかったキーワードを見つけることができた」といった内容の感想を記述している被験者が 3 名いた．評価指標 C においては，12 人中 9 人の被験者が 4 以上の評価を与えた．また，他者が閲覧した Web ページの解析結果を用いて本システムを利用した際に新たに案内したい場所を発見した被験者は 12 人中 9 人であった．

第　5　章
まとめと今後の課題

本研究では，興味を持って閲覧した Web ページに着目し，閲覧 Web ページの解析結果を閲覧者に提示することで, 気付きを支援する手法について提案を行った. また, 本手法を Web アプリケーションとして実装し評価実験を行うことで，本手法により新たな知見を発見できる可能性があることが確認できた.

技術的な課題として，今後は，重要キーワードの選定基準を単に出現頻度で決定するのではなく，高頻度なキーワード以外にも利用者にとって有用な情報を抽出できる仕組みを取り入れる必要がある．例えば，キーワード間の関係を考慮し，キーワード間の共起率を考慮する方法などが考えられる．また，閲覧された Web ページ全体に高い頻度で出現するキーワード以外にも，各閲覧 Web ページのテキストデータからそのテキストデータを代表するキーワードを抽出し利用者へ提示する仕組みとして，TF-IDF のアルゴリズムの導入についても検討する必要がある.

システムの評価についても，本システムを長期間に渡り利用してもらい，有効性についてさらに検証していく必要がある．また，今回の実験に対する本システムの有効性を検証するために用いた評価指標についてもさらなる検討が必要である.

謝辞

本研究を行うにあたって，多くの方々にご協力を頂きました．この場を借りて心から感謝の意を表します．本論文を執筆するにあたり，主指導教官である吉田武稔教授には，適切なご指導や助言を頂いたのみならず，研究活動の支援や機会を広く与えてくださったことに深く感謝致します．また，研究に関するお話以外からも物事の捉え方や考え方について学ぶところは多く，自分自身を成長させることができました．

本研究を行うにあたり，技術的なご指導ならびに広範な知識を持って，様々な意見やアイデアを提供してくださった林正治氏に深く感謝致します．また，研究活動についてだけでなく，様々な相談に乗ってくださり，研究に励むことができました．

研究活動を進めるにあたり，ご指導や助言をしてくださった堀井洋助手，權仁洙客員教授に深く感謝致します．權仁洙客員教授には，研究に不可欠でありながらも，私が苦手とする英語について根気よくご指導してくださり，大変ためになりました．

研究活動の息抜きとして色々な催しを開いてくれた吉田研究室のメンバーに深く感謝致します．最後に，研究を続けるにあたり陰ながら私を支えてくださった友人と家族に深く感謝致します．

参 考 文 献

[1] 堀井洋，林正治，他：メタデータ照合型医療情報通信監視システムの構築，第 26 回医療情報学連合大会，26th JCMI，pp.242-245 (2006).

[2] 九津見洋，内藤榮一，他：ユーザ適応型ホームページ推薦ソフト“ウェブナビゲーター”の開発，電子情報通信学会論文誌，Vol.J84-D-II，No.6，pp.1149-1157 (2001).

[3] 松尾豊，福田隼人，石塚満：ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援，人工知能学会論文誌，18 巻 4 号 E ，pp.203-211 (2003).

[4] 福村真哉，春本要，他：ユーザの視聴傾向に基づく Web コンテンツ個人化提示システム，情報処理学会論文誌，データベース，Vol.45，No.SIG 14(TOD 24)，pp.12-22 (2004).

[5] 河合由起子，官上大輔，田中克己：個人の選好に基づく複数ニュースサイトの記事収集・閲覧システム, 情報処理学会論文誌, データベース, Vol.46, No.SIG8(TOD 26), pp.14-25 (2005).

[6] 戸川聡，金西計英，矢野米雄：WAVISABI：Web 閲覧特性に基づく管理者支援のための利用動向可視化システム，情報処理学会論文誌，Vol.46，No.4，pp.985-994 (2005).

[7] 松村真宏，大澤幸生，石塚満：テキストによるコミュニケーションにおける影響の普及モデル，人工知能学会論文誌，17 巻 3 号 SP-B，pp.259-267 (2002).

[8] 松村真宏：オンラインコミュニティにおけるチャンス発見，人工知能学会誌，18 巻 3 号，pp.295-300 (2003).

[9] 梅田恭子，安田孝美，横井茂樹：知識メモを活用した研究情報共有方式の提案，情報処理学会論文誌，Vol.43，No.11，pp2562-2571(2001).

[10] 倉林則之，山崎達也，他：ネットワークコミュニティにおける関心の類似性に基づいた知識共有の促進，情報処理学会論文誌，Vol.43，No.12，pp.3559-3570 (2002).

[11] コンテンツと独立した動的第三者リンクによる知識共有支援 : 佐藤宏之，神戸雅一，金井敦，情報処理学会論文誌，Vol43，No.11，pp.3407-3418 (2002).

[12] 山田雅信，高橋俊行，他：インクリメンタル PageRank による重要 Web ページの効率的な

収集戦略, 情報処理学会論文誌, データベース, Vol.45, No.SIG 11(ACS 7), pp.465-473 (2004).

[13] Apache Tomcat. http://tomcat.apache.org/

[14] PostgreSQL. http://www.postgresql.org/

[15] Squid Web Proxy Cache. http://www.squid-cache.org/

[16] HTML Parser. http://htmlparser.sourceforge.net/

[17] The automatic creation of literature abstracts：Luhn,H.P., In IBM Journal for Research and Development, Vol.2 No.2 pp.159-165 (1958).

[18] Wen ZHANG, Xijin TANG：Web TEXT MINING ON XSSC, *Knowledge and Systems Sciences, toward Knowledge Synthesis and Creation Proceeding of KSS2006*, LNDS 8, Global-Link, pp.167-175 (2006).

[19] 形態素解析システム Sen. http://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sen.html

[20] 池野篤，濱口圭孝，他：Web 文書集合からの専門用語獲得，情報処理学会論文誌，Vol.47, No.6，pp.1717-1727(2006).