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

戸川 聡[†] 金西計英[††] 矢野米雄[†††]

現在，大学などのキャンパスネットワークでは，Web を基盤として情報が流通している．このため，利用者全体の Web 利用の振舞いがネットワーク運用上の問題となることも多い．しかし，これまでの管理者支援ツールはパケット流量測定など，ネットワークの性能管理や障害管理を支援するものが主流である．そこで本論文では，利用者全体の振舞いから得られる情報の可能性を明らかにし，興味関心を可視化する管理者支援システム WAVISABI を提案する．WAVISABI は，Web マイニングシステムの一種であり，利用者が閲覧した Web コンテンツから閲覧特性を抽出，クラスタリングし可視化する．システムが提供する特徴マップにより，管理者の閲覧動向把握が支援できることを示す．さらに，閲覧動向の変化を示唆することで，異常発見支援の可能性を示す．

WAVISABI: Users Activity Visualization System for Administrator Assistance based on Web Browsing Behavior

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

Campus networks like those of university are occupied mostly with web traffic. Therefore, the behavior of web users often becomes a problem for the network administrator. The most general tools for network management are developed to achieve network performance management or trouble management such as the measurement of traffic. In our research, we have built a WAVISABI system that visualizes the browsing activity of web users. WAVISABI is a type of web mining system. This system extracts features of user interests from the web content that the users browsed, and after that, creates a feature map by clustering and visualizing. The network administrator can quickly understand the web browsing activity of the users of the system by refering to the map.

1. はじめに

ネットワーク管理者は，自組織のネットワーク利用動向を適切に把握しておく必要がある．たとえば，突然サッカーに関連する Web サイトへのアクセスが増えたとき，管理者は，どこかでサッカー大会が始まったと思い当たるかもしれない．そして，トラフィック状態を検討するなど次の作業にとりかかるのだろう．つまり，管理者にとって，利用動向の把握はそれ以降展開する様々な作業の基礎となる重要な作業だといえる．そこで，我々は利用動向の把握を支援するシステムの構築を行っている[1)~5)]．なかでもインターネット利用の中心を占める Web アクセスの動向把握に重点を置くことにした．

ネットワーク管理者の支援といったとき，最終的にはある組織内のスムーズなパケット転送が目的とされる．たしかに，パケットの遅滞ない転送は重要であり，管理者はそのために能力を割いている．ゆえに，適応的なパケット転送を目指すツールが管理者支援として注目される．

一方，我々の提案する管理者支援は，管理者の作業・活動に注目している．管理者があるツールを用いてネットワークを制御することではない．本論文では，管理者の行うネットワーク管理というタスクを分析し，タスクに潜む問題点を明らかにすることから管理者の支援を考える．

本研究の目的は，管理者が対外 Web アクセス動向を把握するための支援システム （WAVISABI: Web-browsing-Activity Visualization System for

† 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

Administrator assistance using users Browsing Information）の提供である．

WAVISABI はパケットの流れを直接制御しない．しかし，管理者の外部アクセスの動向把握は重要である．膨大な情報が流れるインターネット上で，なかでも Web アクセスの内容把握は容易ではない．そうしたとき，ポート 80 番へのアクセスが全体の何パーセントといった情報ではなく，どれくらいの利用者が，何に興味を持って Web アクセスを行っているのかが問題となる．こうした観点からの適切な情報がシステムより提示されるなら管理者の負担は大きく軽減する．

しかし，現状でネットワーク利用動向の把握といった場合，ネットワーク監視ツールの多くは，SNMP [6] による枠組みや，プロトコルアナライザによる解析など，トラフィックの量的監視を目的としている．また，利用動向を把握する際，個人のプライバシに配慮する必要がある．我々は個人の履歴をトレースするのではなく，利用者全体の履歴の統計的処理により，組織全体の傾向を取り出すようにした．

さらに，WAVISABI の目的は管理者支援であり，提供された情報を最終的に解釈するのは管理者自身である．管理者は提供された情報から，異常なアクセスなど様々なことを読み解く．ただし，WAVISABI は情報の遮断のみを目指すものではない．管理者はネットワーク利用の動向把握により，不適切な情報を遮断し，トラフィックが集中するアクセスに対してマルチホーム化などの対策をとる．つまり，利用動向を可視化する管理者支援ツールなどによって，管理者の負担が軽減され，組織のインフラとしてのインターネットサービスの質的向上が実現される．

本論文では，Web 閲覧動向をコンテンツの内容から興味の動向をクラスタリングし，可視化する手法について述べる．さらに，閲覧動向の比較から，何らかの変化をシステムが発見する方法についても考察する．まず本論文では，利用者が閲覧した Web コンテンツを処理対象とする．そして，収集したコンテンツ集合に含まれる閲覧特性を抽出し，特徴マップとして可視化し管理者に提示する．特徴マップは，閲覧特性として抽出したカテゴリとその閲覧頻度を一元的に把握できるよう構成されている．結果として，閲覧動向の定性的把握が支援でき，管理者による閲覧動向把握の負担が減る．さらに，蓄積された閲覧履歴をもとに，閲覧動向の変化について，システムが何らかの示唆が行えると考える．これは，管理者の異常発見の支援につながる．

以下，本論文では 2 章でこれまでの管理者支援ツールの方向性について述べ，3 章で管理者の閲覧動向把握タスクと課題について述べる．4 章で利用動向可視化による監視支援について述べ，5 章で Web 利用動向可視化システム WAVISABI について述べる．6 章で実験と考察について述べ，最後にまとめる．

2. これまでの管理者支援ツールの方向性

現在一般的なネットワーク監視手法として，SNMP を用いる方法がある．OpenView などの SNMP マネージャは，監視対象機器の SNMP エージェントから MIB オブジェクト値を取得し，統計処理ののち管理者に提示する．しかし，MIB から取得できる情報は，インタフェースごとのトラフィック量や機器の障害情報など，性能管理や障害管理に関するものである．

監視のための管理者支援システムとして，見えログ [7] が考案されている．これは計算機ログを頻度解析し，ログ情報から少量の異常事象を抽出することで，管理者によるログ調査を支援している．しかし，見えログは大量のログに埋没するパターンを発見，提示しており，ログを対象としたテキストマイニングといえる．さらに，見えログはサーバなど機器における障害発見支援を目的としている．

これら既存の管理者支援システムでは，トラフィックやログそのものを対象とした解析が行われている．これは主にネットワークや機器を対象とした性能分析，障害分析が目的となっている．しかし，本論文で述べる，利用者の興味関心を直接取り出し，分析し，管理者支援に応用する試みはみられない．また，管理者の作業タスクの分析に基づいた支援ツールもほとんどない．

3. 管理者の閲覧動向把握タスクと課題

管理者の行う閲覧動向把握のタスクを本論文では監視とよぶ．本章では，管理者が行う監視とその課題を明らかにする．

3.1 監視方法

一般的に，管理者は次のような方法で監視を行っている．

HTTP プロキシサーバログの調査 プロキシサーバは，クライアントから指示された URL をもとにコンテンツを取得し，これをクライアントに転送する．同時に，クライアント IP アドレス，実行日時，実行した HTTP コマンド，URL を含むログ情報を記録する．一般にログ情報は大量のテキストで構成される．管理者は grep などのフィルタコマンドや Perl などのスクリプト言語を用いてログを解析する．そ

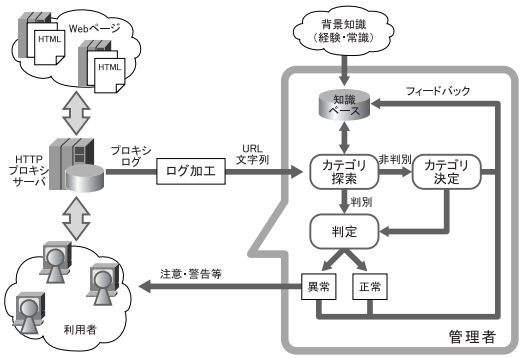

図 1 監視のタスクモデル
Fig. 1 Task model for monitoring.

の結果，利用者が要求した URL とその要求量，リクエスト時間などが抽出できる．

プロトコルアナライザによる調査 管理者は，Sniffer や Ethereal など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．

いずれの方法においても，監視のために管理者が利用可能な情報は，大量の URL とその閲覧量である．

3.2 タスクとしての監視のモデル化

図 1 に，管理者が URL とその閲覧量を用いて行う監視のタスクモデルを示す．このモデルは「カテゴリ探索」「カテゴリ決定」「判定」の各サブタスクと知識ベースから構成される．

本論文におけるカテゴリとは，利用者が閲覧する情報が妥当かどうか判定するための分類情報である．ウェブナビゲータ[8] など利用者の Web 閲覧支援とは違い，すべての閲覧情報を均一に分類する必要はない．

カテゴリ探索 管理者はログから抽出した URL 文字列をキーとして，知識ベースを参照する．知識ベースに当該 URL が存在すれば，その URL が示す情報がどのカテゴリに属するか判別できる．

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 を実際に閲覧し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合せは，新たな知識として知識ベースに登録される．

判定 管理者は獲得したカテゴリ情報から，閲覧された情報を判定する．局所的には，閲覧された情報が妥当かどうか判定する．大局的には，利用者が日常的に行う閲覧動向を把握し，これから乖離する状況を発見する．これらの判定結果とカテゴリ情報は，新たな知識として知識ベースに登録される．

知識ベース 知識ベースは監視タスクでの様々な判断の基準となる監査データである．知識ベースは監視タスクでの獲得知識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，書籍などから獲得する情報，その他日常生活で得られる経験や常識も背景知識として登録される．

監視タスクの遂行において，管理者が直接利用できる情報は，大量の URL とその閲覧量といえる．管理者はこれらの基礎的情報のみで閲覧情報を判定し問題を発見する．監視作業は，管理者の人間としての問題認識能力に大きく依存しているといえる．

3.3 監視時の課題

人間の問題認識能力には限界がある．特に大量情報を受け取ったとき，個別情報を認識し，全体傾向を把握することは難しい．監視が管理者の問題認識能力に依存しているかぎり，その限界が監視時の限界となる．以下，監視タスク実行時の課題を述べる．

カテゴリ探索時の課題 管理者は URL からカテゴリを探索する．しかし，URL 自体は概念的意味を持たない単なるテキストであり，URL のみによるコンテンツ内容の把握は難しい．例として，徳島大学高度情報化基盤センター Web サイト内の URL をあげる．

(1) http://www.ait.tokushima-u.ac.jp/aitdir/node4_ct.html
(2) http://www.ait.tokushima-u.ac.jp/aitdir/node20_ct.html

(1) は組織構成，(2) は利用規則に関する情報を提供している．特に，URL 命名規則が情報の内容に基づくものではなく，サイト内の物理構造に基づく場合，URL からの内容把握はきわめて困難となる．

カテゴリ決定時の課題 管理者はカテゴリ判別不可能な URL を直接閲覧し内容を確認する．しかし，大量の URL を個別に閲覧することは大変な負担となる．

管理者は，監視タスクの遂行において URL を直接処理し，URL が指すコンテンツの内容を確認しながら，全体的な閲覧動向を把握しなければならない．URL を監視対象とすれば管理者の作業負担は膨大であり，閲覧動向の詳細な把握は困難といえる．

4. 利用動向可視化による管理者支援

本章では，利用動向可視化による管理者支援を提案する．まず，監視のための管理者支援について述べる．次に，興味抽出と可視化による監視支援モデルを定義する．そして興味関心モデルである閲覧モデルについて述べ，自己組織化マップによる可視化について述べる．

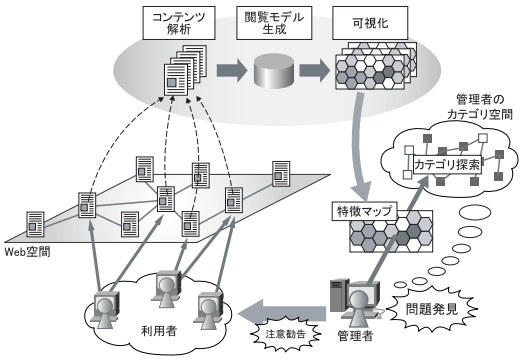

図 2 監視支援モデル
Fig. 2 Monitoring assist model.

4.1 監視のための管理者支援

本論文において，閲覧されたコンテンツが表現する意味的内容を概念とする．

監視作業での判定に必要な情報は，利用者が閲覧した概念は何か，ということである．コンテンツから概念が抽出でき，かつ，それが現実社会の価値基準に近いほど，監視作業を直接支援できるといえる．

一方，自然言語処理の研究分野にて，Web からの情報抽出が試みられており，HTML などの半構造文書から主要語や固有表現の抽出が実現されつつある[9),10)]．しかし，現時点でシステムが抽出した概念を，現実社会の価値基準と違和感なく対応させるには技術的課題が多い．これは本研究においても同様である．

そこで我々は，利用者が閲覧した概念を想起可能な情報提示を試みる．管理者はシステムが提示する情報から，利用者が閲覧した概念を想起できればよい．この結果，カテゴリ判定時の負担が軽減され，監視作業が支援される．管理者は閲覧動向の把握と問題認識に作業リソースを集中できるようになる．

4.2 興味抽出と可視化による監視支援モデル

管理者支援を実現するため，**図 2** に示す監視支援モデルを定義する．このモデルは「コンテンツ解析」「閲覧モデル生成」「可視化」機能から構成される．管理者は可視化された情報を参照し，利用動向を把握する．

4.2.1 コンテンツ解析

利用者が閲覧したコンテンツを解析し，閲覧特性を抽出する．監視のために管理者が把握すべきことは「何」に関する情報が閲覧されているか，ということである．コンテンツが主張する意味構造まで把握する必要はなく，「何」が「どれくらい」閲覧されているかを基本単位としてつかみ，その変化を知ることで利用動向を把握できる．

このため本研究では，コンテンツから名詞に分類される語を抽出する．このうち固有名詞に分類されるものは，コンテンツ概念を直接想起させる可能性が高いため，そのままカテゴリとして扱う．一般名詞に分類されるものは，シソーラスを参照し対応するカテゴリに分類する．あわせてコンテンツ自体の参照量を算出し，決定したカテゴリ情報の出現量とする．

獲得したカテゴリ情報から利用者が「何」の情報を，出現量から「どれくらい」閲覧したか想起できる．これは利用者の興味関心を抽出した閲覧特性といえる．

4.2.2 閲覧モデル生成

抽出したカテゴリ情報とその出現量から閲覧モデルを生成する．管理者が最終的に利用動向を定性的に把握するためには，単位時間において利用動向を定量的に表現する必要がある．基準となる定量的計測があり，その変化を追うことで利用動向の定性的把握が可能となる．

4.2.3 可 視 化

単位時間で集積された閲覧モデルを可視化し，管理者に提示する．本論文で扱う監視は，利用者の Web 利用動向把握とその変化発見の支援である．このため，管理者への提示情報は一目で全体状況が把握できることが望ましい．単位時間の状況が一目で把握できれば，閲覧特性の俯瞰が可能となり，変化の追跡も容易となる．

4.2.4 問 題 発 見

管理者は，閲覧特性が可視化された特徴マップを参照する．特徴マップは単位時間内に出現したカテゴリ情報が俯瞰できるよう構成される．管理者は，自身が持つ判定情報としてのカテゴリと，特徴マップが提示するカテゴリを照合し，利用動向変化をカテゴリレベルで判定できる．

4.3 閲覧モデルの構成

閲覧モデルは単位時間における閲覧特性を定量的に集積しなければならない．このため，モデル生成にはベクトル空間モデル（Vector Space Model: VSM）を適用する．モデルを構成する特徴ベクトルにはカテゴリが対応し，特徴量としてカテゴリ出現量を集積する．特徴ベクトルを $\boldsymbol{x}$，出現量を a_1〜a_n とすると，特徴ベクトルは次式で表される．

$$\boldsymbol{x} = \{a_1, a_2, \ldots, a_n\}$$

閲覧モデルは，生成されたすべての特徴ベクトルを集めたものである．閲覧モデルを $\boldsymbol{D}$，特徴ベクトルを $\boldsymbol{x}_1$〜$\boldsymbol{x}_m$ とすると，閲覧モデルは次式で表される．

$$\boldsymbol{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}^T$$

これにより，自組織の閲覧特性を特徴ベクトル $\boldsymbol{x}$ のベクトル集合で表現できる．結果，カテゴリ間類似度

を特徴ベクトル間の余弦類似尺度のみで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を，ベクトル間類似度で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)]．

4.4 自己組織化マップによる可視化

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．これはカテゴリ情報とそれを含むコンテンツの関係が，多次元空間上の分布として表現できることを意味する．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の把握には困難をともなう．

自己組織化マップ（Self-Organizing Map: SOM）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持した状態で特徴を 2 次元空間に写像する．加えて，似た特性を持つ特徴ベクトルをクラスタ化する[11),13)]．

SOM は 2 層のニューラルネットワークで構成される．入力層に入力される特徴ベクトルを $\boldsymbol{x}$，出力層の各ユニットに連結される参照ベクトルを，

$$\boldsymbol{m}_i = \{\mu_{i1}, \mu_{i2}, \ldots, \mu_{in}\}^T$$

とするとき，以下に SOM アルゴリズムを示す．

(1) 参照ベクトル $\boldsymbol{m}_i, (i = 1, \ldots, n)$ を初期化する．

(2) 次式に従い，全ノードから入力 $\boldsymbol{x}$ への最近傍ノード c を探す．これはユークリッド距離 $\| \boldsymbol{m}_i - \boldsymbol{x} \|$ が最小となるノードである．

$$c = \arg\min_i \| \boldsymbol{m}_i - \boldsymbol{x} \|$$

(3) 次式に従い，探索したノード c の特徴ベクトル $\boldsymbol{m}_c$ とその近傍ノードを更新する．

$$\Delta \boldsymbol{m}_i = \alpha h_{ci} (\boldsymbol{m}_i - \boldsymbol{x})$$

ここで α は学習率係数であり一般に 0〜1 の範囲をとる．h_{ci} は近傍関数であり，次式で示されるガウス関数である．ここで $\boldsymbol{r}_c$，$\boldsymbol{r}_i$ は，それぞれノード c，ノード i の位置ベクトルを示す．

$$h_{ci} = \exp\left(-\frac{\| \boldsymbol{r}_c - \boldsymbol{r}_i \|^2}{2\sigma^2}\right)$$

(4) (2)〜(3) を繰り返す．

本研究では SOM アルゴリズムに閲覧モデルを入力し，クラスタリングを行う．結果，閲覧モデルを 2 次元平面に写像した特徴マップが生成できる．これはモデル特徴を可能な限り保持したまま，認識性を高めた結果といえる．管理者は特徴マップを参照することで，出現量に応じたカテゴリの俯瞰が可能となる．

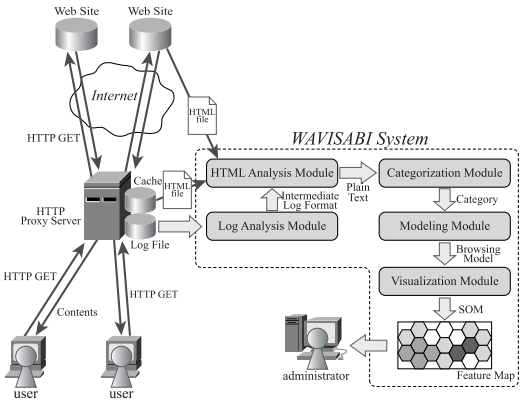

図 3 システム構成
Fig. 3 System architecture.

5. Web 利用動向可視化システム：WAVISABI

本研究で構築した Web 利用動向可視化システム WAVISABI の設計と実装について述べる．**図 3** に本システムの構成を示し，構成モジュールであるログ解析部，HTML 解析部，カテゴリ抽出部，モデル生成部，可視化部について詳述する．

5.1 ログ解析部

本システムでは，利用者が閲覧したコンテンツを収集する材料として，プロキシサーバのログ情報を利用する．すべての対外 Web アクセスはプロキシサーバ経由で行われる前提だが，プロキシを透過型キャッシュとして運用することで，利用者による意図的なプロキシ回避にも対応できる．

処理対象のログ情報は，squid-2.5.STABLE1[14)] のものを Apache[15)] 互換形式で出力したものである．ログ情報には，クライアント IP アドレス，アクセス日時，HTTP コマンド，URL，ステータスコードなどが記録されている．これを入力とし，以下の条件で選別する．

- URL が自組織以外である．
- HTTP コマンドが “GET” である．
- ステータスコードが “200(OK)” もしくは “304(Not Modified)” である．
- URL 末尾が ‘.html’ や ‘.htm’，もしくは ‘/’ で終了している．

この結果，対外 Web アクセスであり，かつ，HTML ファイル取得の成功したログが抽出できる．

選別後のログを整形し，中間形式ログとして出力する．中間形式ログには，クライアント IP アドレス，アクセス日時，URL が格納される．これにより，入力

となるプロキシサーバログのフォーマットが変更された場合でも，ログ解析部の修正のみで対応できる．

我々は今回，利用者が閲覧した HTML ファイルの URL を得るため，まずはログ情報のみから当該情報を抽出できる方法を選択した．今後の課題として，抽出精度向上のため HTTP Responce ヘッダ内の Content-Type フィールドを判定する方法の採用などが考えられる．

5.2 HTML 解析部

中間形式ログを用いて，利用者が閲覧した HTML ファイルを収集する．HTML ファイルのほとんどは，利用されるプロキシサーバから取得できる．しかし，キャッシュ期限切れの場合，URL で示される Web サーバから直接取得する．

取得された HTML ファイルは HTML タグを除去し，プレインテキストに変換する．加えて文字コード変換を行う．

5.3 カテゴリ抽出部

プレインテキストを形態素解析し，キーワードとして一般名詞および固有名詞に属する語を抽出する．形態素解析には茶筌[16] version 2.2.9 を利用している．

得られたキーワードのうち一般名詞に属する語は，シソーラスを参照し該当するカテゴリに変換する．この目的は，キーワードの持つ意味のあいまいさを除去するためと，可視化時のクラスタリングを高速化するためである．

固有名詞に属する語は，それ自体が内容を想起させる可能性が高い情報といえる．たとえば「イチロー」「松井」というキーワードからは「メジャーリーグ」に関する情報を容易に想起できる．このため固有名詞はそのままシソーラスに登録し，個別のカテゴリとして扱う．

シソーラスに登録されるすべてのカテゴリはラベル値を持つ．ラベル値は，カテゴリ空間における登録位置を示すが，カテゴリの持つ意味との関連性はない．

5.4 モデル生成部

閲覧モデルは総カテゴリ数分の特徴ベクトルを集合させたものである．特徴ベクトルの特徴量として，対応するカテゴリの出現量をコンテンツ別に保持する．閲覧モデルは単位時間ごとに生成され，1 つの閲覧モデルが 1 枚の特徴マップに対応する．

5.5 可視化部

得られた閲覧モデルを SOM アルゴリズムにより可視化する．SOM アルゴリズムにより抽出されたカテゴリが自己組織化され，似た特徴量を持つカテゴリがクラスタ化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．

表 1 実験環境
Table 1 Hardware specification for experimental usage.

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

表 2 実験データ件数
Table 2 Amount of experimental data.

種別	件数
実験データ件数	175,650 件
抽出コンテンツ件数	3,975 件
抽出カテゴリ件数	1,880 件

表 3 各部における平均処理時間
Table 3 Avarage time of each module.

処理部名	処理時間
ログ解析部	1 分
HTML 解析部	35 分
カテゴリ抽出部	65 分
モデル生成部	4 分
可視化部	20 分
計	125 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を生成した．ある組織が日常的に使用する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 18 日から同年 3 月 22 日までである．

表 1 に実験環境を示し，**表 2** に実験データ件数および処理過程で抽出されたコンテンツ件数とカテゴリ数を示す．**表 3** に実験環境下における平均処理時間を示す．

6.2 考　　察

6.2.1 特徴マップ

図 4 に実験で生成された特徴マップを示す．

1 つの特徴マップは縦 16 要素，横 20 要素の計 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 1,880 件であるため約 17%の代表的カテゴリが表出している．すべてのカテゴリをそのまま出現させるのではなく代表的カテゴリを出現させることで，全体傾向の俯瞰が容易になっている．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クラスタ化されて表出する．この結果，大量に閲覧

図 4 特徴マップ（全体図）
Fig. 4 Overall view of feature map.

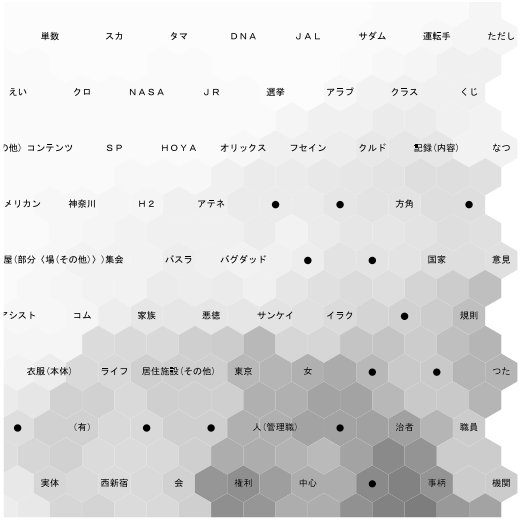

図 5 特徴マップ（3 月 18 日）
Fig. 5 Feature map for March 18th.

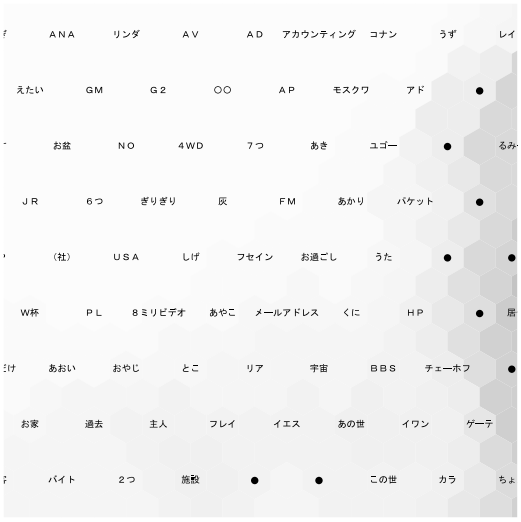

図 6 特徴マップ（3 月 19 日）
Fig. 6 Feature map for March 19th.

されたカテゴリの把握が可能となっている．

例としてイラク戦争に関連する事例を示す．イラク戦争は 2003 年 3 月 18 日，米国ブッシュ大統領からイラクに最後通告され，同年 3 月 20 日開戦した（日時はともに日本時間）．

図 5 は 3 月 18 日の特徴マップのうち，イラク戦争の関連部分を切り出したものである．「イラク」「フセイン」「バグダット」などの要素が確認できるが，それぞれクラスタとして認識できるほど閲覧されていないといえる．

図 6 は開戦前日の特徴マップである．このマップでは「フセイン」という要素しか関連する項目は表出し

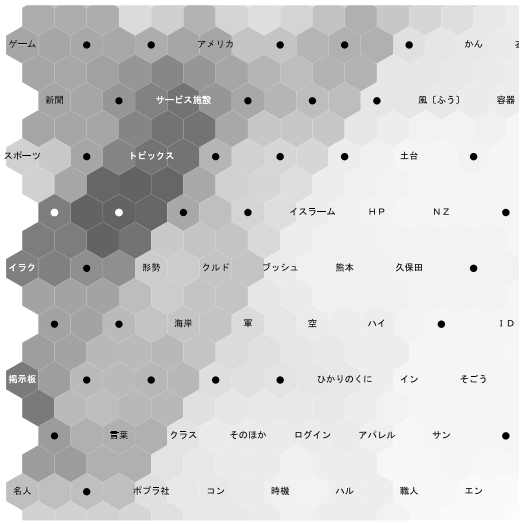

図 7 特徴マップ（3 月 20 日）
Fig. 7 Feature map for March 20th.

ていない．**図 7** は開戦日の特徴マップである．「イラク」を中心に半径 2〜3 要素のクラスタを認識できる．また，近傍に位置する「ブッシュ」「クルド」「アメリカ」などもそれぞれ半径 2 要素ほどのクラスタが形成されている．

このように，特徴マップを連続的に参照することで，管理者は単位時間ごとの Web 利用動向とその推移を把握できる．

6.2.2 特徴マップの管理作業への適用

本項では，特徴マップから獲得した情報のネットワーク管理作業への適用について考察する．

管理者は本システムから得られる特徴マップや作業上扱うログ情報を日常的に閲覧することで，管理するネットワークの定常状態を知る．加えてこれらの作業を継続的に行うことで，定常状態と特徴マップ参照時点との差異が認められないことを知る．すなわち，異変が発生していないことを「確認」できる．

同様の作業から，定常状態と特徴マップ参照時点との差異を認識する．この時点で，管理者は日常とは違うことを「発見」する．たとえばその差異が性的表現などの通常では好ましくないカテゴリの表出だとすると，何らかの目的外利用が行われていると推測でき，その結果異常発見につながる．この場合管理者は，フィルタルール見直しなどの作業に着手できる．さらに利用特性の変化などの差異にも気づく．たとえば，毎週月曜日に経済関連の Web アクセスが頻繁に発生しているとする．大学などの場合，当該曜日に経済関連の授業が展開されていると推測できる．その場合管理者は，MRTG などを用いてトラフィック状態を確認する．もし，クラスタ化されたトラフィックが学内の他のトラフィックと比べ突出していることが分かれば，管理者は学内ネットワークの物理的構成に問題がある，あるいは，対外接続線に問題があることに気づく．その結果，マルチホーム接続化の検討や，その講義専用のプロキシサーバ運用など，次の管理作業を計画，実施するための契機となる．

WAVISABI は利用者のトラフィックを直接制御することはない．しかし，特徴マップから得られる情報から，ネットワーク利用の現状が把握でき，またその変化がトリガとなり，次の管理作業にとりかかる契機を与えることができる．

6.3 カテゴリ指標抽出による利用動向変動の示唆

本節では，単位時間ごとに生成される閲覧モデルのモデル間変動を抽出することで，利用動向変動の示唆を試みる．

WAVISABI は，利用者が閲覧した Web コンテンツ集合からカテゴリを抽出し，特徴マップとして可視化する．しかし，現状では特徴マップ間の利用動向変化は，管理者が特徴マップから直感的に読み取る必要がある．

そこで我々は，まず，閲覧モデルに含まれるカテゴリ情報とその出現頻度から指標を算出する．本論文ではこれをカテゴリ指標とよぶ．次に，カテゴリ指標の変動過程を示し，管理者が閲覧動向の変動を定量的にとらえられるようにする．

6.3.1 カテゴリ指標の算出

閲覧モデルに使用されるすべてのカテゴリには，それぞれラベル値が付与されている．これは，出現する全カテゴリに一様かつ一意に割り当てられている．カテゴリ指標は，閲覧モデルに含まれる各カテゴリの出現頻度と，そのカテゴリに付与されるラベル値の積を最大ラベル値で割り，0〜1.0 の範囲に正規化したものである．

閲覧モデルに含まれる総カテゴリ数を n，カテゴリの出現頻度を C_f，カテゴリに付与されるラベル値を C_v，ラベル値の最大値を C_{max} とすると，カテゴリ指標 C_{idx} は次式で表現できる．

$$C_{idx} = \frac{\sum_{i=1}^{n} C_v C_f}{C_{max}}$$

これからカテゴリ指標は，全カテゴリ空間における当該閲覧モデルの位置を示すものといえる．

6.3.2 カテゴリ指標変動過程の提示

特徴マップの作成実験に使用した閲覧モデルを用いて，カテゴリ指標を算出した．ここで閲覧モデルは 1 時間単位に生成し，それぞれカテゴリ指標を算出した．

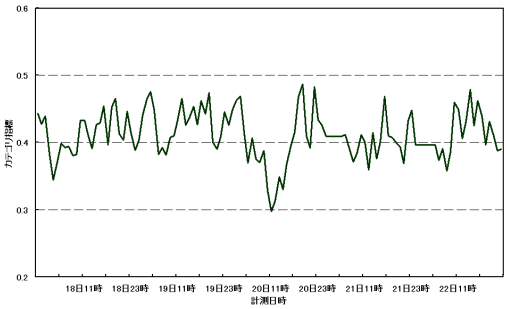

図 8 カテゴリ指標変動（原データ）
Fig. 8 Variation of original category index.

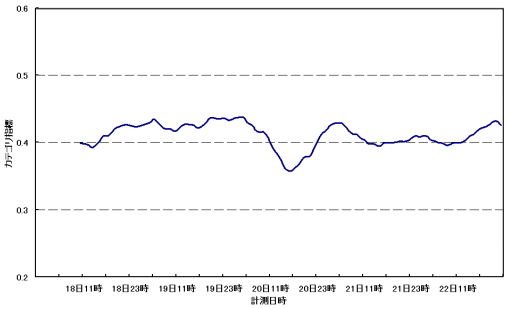

図 9 カテゴリ指標変動（移動平均）
Fig. 9 Variation of moving averaged category index.

得られたカテゴリ指標を時系列に従いグラフ化したものが**図 8** である．実験期間におけるカテゴリ指標の変動は，おおむね 0.35〜0.5 の間を振幅していることが分かる．しかし，この状態では値の上下動が激しく，仮に大幅な動向変化が発生したとしても他の小規模な変動に埋没する可能性が高い．

そこで，カテゴリ指標の系列から移動平均を生成し，小規模変動の平滑化を試みる．**図 9** は，12 時間分のカテゴリ指標から単純移動平均を算出しグラフ化したものである．図 8 に比べ，平均的な指標変動がおおむね 0.39〜0.43 間の振幅に平滑化されていることが分かる．その中でカテゴリ指標が 0.35 程度まで低下している個所が認められる．この変化は 3 月 20 日 12:00 頃から始まり，同日 22:00 頃に定常状態に戻っている．グラフの傾斜量を連続的に観測すれば，定常からの利用動向の乖離をシステムが示唆できると考えている．

なお，日本時間の 3 月 20 日 11:33 に，米軍はイラク攻撃を開始し，イラク戦争が開始された．これとほぼ同時に，ニュースサイトなどでいっせいにイラク攻撃開始が報じられ，相当数の利用者が当該ニュースを閲覧したと想像できる．このため，Web 利用動向がイラク戦争関連ニュースに遷移し，結果としてカテゴリ指標系列が定常状態から乖離したと考えられる．

7. おわりに

本論文では，管理者が行う新たなネットワーク監視として Web 利用動向の監視の必要性を述べ，監視タスクの分析から，監視作業における管理者の作業負担を明確にした．利用者が閲覧したコンテンツからの興味関心を抽出，可視化する監視支援モデルを定義し，これに基づき構築した利用動向可視化システム WAVISABI を提案した．さらに，実験データにより生成した特徴マップについて議論し，利用動向可視化による管理者支援の有効性を明らかにした．加えて，カテゴリ指標変動による利用動向変化の示唆の可能性について述べた．

今後は，考察で示したカテゴリ指標による利用動向変化の検出をシステムに実装し，同機構の評価，検証を行う．さらに，本論文で提案した手法の一部をトラフィック解析に応用し，不適切トラフィックの検出支援環境などを構築していく予定である．

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）ならびに基盤研究（C）（2）（No.16500591）の補助を受けた．

参考文献

1) Togawa, S., Kanenishi, K. and Yano, Y.: Web Browsing Activity Visualization System for Administrator Assistance, *Proc. 2002 IEEE Intl. Conf. on Systems, Man and Cybernetics*, published by CD-ROM only (2002).
2) 戸川 聡，金西計英，矢野米雄：Web 閲覧の個人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の構築，情報処理学会研究報告（2002-DSM-28），Vol.2002, No.118, pp.49–54 (2002).
3) Togawa, S., Kanenishi, K. and Yano, Y.: Web Browsing Activity Visualization System for Administrator Assistance Using Browsing Information, *Proc. 10th Intl. Conf. on Human-Computer Interaction*, Vol.1, pp.863–867 (2003).
4) 戸川 聡，金西計英，矢野米雄：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，インターネットコンファレンス 2003 論文集，pp.77–86 (2003).
5) Togawa, S., Kanenishi, K. and Yano, Y.: Web Browsing Activity Visualization System for Administrator Assistance Using Users' Browsing Behavior, *Proc. 18th Intl. Conf. on Advanced Information Networking and Applications*, Vol.2, pp.279–284 (2004).
6) Miller, M.A.: SNMP インターネットワーク管理，翔泳社 (1998).

7) 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).
8) 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).
9) 山田泰寛，池田大輔，坂本比呂志，有村博紀：WWW からの情報抽出，人工知能学会誌，Vol.19, No.3, pp.302–310 (2004).
10) Chang, G., Healey, M.J., McHugh, J.A.M. and Wang, J.T.L.: Web マイニング，共立出版 (2004).
11) Kohonen, T.: *Self-Organizing Maps, 3rd Edition*, Springer-Verlag (2001).
12) Van Hulle, M.M.: 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) Squid Web Proxy Cache. http://www.squid-cache.org/
15) Apache HTTP Server Project. http://httpd.apache.org/
16) 形態素解析システム茶筌. http://chasen.aist-nara.ac.jp/
17) SOM_PAK and LVQ_PAK. http://www.cis.hut.fi/research/som_lvq_pak.shtml

(平成 16 年 7 月 12 日受付)
(平成 16 年 10 月 4 日採録)

戸川 聡（学生会員）

1989 年阿南工業高等専門学校機械工学科卒業．同年テック情報（株）入社．1993 年四国大学情報処理教育センター勤務．2003 年同大学経営情報学部講師を併任．主にキャンパスネットワークの設計・構築・運用に従事．この間，2000 年徳島大学工学部知能情報工学科卒業．2003 年同大学大学院博士前期課程修了．現在，同大学大学院博士後期課程在学中．IEEE CS，電子情報通信学会各会員．

金西 計英（正会員）

1986 年徳島大学教育学部卒業．1988 年鳴門教育大学大学院修士課程修了．博士（工学）．1988 年関西学院大学総合教育研究室実験助手．1991 年金沢工業大学 CAI 室助手．1993 年四国大学短期大学部講師．1999 年徳島大学大学開放実践センター講師．2002 年同大学高度情報化基盤センター助教授．知的学習支援システムの研究に従事．ヒューマン・コンピュータ・インタラクションに興味を持つ．教育システム情報学会編集委員．電子情報通信学会，日本教育学会，日本ソフトウェア科学会，日本認知科学会各会員．

矢野 米雄（正会員）

1969 年大阪大学工学部通信工学科卒業．1974 年同大学大学院工学研究科博士課程修了．工学博士．同年徳島大学工学部助手．現在，同大学工学部知能情報工学科教授，工学部長．その間，1979〜1980 年，1981 年に米国イリノイ大学 CERL 客員研究員．また，1997 年カナダアカディア大学，米国マサチューセッツ大学文部省在外研究員．教育工学，e-Learning，知的 CAI の研究に従事．1998 年教育システム情報学会論文賞，1999 年 Top Paper Award（WebNet99），2001 年人工知能学会研究奨励賞，2002 年エレキテル尾崎財団源内大賞を受賞．日本教育工学会，教育システム情報学会，AACE，IEEE 各会員．