Web 閲覧特性に基づいた利用動向可視化による管理者支援システム

戸川　聡[†]　金西計英[††]　矢野米雄[†††]

本研究では，Web 閲覧動向可視化による管理者支援システムとその支援モデルを提案する．現在，大学などのキャンパスネットワークを流通するトラフィックの大半が Web 技術を基盤としており，ネットワーク上の様々な情報流通は Web を仲介として行われている．一方，現在一般的なネットワーク監視は，性能管理，障害管理の観点から実施されている．既存のネットワーク監視手法はネットワークの安定運用維持という観点からは有効だが，Web 閲覧に起因する問題発見への有効性はない．我々は，ネットワーク管理要素として Web 閲覧動向監視の必要性を述べ，この監視作業の支援モデル VANBETA を提案する．VANBETA は監視対象組織の Web 閲覧動向をカテゴリレベルで抽出し可視化することで，管理者の閲覧動向監視作業を支援する．さらに，有効性検証のため試作したシステムの構成と実験結果について述べ，我々の提案する支援モデルの有効性を明らかにする．

Administrator Assistance System based on Web Browsing Characteristic Using Users Activity Visualization

Satoshi Togawa,[†] Kazuhide Kanenishi[††] and Yoneo Yano[†††]

1. はじめに

Web の爆発的普及は様々な情報の入手を容易にした．現在，Web 利用者が入手可能な情報は，時事ニュースから原子爆弾の製造方法まで多岐に渡る．さらに，e-Learning やオンラインショッピング，e-Japan 戦略[1)] など，様々な教育活動や商業活動，公共サービスが Web 技術を基盤として盛んに推進，展開されている．今日，インターネットを流通するトラフィックの大半が Web に関係すると言っても過言ではない．

一方，企業内 LAN や大学のキャンパスネットワークは利用規定に基づき運用されている．一般に利用規定には，例えば教育研究目的に限定した利用のみ許可するなど，そのネットワークの利用目的や禁止事項が定義される．これらのネットワークの利用者は利用規定を遵守すべきであり，規定から逸脱する Web 利用は避けなければならない．しかし現実には利用者個々のモラルに依存しており，Web 利用時に規定が遵守されているとは言いがたい．

† 徳島大学大学院工学研究科
Graduate School of Engineering, University of Tokushima
†† 徳島大学高度情報化基盤センター
Center for Advanced Information Technology, University of Tokushima
††† 徳島大学工学部
Faculty of Engineering, University of Tokushima

ネットワーク管理者が利用者による規定外 Web 閲覧の制限を試みる場合，SFS/LB[2)] 等の閲覧制限機構を利用できる．これらは HTTP プロキシサーバやネットワークゲートウエイとして実装され，フィルタ情報やレイティング情報をもとに閲覧制限を実現する．一般にフィルタ情報には，利用者に閲覧させたくない URL 文字列やキーワード一覧が登録される．

URL 文字列を用いる方法では，フィルタ情報に登録した URL 文字列が利用者閲覧の URL に前方一致した場合，該当 Web ページ転送を遮断する．しかし変化の激しいインターネット社会ではフィルタ情報の陳腐化が早く，フィルタ情報を最新状態に維持することは難しい．また，Web 技術を用いて構築されるサービスが爆発的に増加しているため，URL を基盤とするフィルタ情報の網羅性に疑問が残る．

キーワードを用いる方法では，参照 Web ページ内における対象キーワードの出現頻度を判定材料とするものがある．キーワード出現頻度が管理者の設定する閾値を超えたとき，システムは Web ページ転送を遮断する．この方式ではシステムが閲覧を制限する Web ページを自律的に選択できる．しかし閾値設定状態により，本来閲覧を制限すべき Web ページが制限対象外となる誤判定の可能性が残る．

よって既存の閲覧制限手法では厳密な閲覧制限は難しい．完全な閲覧制限を実施するためには，管理者自らログ情報を調査し規定外 Web 閲覧を発見しなけれ

ばならない．これは，現時点において規定外 Web 閲覧を完全に排除するには，最終的に人力に頼らなければならないことを意味する．

一般にログ情報は膨大なテキスト情報により構成されている．ログ情報調査は必要な情報を抽出するために多大な労力が必要であり，管理者への負担が大きい[6]．こうして得られる調査結果は全体のログ情報から規定外利用に関係する部分のみを抜き出したサブセット情報であり，これもまた大量のテキスト情報で構成される．認知科学的観点から，一般に人間は大量のテキスト情報のみを用いて，その全体像を把握することは得意ではない．

このため，ネットワーク管理者は利用者モラルにのみ頼ることはできない．我々は，今，管理者には新たな観点でのネットワーク監視が求められている，と考えている．既存の障害監視，性能監視面のみではなく，Web 利用動向管理の観点からも監視が必要である．日常から利用者の Web 閲覧動向を監視することで，問題兆候の早期発見が可能となる．

現在，一般的なネットワーク監視手法として SNMP[3] を用いる方法がある．OpenView[4] 等の SNMP マネージャは，監視対象機器に実装される SNMP エージェントを介し MIB オブジェクト値を取得する．SNMP マネージャは取得したデータを統計処理し管理者に提示する．しかし MIB オブジェクト値から得られる情報は，あるインタフェイスのトラフィック量や機器の障害情報など，主に性能管理と障害管理に関するものである．

また，管理支援システムとして SPLICE/NM[5] や見えログ[6] が考案されている．SPLICE/NM は監視対象機器の障害診断と復旧作業自動化を実現している．見えログは計算機のログ情報を頻度解析し，大量のログ情報から少量の異常事象を抽出し可視化することで，管理者によるログ調査を支援している．しかしこれらのシステムは，監視対象の障害発見支援や復旧自動化を試みるものであり，ともに障害管理支援が目的である．

これら既存の管理支援システムでは，性能管理，障害管理面での支援機能は実現されているが，Web 閲覧動向管理という観点での支援機能は実現されていない．

そこで我々は，監視作業における管理者の認知的負担軽減を目的とした，閲覧動向可視化による監視支援：VANBETA (Visual Analyze Behavior Tracking Architecture) を提案する．

VANBETA は，利用者群が閲覧した Web ページ集合が含有するカテゴリ情報を抽出し可視化する．可視化された情報は特徴マップとして管理者に提供される．管理者は特徴マップを参照し，不適当なカテゴリ閲覧を発見できる．また，特徴マップを時系列に従って連続参照し，閲覧されるカテゴリ傾向や遷移を認識できる．その結果，監視対象組織の閲覧動向が俯瞰でき，日常では出現しない異質な振る舞いを発見できる．

また，特徴マップは部局など組織単位の閲覧動向を可視化する．組織単位を処理対象とすることで，特定個人の利用履歴を隠蔽でき個人プライバシに配慮している．

以下，本稿では 2 章で Web 閲覧動向可視化の現状と課題について述べ，3 章で閲覧動向可視化による監視支援について述べる．4 章で閲覧モデルの構成と可視化ついて述べ，5 章で監視支援モデル VANBETA に基づき試作したプロトタイプシステム WASABI について述べる．6 章で実験と評価について述べ，最後に 7 章でまとめる．

2. Web 閲覧動向監視の現状と課題

本章では管理者が行う Web 閲覧動向監視について，その現状と問題点を明らかにする．まず Web 閲覧動向を監視する場合，現時点で利用可能な手法について述べ，監視に適合した Web 情報カテゴライズについて述べる．次に監視作業における管理者の認知活動について述べる．そして定義した認知モデルに基づき，監視作業時の課題を明らかにする．

2.1 監視手法

現在，管理者が対象組織の Web 閲覧動向を監視する場合，以下の手法を利用できる．

HTTP プロキシサーバログの調査： 一般に HTTP プロキシサーバはクライアントからのリソース取得要求に従い，URL で示される Web サーバとファイルパスから目的リソースを取得する．取得したリソースは要求元クライアントに転送される．同時に下記項目を含む情報をログ情報として記録する．

- クライアント IP アドレス
- 実行日時
- 実行 HTTP コマンド（要求 URL 情報を含む）
- ステータスコード

一般にログ情報は大量のテキスト情報で構成される．管理者は，grep, sed, awk, sort などのフィルタコマンドや Perl 等のスクリプト言語を用いてログ情報を加工できる．ログ情報を加工することで，そのプロキシサーバを使用する利用者群が，いつ，どこから，どの URL で示される情報を，どの程度閲覧したかを抽出できる．

HTTP プロキシサーバは，Web ブラウザ側で明示的に使用宣言する場合が一般的である．しかし HTTP 通信を強制的に透過型プロキシへ誘導するなどの方法により，監視対象組織のすべての HTTP 通信を捕捉することも可能である．

プロトコルアナライザによる調査： 管理者は，Sniffer[7] や Ethereal[8] 等のプロトコルアナライザを用いて，観測点を流通する IP パケットを捕捉できる．捕捉した IP パケットを HTTP に限定して解析することで，HTTP プロキシサーバログ調査による方法と同等の情報を抽出できる．さらに高機能なツールであれば，プロトコル別利用状況などの統計情報を出力することも可能である．

いずれの手法も管理者が得る主な情報は，利用者群が閲覧する URL とその閲覧頻度であると言える．現在利用可能な手法では，管理者はこれらの情報から利用者群が閲覧するリソースのカテゴリ情報を認識しなければならない．

2.2 監視のための Web 情報カテゴライズ

Web 情報のカテゴライズは，利用者のための Web 閲覧支援の観点により早くから取り組まれてきた．代表的なものとして情報検索サービスの Yahoo![9] は，利用者からの推薦による Web 情報収集と人的資源による分類作業により膨大なカテゴリ情報に基づくディレクトリを構築している．

同様に，利用者への情報フィルタリングや情報推薦の観点から，ウェブナビゲータ[10] などのシステムが盛んに構築，提案されている．これらはユーザプロファイルに適応し，個人特性や嗜好に応じた Web 情報を分類，推薦する．

これら利用者のための閲覧支援を目的とした Web 情報カテゴライズは，Web 空間上に存在する情報を網羅的に分類する必要がある．このため収集した Web 情報の成長にともないカテゴリ情報も成長し，場合により階層構造を構成する．これら利用者向けの Web 情報カテゴライズは，網羅的な情報群から特定トピックを探索する場合は有効である．

しかし Web 閲覧動向監視を目的とする場合，問題兆候の発見を促進するカテゴライズが必要となる．閲覧支援目的のカテゴライズのような網羅的，かつ厳密な分類ではない．全体傾向を直感的に把握可能な一覧性の高い分類が必要となる．見通しの高いカテゴライズにより，利用者群が閲覧した Web 情報群の俯瞰が可能となる．

このため個々のカテゴリには厳密性を問わない．曖昧さを含むことよりも，全体傾向とそれに含まれる部分的な閲覧傾向を想起しやすいカテゴリ抽出を行うことで，閲覧傾向の把握が容易となり問題兆候の発見が支援される．

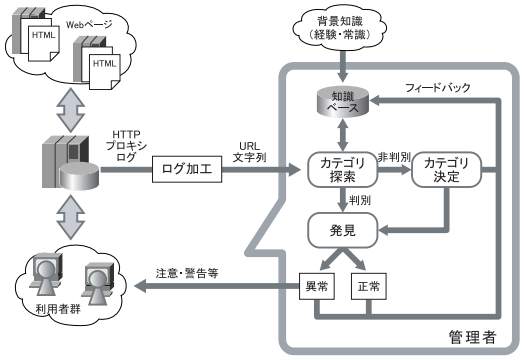

図 1 Web 閲覧動向監視における認知モデル

2.3 監視作業の認知モデル

2.1 で述べた手法により監視作業を行う場合，管理者はどのような認知活動を行うかを明らかにする．

まず，HTTP プロキシサーバログ調査による方法を用いる場合，監視作業における管理者の認知活動は**図 1** に示すモデルに従うと考えられる．

この認知モデルは「カテゴリ探索」「カテゴリ決定」「発見」の各機能と知識ベースから構成される．以下にそれぞれの詳細を述べる．

カテゴリ探索 管理者は HTTP プロキシログを加工し，利用者群が閲覧した URL 文字列と閲覧頻度を得る．カテゴリ探索機能は得られた URL 文字列を元に知識ベースを参照し，その URL が示す情報のカテゴリを判別する．

カテゴリ決定 カテゴリが判別できなかった場合，管理者はその URL で示される情報を実際に閲覧し該当カテゴリを決定する．こうして獲得された URL とカテゴリ情報の組み合わせは，新たな知識として知識ベースにフィードバックされる．

発見 局所的には，ネットワーク運用方針から逸脱するカテゴリの情報が閲覧されていないか発見する．大局的には，まず閲覧されているカテゴリ情報の傾向と遷移を把握し正常状態を認識する．その上で正常状態から逸脱する異常な閲覧動向を発見する．これはメタレベルでの現状認識と問題発見である．局所レベルでの異常情報と，大局的な問題発見に用いられた原因情報（カテゴリ情報の傾向と遷移等）は，新たに獲得した知識として知識ベースにフィードバックされる．

知識ベース 知識ベースは監視活動における判断や認識の基準となる監査データである．知識ベースは監

視作業で得られる獲得知識のフィードバックのみで構築されるのではない．管理者自身が日常的に行うWeb 閲覧，テレビや雑誌，書籍等から獲得する情報，その他日常生活で得られる経験や常識なども背景知識としてフィードバックされ，知識ベースは強化される．

プロトコルアナライザを用いる方法も，基本的にはログ解析による方法と同等の調査結果を得ることができる．よって，この調査方法における認知モデルにも図 1 で示すモデルを適用できる．

Web 閲覧動向監視において，その作業遂行のために管理者が直接利用可能な情報は，URL とその閲覧頻度である．管理者はこれらの情報から高度な認知活動と推論を行い問題を発見する．以上から，Web 閲覧動向監視作業は管理者の認知能力に大きく依存していると言える．

2.4 監視作業の課題

人間の認知能力には限界がある．特に大量情報を受け取った場合，そのすべてを正確に認識し処理することは難しい．現在の Web 閲覧動向監視作業が管理者の認知能力に依存している以上，その認知能力の限界が監視作業遂行時の限界となる．

以下，我々が定義した認知モデルに基づき監視作業遂行時の課題を述べる．

カテゴリ探索機能での課題（1）　管理者が最初に受け取る情報は，ログ情報から抽出した URL 情報である．URL はインターネット上に存在するリソースへのポインタであり，それ自体は単なる文字列に過ぎない．URL 文字列のみから，そのポイント先に存在するリソースが有する情報のカテゴリを判別することは難しい．さらにこの傾向は URL 文字列が長くなるほど強くなる．

例として以下の URL を示す．

(1)　http://www.tokushima-u.ac.jp/

(2)　http://www.ait.tokushima-u.ac.jp/main/node38_ct.html

URL（1）は徳島大学に関する総合情報を示すことが推測できる．しかし URL（2）は徳島大学の “何に関する” 情報を示しているか推測することは困難である☆．さらに個人のホームページなど，インターネットサービスプロバイダの Web サーバ名から始まる URL では，その内容推測は極めて困難である．

カテゴリ探索機能での課題（2）　管理者はログ加工の結果，大量の URL 情報を得る．URL 自体は単なる文字列であり，文脈的に意味のないテキスト情報で構成される．一般に人間は，大量の無意味なテキスト情報から，含有するカテゴリ情報を抽出することは得意ではない．

カテゴリ決定機能での課題　カテゴリ決定では管理者が判別不可能な URL に関し，その URL で示される情報を直接閲覧して内容を確認する．しかし大量の URL を個別に閲覧し内容を確認することは大変な労力を必要とする．

発見機能での課題　局所異常の発見，すなわち短期の監視期間におけるネットワーク運用方針から逸脱する閲覧カテゴリの発見は，人間の認知能力内で大部分の検出が可能と考えられる．しかし大局的な閲覧動向認識と異常発見は認知能力の限界により認識困難と考えられる．これは長期監視期間において閲覧カテゴリの遷移傾向を時系列で認識し，その傾向変化から異常状態を発見することを指す．

3. 閲覧動向可視化による監視支援

本章では，閲覧動向可視化による監視支援の枠組みである VANBETA について述べる．まず，Web 閲覧動向の監視支援のためカテゴリ抽出と閲覧動向可視化の必要性について考察する．次に VANBETA による閲覧動向可視化の方法を述べる．

3.1 カテゴリ抽出と可視化

2.4 では，Web 閲覧動向監視における管理者の認知能力の上限が，閲覧動向監視能力の限界であることを述べた．この限界をもたらす主要因は，管理者は膨大な URL 情報をもとに，それぞれの URL が指す情報のカテゴリを認識しなければならない，ということにある．これが閲覧動向監視での問題認知に大きな負担となることは容易に想像できる．

Web 閲覧動向監視において最終的に必要な情報は，利用者群の閲覧した Web ページはどのカテゴリに属しているか，ということである．閲覧された Web ページ集合から含有されるカテゴリ情報を抽出し管理者に提示することができれば，管理者の認知動作は提示されたカテゴリ情報を認識するだけでよくなる．利用者群が閲覧した URL 集合からの該当カテゴリ推測が不必要となり大幅な負担軽減となる．

しかし利用者群が日々閲覧する Web ページ量は膨大である．この Web ページ集合に含まれるカテゴリ情報を抽出すれば，これもまた膨大な情報量を有する．得られたカテゴリ情報を組織化し，その特徴のみ抽出できれば，管理者の認知能力範囲内での情報提示が可

☆ この URL は徳島大学高度情報化基盤センターの所在地情報を示す．

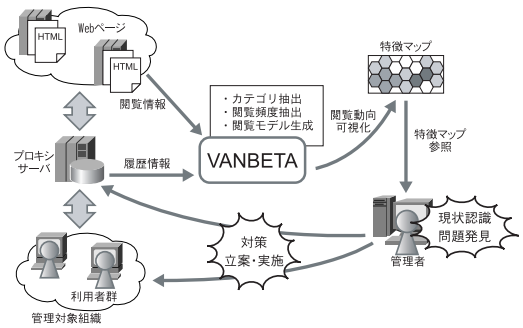

図 2　可視化による監視支援モデル

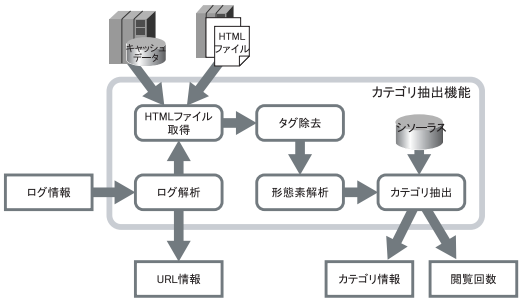

図 3　カテゴリ抽出機能

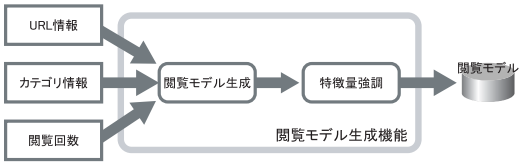

図 4　閲覧モデル生成機能

能となる．同時に提示情報を可視化すれば，利用者群が主にどのようなカテゴリの情報を閲覧しているか俯瞰できる．これにより閲覧傾向の全体像を容易に把握できる．

さらに可視化された情報を時系列にしたがい連続参照すれば，日常の閲覧傾向を把握できる．その上で閲覧傾向の遷移を観察すれば，日常状態から逸脱する異常な閲覧傾向を発見できる．

管理者による Web 閲覧動向監視の目的は，ネットワーク運用方針から逸脱するカテゴリの Web 閲覧を発見し必要な対策を施すことにある．我々は，以上の機能を実現することで，管理者は利用者群の Web 閲覧動向を直感的に把握できると考えている．これにより管理者の認知的負担が軽減され，利用者群の Web 閲覧動向が深く理解できるようになる．その結果，Web 閲覧における問題動向の発見が支援できる，と考えている．

3.2　可視化による監視支援モデル：VANBETA

本節では，可視化による監視支援の枠組みである VANBETA の振る舞いを述べる．**図 2** に VANBETA の構成を示す．VANBETA は「カテゴリ抽出」「閲覧モデル生成」「可視化」「発見支援」の各機能で構成される．以下，各機能の詳細を述べる．

3.2.1　カテゴリ抽出機能

カテゴリ抽出機能は利用者群が閲覧した Web ページ集合から，含有されるカテゴリ情報を抽出する．同時に該当カテゴリの閲覧回数も抽出する．また，その過程で得られる利用者参照の URL 情報も保存する．

カテゴリ抽出機能の詳細を**図 3** に示す．カテゴリ抽出機能への入力は，利用者群が使用する HTTP プロキシサーバのログ情報である．このログ情報には利用者群が行った Web 閲覧の履歴が蓄積されている．まずログ情報に含まれる URL 情報を抽出する．これをもとに利用者群が閲覧した HTML ファイルを収集する．得られた HTML ファイルからタグ情報を除去しプレインテキストに変換する．これを形態素解析し一般名詞と固有名詞に属する語をキーワードとして抽出する．抽出されたキーワードは，シソーラスを用いて該当するカテゴリに変換する．同時にキーワード抽出元 HTML ファイルの取得時に用いた URL の参照回数から，該当カテゴリの閲覧回数を得る．

こうして得られたカテゴリ情報とその閲覧回数は，利用者群が閲覧する Web 情報はどのようなカテゴリの情報を含有するか，ということを定量的に表現するものとなる．言い換えれば，利用者群の興味関心を定量的に抽出した，と言える．

3.2.2　閲覧モデル生成

カテゴリ抽出で得られたカテゴリ情報と URL 情報，閲覧回数から閲覧モデルを生成する．閲覧モデル生成機能の詳細を**図 4** に示す．

モデル生成には，情報検索分野で多用されるベクトル空間モデル (Vector Space Model:VSM) を使用する．抽出カテゴリを 1 ベクトルとし，ベクトル要素には利用者群閲覧の URL 情報を設定し，その特徴量にはカテゴリ閲覧回数を設定する．我々はこのベクトル 1 つをカテゴリベクトルと呼ぶ．

これにより，利用者群の興味関心をカテゴリベクトル個数分のベクトル集合で表現できる．その結果，余弦類似尺度の計算のみでカテゴリ間の距離関係を算出でき，利用者の興味関心間の相対的な位置関係の計測をベクトル間の距離計算に置き換えることができる．またベクトル空間モデルは，後述する可視化手法である自己組織化マップへの入力として親和性が高い[11)12)]．

得られた閲覧モデルは特徴量を強調するため，重み

付け処理を施す．重み付けによりモデルが持つ特徴を際立たせ，特徴をより深く把握できる．重み付けには，各カテゴリベクトルに集約されている閲覧回数を n 倍し，多く閲覧されたカテゴリを強調することとした．

3.2.3 可　視　化

閲覧モデルを可視化する．自己組織化マップ (Self Organizing Maps:SOM) を用いて，多次元ベクトル集合で表現される閲覧モデルの特徴を二次元平面に写像し，特徴マップを生成する．

生成された閲覧モデルは多次元ベクトル集合として構成されている．これは出現カテゴリとそれを参照する URL との関係が，多次元空間上での分布として表現されることを意味する．人間は基本的に 3 次元までの空間は直観的に把握可能だが，それ以上の多次元空間把握には困難を伴う．自己組織化マップは，多次元空間におけるデータ相互の距離関係を可能な限り保った状態で 2 次元空間に写像することができる．この結果，2 次元に写像された特徴マップを生成するため，すべての要素を一目で把握できる．

3.2.4 時系列提示による発見支援

3.2.3 で生成される特徴マップは，ある観測単位において利用者群が閲覧したカテゴリ情報とその閲覧量を可視化したものである．すなわち観測時における利用者群の興味関心を，その閲覧量とともに可視化している，と言える．特徴マップを時系列に従い連続参照すれば，閲覧動向の変化を認識できる．特徴マップ参照により日常における閲覧動向変化を把握すれば，正常状態とは異なる異常な閲覧動向の発見が容易となる．

4. 閲覧モデル構成と可視化

4.1 閲覧モデル構成

本節では VANBETA が生成する閲覧モデルの構成について述べる．閲覧モデルは以下の要素から構成される．

(1)　カテゴリベクトル
(2)　URL 情報
(3)　閲覧回数

カテゴリベクトルは，抽出したカテゴリ情報と参照された URL 情報を多次元ベクトルとして表現したものである．カテゴリベクトルを $\boldsymbol{a}$，URL 情報を a_1〜a_n とすると，カテゴリベクトルは次式で表現できる．

$$\boldsymbol{a} = \{a_1, a_2, \ldots, a_n\} \tag{1}$$

カテゴリベクトルの特徴量として，当該カテゴリの閲覧回数を URL 別に設定する．カテゴリベクトルは，当該カテゴリがどの URL に含まれているかを示すと同時に，利用者群が当該カテゴリをどの程度閲覧したかを集積したものと言える．

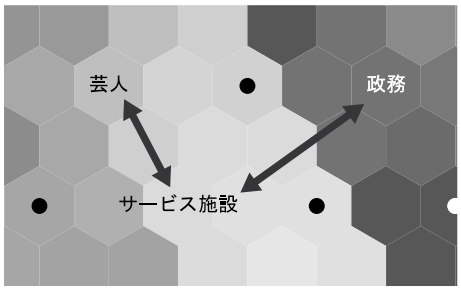

図 5　特徴マップ例

閲覧モデルは，生成されたすべてのカテゴリベクトルを集合させたものである．閲覧モデルを $\boldsymbol{A}$，カテゴリベクトルを $\boldsymbol{a_1}$〜$\boldsymbol{a_m}$ とすると，閲覧モデルは次式で表現できる．

$$\boldsymbol{A} = \{\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_m}\}^T \tag{2}$$

カテゴリベクトルは，カテゴリ抽出機能で抽出された総個数分生成される．したがって閲覧モデルは，生成した全カテゴリベクトルが集められたベクトル集合である．

n を当該カテゴリを参照した URL 数，m をカテゴリベクトル全生成数とすると，閲覧モデルは次式で表現できる．

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \tag{3}$$

4.2 自己組織化マップによる可視化

Kohonen により提唱された自己組織化マップ (Self-Organizing Map:SOM) は，2 層のニューラルネットワークで構成される教師なし競合学習モデルである．SOM はデータ間の幾何学的構造を学習アルゴリズムにより発見し 2 次元平面に写像する．このため特徴のよく似たデータは特徴マップ上の近傍領域に配置される．また同時にデータのクラスタリングも行う．

生成される特徴マップ例を**図 5** に示す．出現概念間の関連度が相対距離で示されると共にセル色の濃淡で示される．一般に色の淡いセル間は関連が高く，濃いセル間は関連が低い．図 5 の場合,「芸人」と「サービス施設」の関連度は高いが,「サービス施設」と「政務」の関連度は低いことがわかる．

5. Web 閲覧動向可視化システム: WASABI

本論文で提案する監視支援モデル VANBETA の有効性を検証するため，試作システム WASABI (Web-

browsing-Activity visualization System for Administrator assistance using Browsing Information）を作成した．

試作システム概要を**図 6** に示す．

本システムは大きく分けて，ログ解析部，HTML 解析部，カテゴリ抽出部，モデル生成部，可視化部より構成される．以下，各部の説明を行う．

5.1 ログ解析部

Web 閲覧者が利用する HTTP プロキシサーバのログ情報を取り込み，利用者が参照した URL の抽出を行う．抽出された URL 情報は HTML 解析部に渡される．

5.2 HTML 解析部

URL 情報を元に利用者が参照した HTML ファイルを取得する．利用者参照の HTML ファイルは HTTP プロキシサーバ内のキャッシュ情報として保持される．このため HTTP プロキシから取得可能な場合は HTTP プロキシから取得する．プロキシから取得できない場合，URL に示される本来の Web サーバからファイル取得を試みる．取得された HTML ファイルは HTML タグ除去を行い Plain Text に変換する．また漢字コード変換も行う．

5.3 カテゴリ抽出部

得られた Plain Text を元に形態素解析を行う．形態素解析には茶筅[15] version 2.2.9 を利用した．出力のうち一般名詞及び固有名詞に属し，かつ語の長さが全角 2 文字以上のものをキーワードとして抽出する．

得られたキーワードを元にシソーラスを参照し該当するカテゴリを抽出する．シソーラスには保持するカテゴリ毎にラベル付けされている．カテゴリを抽出するごとに得られたラベル値平均を算出し，その HTML ファイルのカテゴリ値とする．あるキーワードから複数のカテゴリが探索された場合，候補カテゴリが有するラベル値と HTML ファイルのカテゴリ値の距離が一番短いものを選択する．

5.4 モデル生成部

ベクトル空間モデルで定義される閲覧モデルを生成する．抽出カテゴリ 1 つに対しそのカテゴリを参照する URL 数が次元となる多次元ベクトルを生成する．本稿ではこれをカテゴリベクトルと呼ぶ．ベクトルの各要素には URL 別のカテゴリ参照回数が特徴量として保持される．カテゴリベクトルは抽出される総カテゴリ数分生成され，ベクトル空間モデルで構成される閲覧モデルを生成する．

5.5 可 視 化 部

得られた閲覧モデルを SOM アルゴリズムを用いて可視化する．SOM アルゴリズムにより抽出されたカテゴリ群が自己組織化され，似た特徴量を持つカテゴリが集約された特徴マップが生成される．管理者は得られた特徴マップを参照することで，管理組織の利用者が参照する Web 閲覧動向の俯瞰が可能となる．

なお SOM による可視化部分には Kohonen の研究グループで開発，配布されている SOM_PAK[16] を利用する．

表 1 実験環境

CPU	Intel Pentium 4 2.4GHz
Memory	640 Mbytes
HD	40 Gbytes
OS	Linux (kernel 2.4.18)

表 2 実験データ件数

種別	件数
実験データ件数	103,968 件
抽出 URL 件数	3,032 件
抽出カテゴリ件数	1,401 件

表 3 各部における平均処理時間

処理部名	処理時間
ログ解析部	1 分
HTML 解析部	30 分
カテゴリ抽出部	60 分
モデル生成部	3 分
可視化部	15 分
計	109 分

6. 実験と考察

6.1 実　　験

本システムに実験データを入力し特徴マップ生成を行った．ある組織が日常的に使用する HTTP プロキシサーバから使用許諾を得てログ情報を採取し実験データとした．ログ採取期間は 2003 年 3 月 25 日から同年 3 月 27 日までである．**表 1** に実験環境を示す．

表 2 に実験データ件数及び処理過程で抽出された URL 件数，カテゴリ数を示す．また，**表 3** に前述条件下におけるシステム各部の平均処理時間を示す．

6.2 考　　察

6.2.1 特徴マップ

図 7 に実験で生成された特徴マップを示す．

1 つの特徴マップは 20 × 16 の 320 要素を持つ．それぞれの要素には比較的多く閲覧されたカテゴリが表出する．今回の実験対象となった抽出カテゴリ数は 1,401 件であるため約 1/4 の代表カテゴリが表出していると言える．全てのカテゴリをそのまま出現させるのではなく代表カテゴリを出現させることで，全体傾向の俯瞰が可能となっている．さらに，特に多く閲覧されたカテゴリは要素の自己組織化により複数のクラ

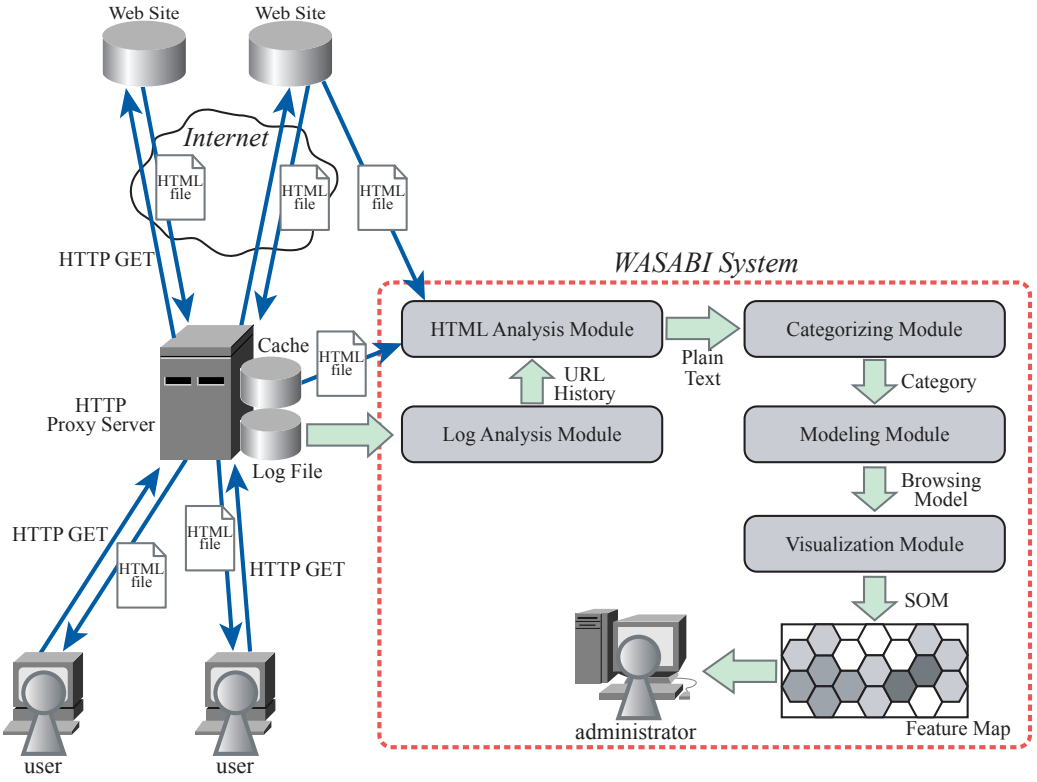

図 6　システム概要

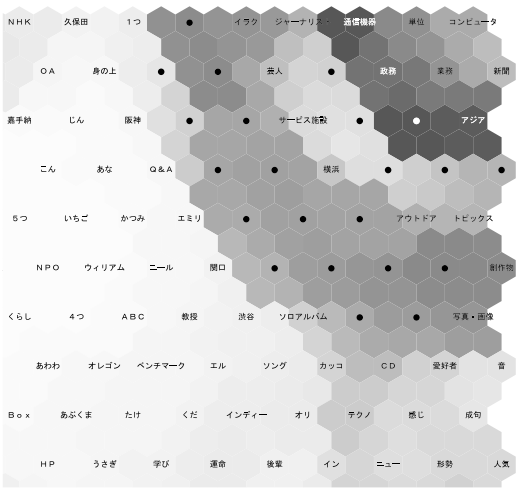

図 8　特徴マップ (3 月 25 日)

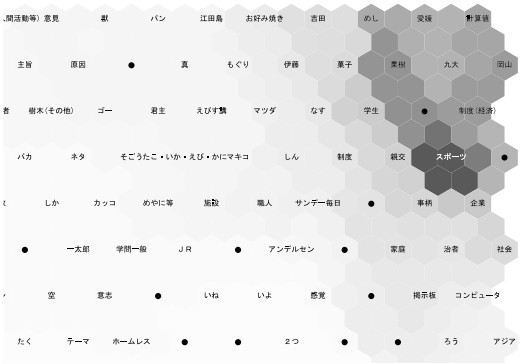

図 9　特徴マップ (3 月 26 日)

スタとして表出している．これにより，より多く閲覧されたカテゴリの的確な把握が可能となっている．

図 8〜**図 10** に，実験日ごとに生成した各特徴マップから，クラスタとして認識できる部分を示す．図 8 では「コンピュータ」「単位」などを中心とする半径 6 要素程度のクラスタを認識できる．**図 9** では「計算値」「愛媛」「岡山」などを中心とする半径 3 要素のクラスタを認識できる．さらにクラスタ内には「スポーツ」を認識できる．図 10 では「コンピュータ」「スポーツ」などを中心とする半径 5〜7 要素のクラスタを認識できる．これらから，実験期間である 3 月 25 日から 3 月 27 日において実験対象組織では「コンピュータ」「スポーツ」に分類される情報に対して比較的強い関心を有した，と推測できる．

生成した各特徴マップにおいて，クラスタとして自己組織化された要素数は 16〜84 要素である．特徴マッ

図 7 特徴マップ（全体図）

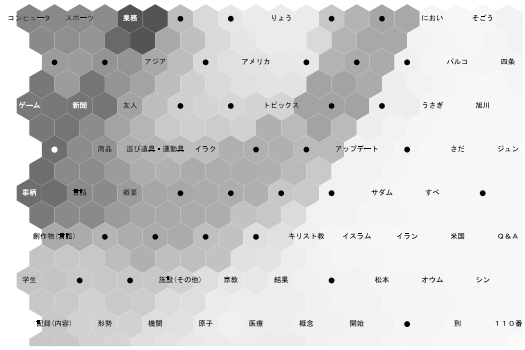

図 10 特徴マップ (3 月 27 日)

プが有する全要素数が 320 要素であることから各特徴マップにおいてクラスタ化されたカテゴリは，特徴マップ全体の約 5%〜26%であることがわかる．これから，今回の実験では最大で約 1/4 の Web 閲覧が何らかの同一傾向を示している，と言える．

また，図 8 と図 10 において,「イラク」「サダム」「アメリカ」「イスラム」などのカテゴリを認識できる．これから，日本時間 2003 年 3 月 20 日から開始されたイラク戦争に関連した情報に関しても継続的な関心を有している，と推測できる．

このように，対象日ごとの特徴マップを連続的に参照することで，管理者は日々の閲覧動向の傾向と推移を概観できる．

6.2.2 処 理 時 間

今回の実験における総処理時間は 109 分必要であった．以下，各処理部ごとに考察を行う．

ログ解析部，モデル生成部での処理時間はそれぞれ 1〜3 分であり処理件数を考慮するとほぼ妥当な時間であると考えられる．HTML 解析部にかかる時間は 30 分である．これを HTML 解析部での処理件数となる抽出 URL 件数 3,032 件で割ると 1 件当たり処理時間は約 0.59 秒となる．HTML 解析部で行う HTML ファイル取得，HTML タグ除去などの処理時間を考えると 0.59 秒/件の処理速度はほぼ実時間で処理されていると言える．

カテゴリ抽出部にかかる時間は 60 分である．今回の実験において，ある HTML ファイルが持つカテゴリ値算出のために，カテゴリ決定処理を 1 ファイル当たり 2 回実行する必要があった．これは HTML ファイルが含有するキーワードのカテゴリを決定するとき，複数候補を有するキーワードが存在したためである．これらあいまいなキーワードのカテゴリを決定するために，確定したカテゴリ値合計での平均を算出し，キーワードから探索されたそれぞれのカテゴリ値との最短距離を有するカテゴリを決定カテゴリとしたためである．このようなあいまいなキーワードは，1 つの HTML ファイル当たり 40〜50%存在したため，すべてが 1 回でカテゴリ決定される場合と比べ約 1.5 倍の

処理時間が必要となった．

可視化部にかかる処理時間は 15 分である．このうち本システムで利用している SOM_PAK の SOM 学習時間が 90%以上を占める．SOM 学習時間を決定する大きな要素は学習回数である．今回は 2 回目の学習を 1 万回行い，これにかかる時間が約 12 分であった．学習回数の減少に従い処理時間も減少するが，必要以上の学習回数削減は適切な学習効果を得られない．学習回数と学習効果のトレードオフ関係から最適なパラメータを決定する必要がある．

7. おわりに

本稿では，ネットワーク管理の一環として Web 閲覧動向監視の必要性について述べた．管理者が行う閲覧動向監視作業を支援するため，組織単位の利用者群における Web 閲覧動向を抽出し可視化する支援モデル VANBETA を提案した．また，提案した支援モデルの検証のために，試作システム WASABI を構築し実験を行った．実験データを用いて可視化した特徴マップを示し考察を行うとともに，管理者が特徴マップを参照することで利用者の Web 閲覧動向の把握が容易となることを示した．

今後は考察で示した改良点を導入するとともに，他の分類手法による分類結果と比較することで本システムの評価を行う．

謝辞 本研究の一部は，日本学術振興会科学研究費基盤研究 (B)(2) 一般 (No.13480047) の補助を受けた．

参 考 文 献

1) e-Japan 戦略 II，首相官邸. http://www.kantei.go.jp/jp/singi/it2/kettei/030702ejapan.pdf
2) 有害情報のレイティングとフィルタリング，財団法人インターネット協会. http://www.iajapan.org/rating/
3) Mark A. Miller: SNMP インターネットワーク管理, 翔泳社 (1998).
4) HP OpenView®. http://www.jpn.hp.com/openview/
5) 岡山聖彦, 山口英, 宮原秀夫: 作業のスクリプト記述に基づいたネットワーク管理支援システム SPLICE/NM の設計と実装, 電子情報通信学会論文誌, Vol. J81-D-I, No. 8, pp. 1014–1023 (1998).
6) 高田哲司, 小池英樹: 見えログ—情報視覚化とテキストマイニングを用いたログ情報ブラウザ, 情報処理学会論文誌, Vol. 41, No. 12, pp. 3265–3275 (2000).
7) Sniffer®. http://www.sniffer.com/
8) Ethereal. http://www.ethereal.com/
9) Yahoo!®. http://www.yahoo.com/
10) 久津見洋, 内藤栄一, 荒木昭一, 江村里志: 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” の開発, 電子情報通信学会論文誌, Vol. J84-D-II, No.6, pp. 1149–1157 (2001).
11) T. Kohonen: Self-Organizing Maps (3rd Edition), Springer-Verlag (2001).
12) Mark M. Van Hulle: 自己組織化マップ—理論・設計・応用, 海文堂 (2001).
13) 徳高平蔵, 岸田悟, 藤村喜久郎: 自己組織化マップの応用—多次元情報の 2 次元可視化，海文堂 (1999).
14) 塩澤秀和, 西山晴彦, 松下温: 「納豆ビュー」の対話的な情報視覚化における位置づけ, 情報処理学会論文誌, Vol. 38, No. 11, pp. 2331–2342 (1997).
15) 形態素解析システム 茶筅，http://chasen.aist-nara.ac.jp/
16) SOM_PAK and LVQ_PAK, http://www.cis.hut.fi/research/som_lvq_pak.shtml
17) Togawa, S., Kanenishi, K. and Yano, Y.: Web Browsing Activity Visualization System for Administrator Assistance, *2002 IEEE Int'l Conf. on Systems, Man and Cybernetics*, published CD-ROM only (2002).
18) Togawa, S., Kanenishi, K. and Yano, Y.: Web Browsing Activity Visualization System for Administrator Assistance Using Browsing Information, *Proc. of HCII2003*, Vol.1, pp.863–867 (2003).