第2回「ネットワーク生態系と空間デザイン」シンポジウム

2004年3月5日　於　ATR人間情報科学研究所

Webサイトの競合ダイナミックス

木村　昌弘　　斉藤　和巳　　上田　修功

NTT コミュニケーション科学基礎研究所

World Wide Web

- ハイパーリンクで繋がれたサイト群の巨大なネットワーク
- 巨大な情報空間であり，コミュニケーションの新重要メディア

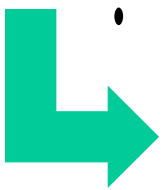
→ ネットワーク生態系の最重要例の一つ

社会学や経済学の観点では，

- サイト管理者がユーザに情報商品を提供する

グローバルなマーケット

- あるサイトへの訪問者数はそのサイトの成功指標



「類似したサービスを提供するWebサイト群は，
各自のサイトへの訪問者数を増やすために，
互いに**競合**している．」

重要

マーケットを形成するWebサイト群に対して，

各サイトの**訪問者数の変動** → モデリング＆解析

競合ダイナミックスの観点

Webマーケットの競合構造

最新ニュースを提供するWebサイト集合の一例

競合はあまりない

“asahi.com”
“NIKKEI NET”
“YOMIURI ON-LINE”

主に日本人ユーザ集団に対して競合

“The New York Times on the Web”
“washingtonpost.com”

主にアメリカ人ユーザ集団に対して競合

マーケットを形成するサイト群 → **競合グループ**に分類される

重要な課題

サイト集合の競合構造を，サイト訪問者数の変動データから抽出

（サイト管理者のビジネス戦略立案への貢献）

必要性は高い

サイト競合ダイナミックスのモデリング ← 競合構造を組み込む

関連研究

Webサーファーのアクセスパターンに関する統計的規則性の発見

例) サイト毎の訪問者数分布がべき分布

定性的説明

Webユーザの振る舞いに関して：

・統計的理論　Huberman *et al* (1998), Adamic & Huberman (2000)

・エージェントモデル　Liu *et al* (to appear)

類似したサービスを提供するWebサイト群に関して：

サイト訪問者数ダイナミックスのモデル　Maurer & Huberman (2003)

定性的説明

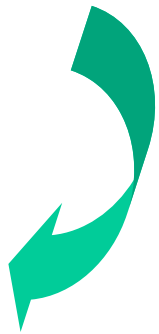
サイトの競合関係がそのマーケットの性質に及ぼす影響

観測データの予測モデル（定量的モデル）ではない

関連研究 (続き)

あるWebマーケットに対して，
各サイトの訪問者数：

観測データ　→　近未来予測



考えられる

ブラックボックスモデルの適用
(ARモデル，ニューラルネットなど)

現象の根本的な構造やメカニズムを陽に表現しない

例えば：

- マーケットの競合構造
- サイトの競合ダイナミックス

研究の目的

マーケットを形成するサイト集合に対して，

各サイトへの訪問者数の短期変動現象のモデリング

観測時系列データ　→　予測モデル（定量的モデル）：

- サイト間の相互作用の構造とメカニズムを，
  サイト集合の**競合構造**と**競合ダイナミックス**の見地から表現
- 観測時系列データを**定量的**に説明し，さらに，
  各サイトの近未来の**シェア**を**予測**

- 新たな確率的ダイナミックスモデルの提案
- その学習アルゴリズムの導出

サイト訪問者数変動のモデリング

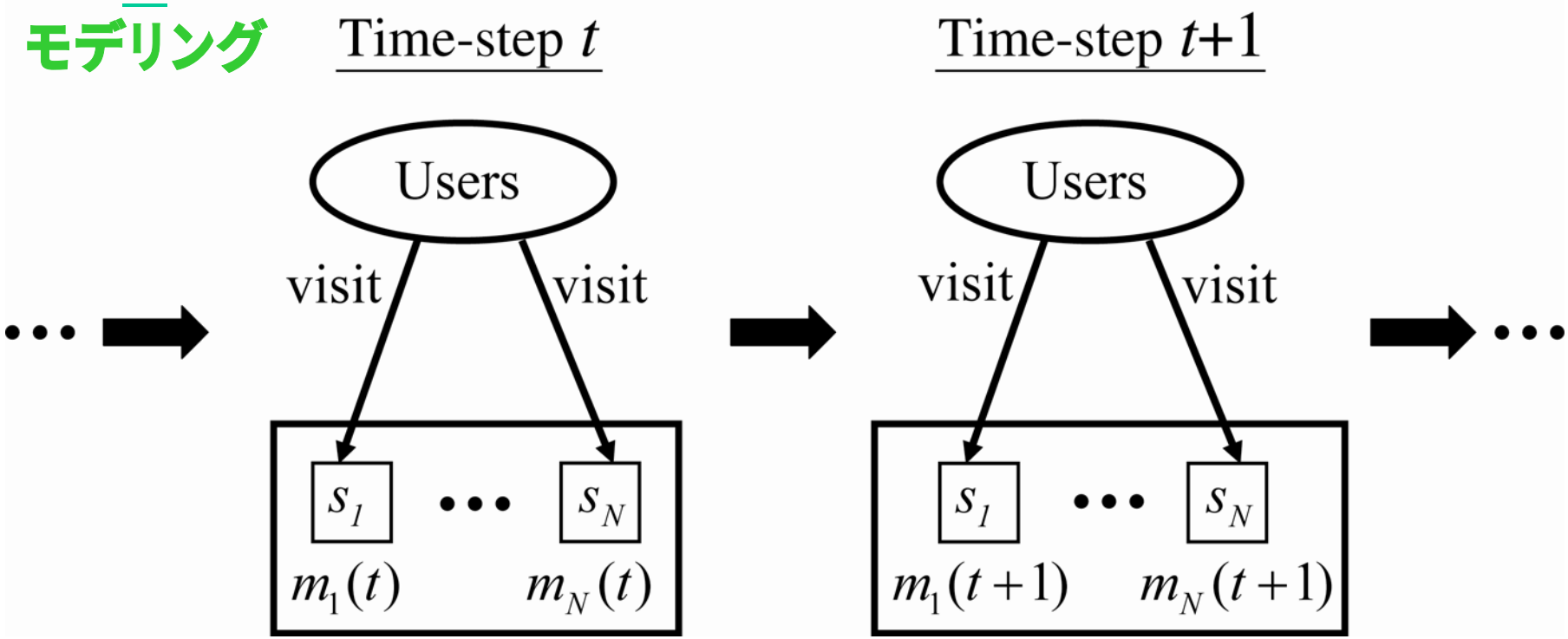
$S = \{s_1, \cdots, s_N\}$ ： マーケットを形成するサイト集合

$m_i(t)$ ： サイト s_i への時間ステップ t における訪問者数

$\boldsymbol{m}(t) = (m_1(t), \cdots, m_N(t))$ ： 時間ステップ t における
マーケット S の訪問ベクトル

ダイナミックス

モデリング



シェアダイナミックス

$M(t)=\sum_{i=1}^{N} m_i(t)$： S への時間ステップ t における総訪問者数

ダイナミックス

S を含むもっと大きい社会システムを調べて，
あらかじめ別にモデリングすべき．

$M(t)$ は与えられると**仮定**し，
各サイト s_i の**シェア**を予測する問題を考える．

$x_i(t)=m_i(t)/M(t)$： s_i の時間ステップ t における**シェア**

S の時間ステップ t におけるシェアベクトル

$$\boldsymbol{x}(t)=\left(x_1(t),\cdots,x_N(t)\right)\in \Delta^{N-1}$$

単体上の確率ダイナミックスのモデリングが必要

S の競合構造

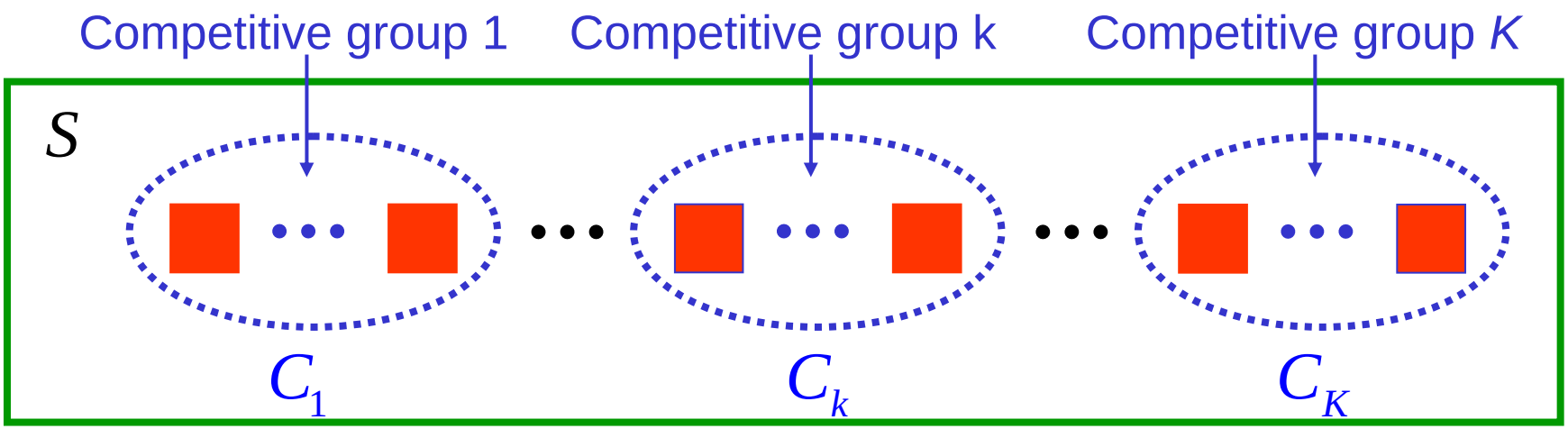
仮定：

S は K 個の競合グループに分割される

$$S = \bigcup_{k=1}^{K} C_k, \quad \text{(disjoint union)}$$

$$C_k = \{ s_i ;\ \lambda_i = k \} : \text{競合グループ（ジャンル）} \quad (k = 1, \cdots, K)$$

$$\lambda = (\lambda_1, \cdots, \lambda_N); \quad \lambda_i \in \{1, \cdots, K\},\ i = 1, \cdots, N$$



提案モデル

確率過程

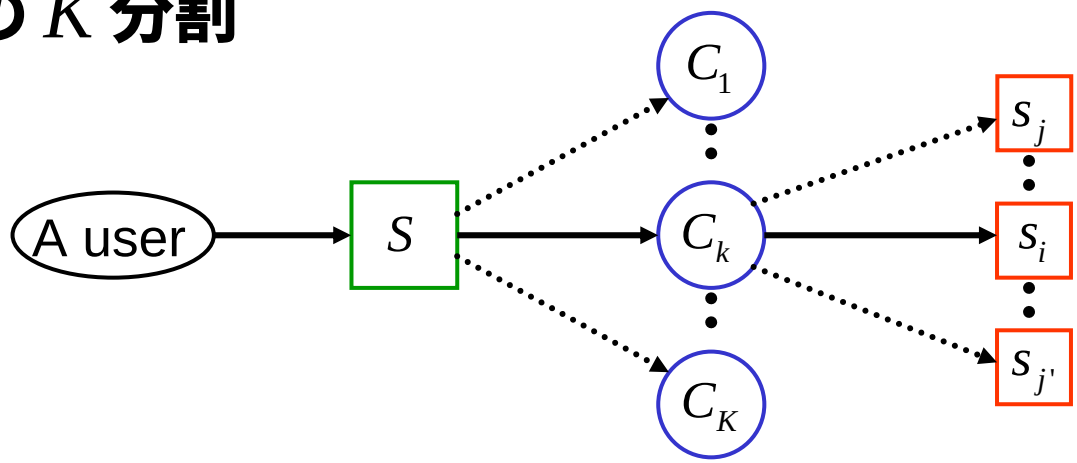
$$P\left(\boldsymbol{m}(t+1)\,|\,\boldsymbol{m}(t), M(t+1), \boldsymbol{\Theta}, \boldsymbol{\lambda}\right) \propto \prod_{i=1}^{N} \left\{P(s_i\,|\,\boldsymbol{m}(t), \boldsymbol{\Theta}, \boldsymbol{\lambda})\right\}^{m_i(t+1)}$$

(多項分布)

$\boldsymbol{\Theta} = \left((\theta_1, \cdots, \theta_K), \boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_K\right)$： **パラメータベクトル**

$\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_N)$： S **の** K **分割**

S **の一般ユーザが**
$s_i \in C_k$ **を訪問する行動**



$$P(s_i\,|\,\boldsymbol{m}(t), \boldsymbol{\Theta}, \boldsymbol{\lambda}) = P(s_i\,|\,C_k, \boldsymbol{m}(t), \boldsymbol{\Theta}, \boldsymbol{\lambda}) P(C_k\,|\,\boldsymbol{m}(t), \boldsymbol{\Theta}, \boldsymbol{\lambda})$$

ただし,

$$P(C_k\,|\,\boldsymbol{m}(t), \boldsymbol{\Theta}, \boldsymbol{\lambda}) \triangleq \theta_k$$ **時間に依らない定数**

$P(s_i \mid C_k, \boldsymbol{m}(t), \Theta, \lambda)$ の定義

$P(s_i \mid C_k, \boldsymbol{m}(t), \Theta, \lambda) \triangleq f_{k,i}(\boldsymbol{x}_k(t); \Phi_k)$ **シェアダイナミックス**

C_k 内でのシェアベクトル

$\|$

$\overline{x_{k,i}}(t+1)$　　$x_{k,i}(t)$ ：s_i の C_k 内でのシェア

定義

$$f_{k,i}(\boldsymbol{x}_k(t); \Phi_k) = (1-\xi_k)\, g_k(\boldsymbol{x}_k(t); \alpha_k, \beta_k, \eta_k) + \xi_k \frac{1}{N_k} \quad (N_k = |C_k|)$$

Replicator dynamics

Uniform dynamics

$$\frac{\overline{x_{k,i}}(t+1) - x_{k,i}(t)}{x_{k,i}(t)} = A_{k,i}(\boldsymbol{x}_k(t)) - \overline{A_k}(\boldsymbol{x}_k(t))$$

s_i の C_k 内での**魅力度**

C_k の平均魅力度

サイト魅力度の定義

s_i の C_k 内での魅力度

$$A_{k,i}(\boldsymbol{x}_k(t)) = (1-\eta_k)\,\alpha_{k,i}\,x_{k,i}(t) + \eta_k\,\beta_{k,i}\,(1-x_{k,i}(t))$$

一般受けする魅力度
(一般価値的魅力度)

専門家受けする魅力度
(希少価値的魅力度)

パラメータベクトル $\boldsymbol{\Theta}$:

魅力度: $0 < \alpha_{k,i}, \beta_{k,i} < 1; \quad \sum_i \alpha_{k,i} = \sum_i \beta_{k,i} = 1$

混合度: $0 < \eta_k < 1, \quad 0 < \xi_k < 1$

$$\rightarrow \boldsymbol{\Phi}_k = \left(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \eta_k, \xi_k\right)$$

選択率: $0 < \theta_k < 1; \quad \sum_{k=1}^{K} \theta_k = 1$

$$\rightarrow \boldsymbol{\Theta} = \left((\theta_1, \cdots, \theta_K), \boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_K\right)$$

学習アルゴリズム

$\{\boldsymbol{m}(0),\cdots,\boldsymbol{m}(T)\}$: S の訪問ベクトルの観測時系列データ

最適な確率ダイナミックスモデル

$$P\left(\boldsymbol{m}(t)\,|\,\boldsymbol{m}(t-1),M(t),\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\lambda}},\hat{K}\right),\quad (t\geq 1)$$

を推定する.

パラメータベクトル　分割ベクトル　分割数

戦略:

- **最尤推定** $\rightarrow \hat{\boldsymbol{\Theta}},\hat{\boldsymbol{\lambda}}$
- **サイトの1期先シェアの平均予測誤差の最小化** $\rightarrow \hat{K}$

Θ の推定

$\{\boldsymbol{m}(0),\cdots,\boldsymbol{m}(T)\}$, K, λ が与えられたとき,

最適パラメータベクトル $\hat{\Theta}$ は

$$L(\Theta) = -\log P(\boldsymbol{m}(0),\cdots,\boldsymbol{m}(T)) - \log P(\Theta)$$

を最小化することにより計算する.

$\Theta \in \Delta^{K-1} \times \Delta^{2N_1} \times \cdots \times \Delta^{2N_K}$ であるので,

$$P(\Theta) \propto \prod_j \Theta_j, \quad \text{(Dirichlet 分布の独立積)}$$

と仮定 → *Laplace smoothing*

$$\therefore L(\Theta) = -\sum_i a_i \log\left(\sum_j b_{ij}\Theta_j\right), \quad a_i, b_{ij} > 0.$$

- 極小解 ＝ 大域最適解
- EMアルゴリズムで効率的に計算できる

λ の推定

$\{\boldsymbol{m}(0),\cdots,\boldsymbol{m}(T)\}$, K が与えられたとき,

最適分割ベクトル $\hat{\lambda}$ は

$$\lambda \mapsto L(\Theta(\lambda)),$$

を最小化することにより計算する. ただし,

$\Theta(\lambda)$: 分割 λ に対する Θ の最適値

戦略:

K 分割 $\lambda=(\lambda_1,\cdots,\lambda_N)$, $(\lambda_i\in\{1,\cdots,K\}, i=1,\cdots,N)$

を局所的に最適な方向へ変更

⟶ $\hat{\lambda}$ を探索

予測性能の評価尺度

$\{\boldsymbol{m}(0),\cdots,\boldsymbol{m}(T)\}$:訪問ベクトルの観測時系列データ

$T_0 > 0$ を固定し，

$1 \le \forall \Delta T \le T_0$ に対して，

実験 $\rightarrow (T = 50, \quad T_0 = 10)$

$\{\boldsymbol{m}(0),\cdots,\boldsymbol{m}(T-\Delta T)\}$ を学習データとして，

時間ステップ $T-\Delta T+1$ でのシェアベクトル

$$\boldsymbol{x}(T-\Delta T+1) = \frac{\boldsymbol{m}(T-\Delta T+1)}{M(T-\Delta T+1)}$$ を予測

予測誤差

$$\mathcal{E}(\Delta T) = \frac{1}{2}\sum_{i=1}^{N}\left|\frac{m_i(T-\Delta T+1)}{M(T-\Delta T+1)} - \hat{x}_i(T-\Delta T+1)\right|.$$

観測値　　予測値

平均予測誤差 $\bar{\mathcal{E}} = \frac{1}{T_0}\sum_{\Delta T=1}^{T_0}\mathcal{E}(\Delta T)$ で性能評価

予測タスクにおける比較対象

- ガウスノイズに基づいた確率過程モデル → ✕

(∵) 必ずしも $\Delta^{N-1} \to \Delta^{N-1}$ とは写像しない

→ あるサイトのシェアを負にし得る

ARモデルやニューラルネットの単純な適用 → ✕

- **単純多項分布法 (NMM)**
- **平行移動法 (PDM)**

比較すべき単純な従来法として採用

予測タスクにおける比較対象 (続き)

- **単純多項分布法 (NMM)**

  仮定：「訪問ベクトル $m(t)$ は，ある多項分布にしたがって時間 t とは独立に生成される．」

  学習データ $\left\{ \boldsymbol{m}(0), \cdots, \boldsymbol{m}(T - \Delta T) \right\}$ から

  その多項分布を最尤推定し，

  $$\omega = \left(\omega_1, \cdots, \omega_N \right)$$ ：推定された多項分布のパラメータ

  時間ステップ $T - \Delta T + 1$ でのシェアベクトルの予測値を

  $$\hat{\boldsymbol{x}}(T - \Delta T + 1) = \omega$$

  とする．

予測タスクにおける比較対象 (続き)

- **平行移動法 (PDM)**

  学習データ $\left\{ \boldsymbol{m}(0), \cdots, \boldsymbol{m}(T - \Delta T) \right\}$ から

  時間ステップ $T - \Delta T + 1$ でのシェアベクトルを予測するのに

  時間ステップ $T - \Delta T$ でのシェアベクトルの観測値で予測する.

  すなわち,

$$\hat{\boldsymbol{x}}(T - \Delta T + 1) = \boldsymbol{x}(T - \Delta T)$$

分類性能の評価尺度

$\lambda = \{C_k\}_{k=1}^{K}$: S の真の K 分割

$\hat{\lambda} = \{\hat{C}_l\}_{l=1}^{L}$: $\{\boldsymbol{m}(0), \cdots, \boldsymbol{m}(T-\Delta T)\}$ から推定された S の K 分割

各 $\hat{C}_l$ に対して,

$C_{k[l]}$: 期間 $[0, T-\Delta T]$ にサイト集合 $C_k \cap \hat{C}_l$ を訪問した
ユーザ数が最も多かった C_k

$\hat{\lambda}$ の λ に対するmicro-averaged precision $AP(\Delta T)$

$$\frac{\text{期間 } [0, T-\Delta T] \text{に } \hat{C}_l \cap C_{k[l]} \text{ を訪問したユーザ数}}{\text{期間} [0, T-\Delta T] \text{ に訪問した総ユーザ数}} \times 100$$

分類タスクにおける比較対象

Webサイト集合を関連するサイト集合に分類

→ しばしば要求される

多くの研究

テキストやリンクの静的情報を使って，
トピック的に関連するサイト群やコミュニティーを抽出する研究

適当ではない

ユーザに関する動的情報に基づいた競合サイトのグループ化

一方，経済物理学の分野で，Mantegnaは：

「株価の時系列データから同業種の会社群を抽出することに成功」

同一視

サイトの訪問者数変動 ⇔ 会社の株価変動

比較対象として，Mantegnaの手法を考える

分類タスクにおける比較対象 (続き)

比較対象とする従来法

1. Mantegnaの手法を用いて,
   学習データ $\{ \boldsymbol{m}(0), \cdots, \boldsymbol{m}(T-\Delta T) \}$ に基づいて,
   各サイトを球面上の点と同一視する.

$$s_i \leftrightarrow \boldsymbol{w}_i = \left(w_i(1), \cdots, w_i(T-\Delta T) \right), \quad (i = 1, \cdots, N),$$

   ここに,

$$w_i(t) = \left(y_i(t) - \mu_i \right) / \sigma_{i,}$$

$$y_i(t) = \log m_i(t) - \log m_i(t-1).$$

2. Spherical K-means法を用いて,

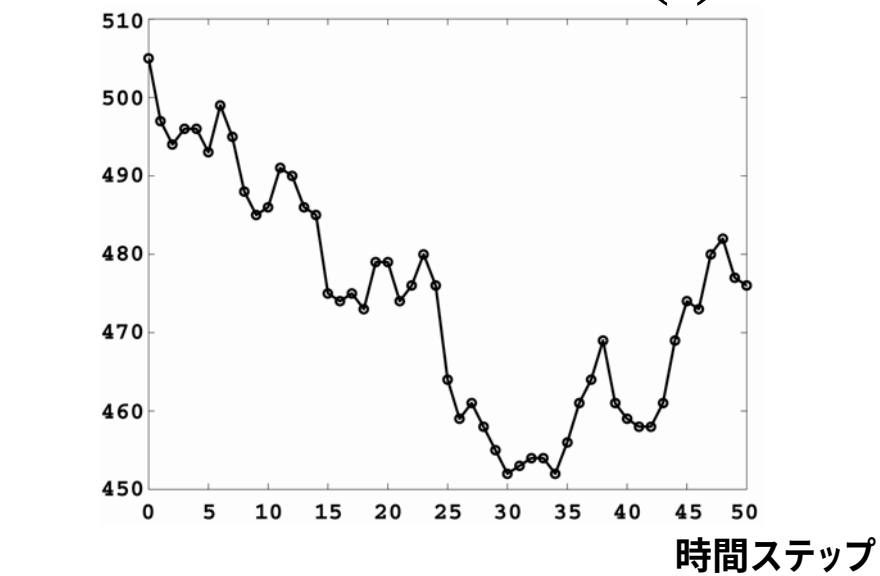
   $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N$ をK分割する.

   $\Leftrightarrow$ $s_1, \cdots, s_N$ をK分割する.

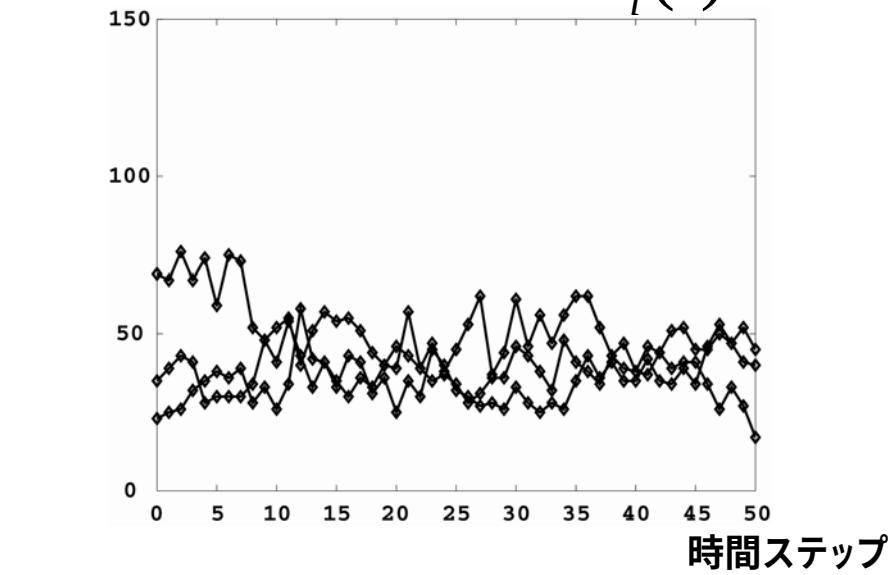
注)
Mantegnaのオリジナル:
Euclid距離に基づいて
デンドログラムを構築

人工データによる実験評価（１０サイト）

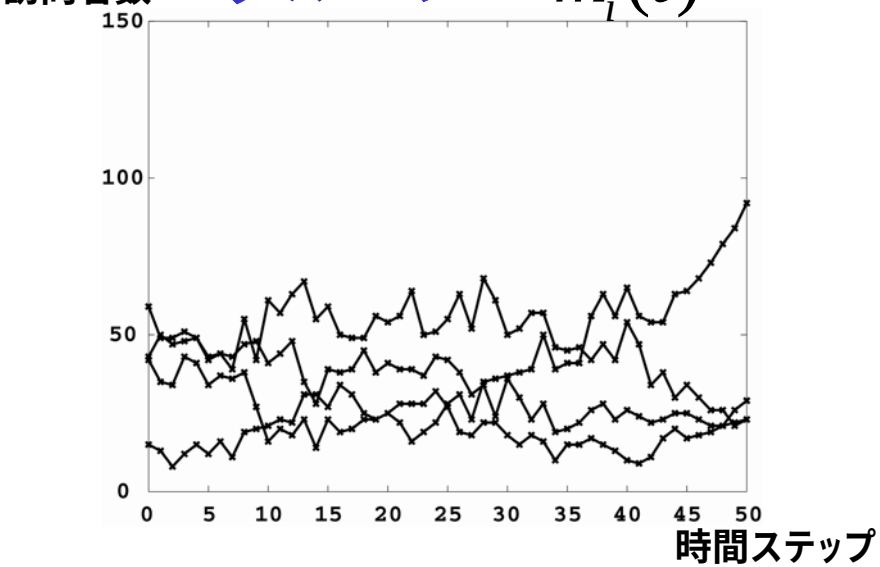
総訪問者数 $M(t)$



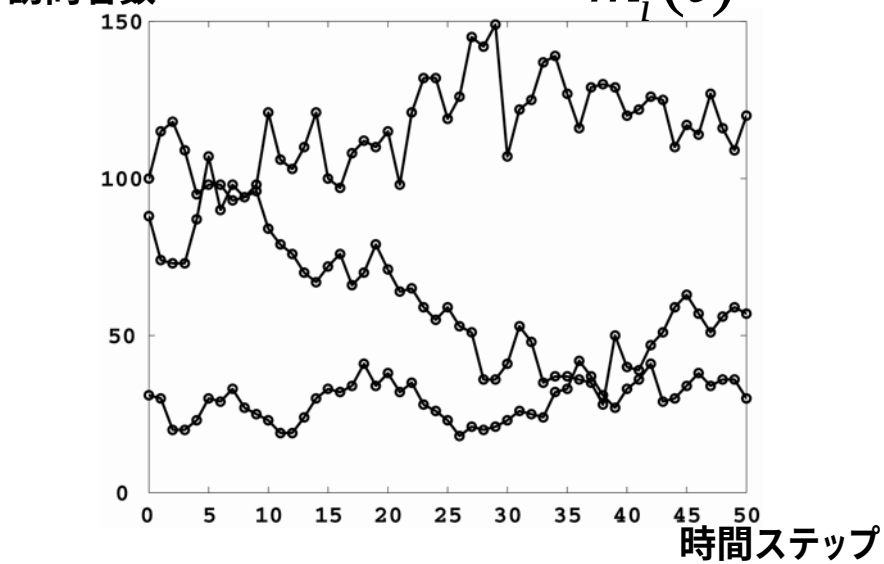
グループ1 $m_i(t)$



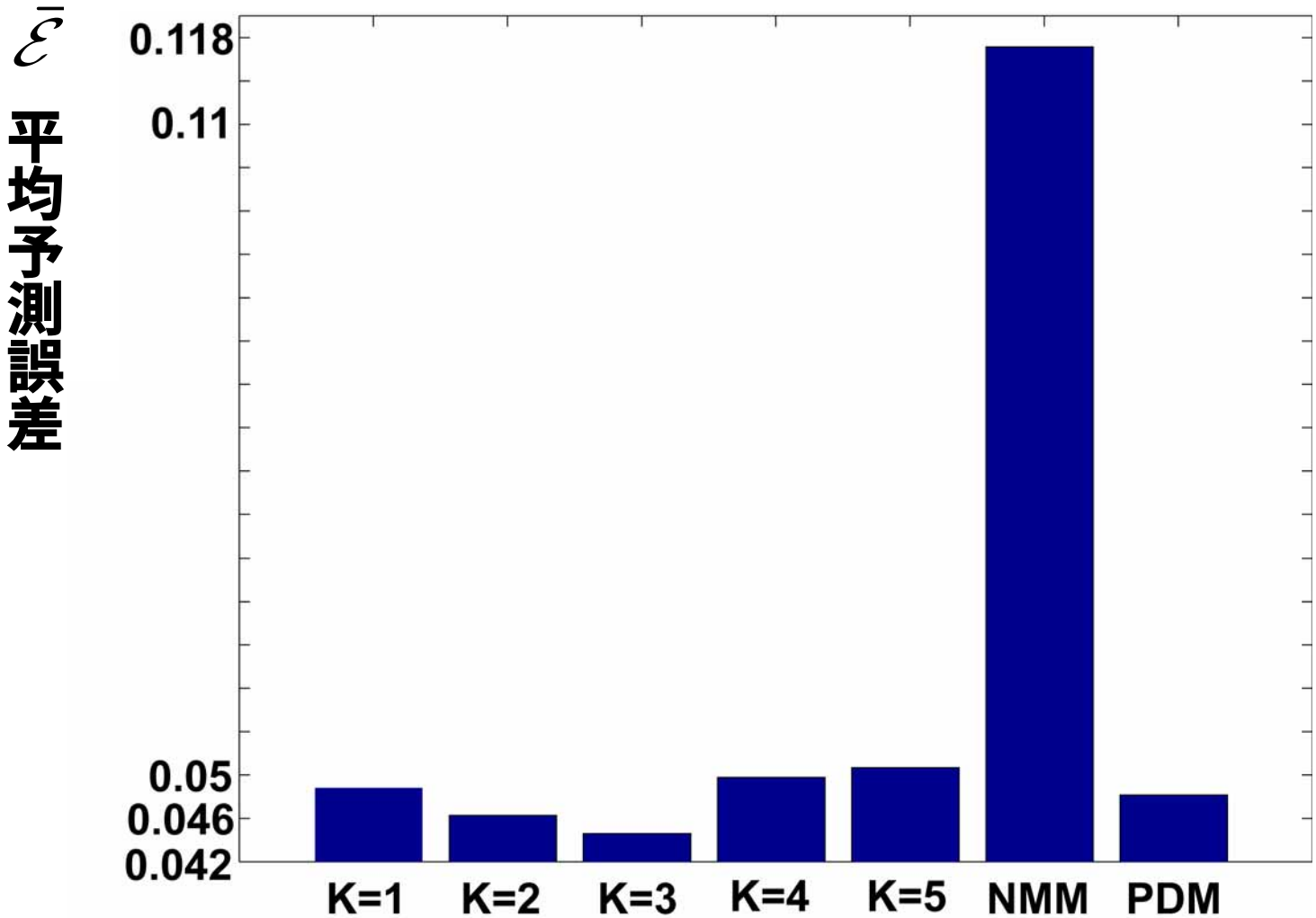
グループ2 $m_i(t)$



グループ3 $m_i(t)$



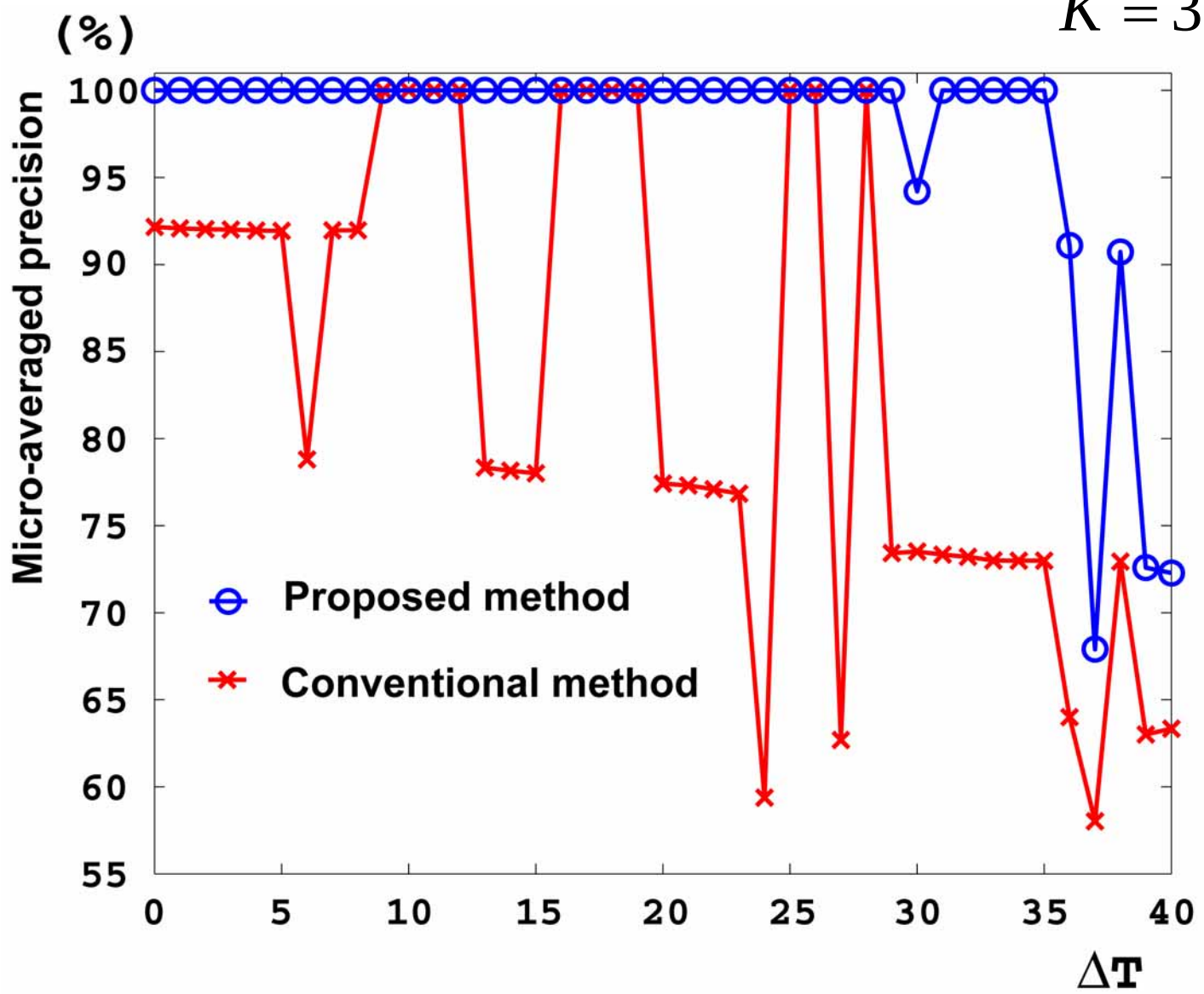
予測性能の比較 (人工データ)



分類性能の比較（人工データ）

$K = 3$

$AP(\Delta T)$



注） $\{\boldsymbol{m}(0), \cdots, \boldsymbol{m}(T - \Delta T)\} \rightarrow AP(\Delta T)$

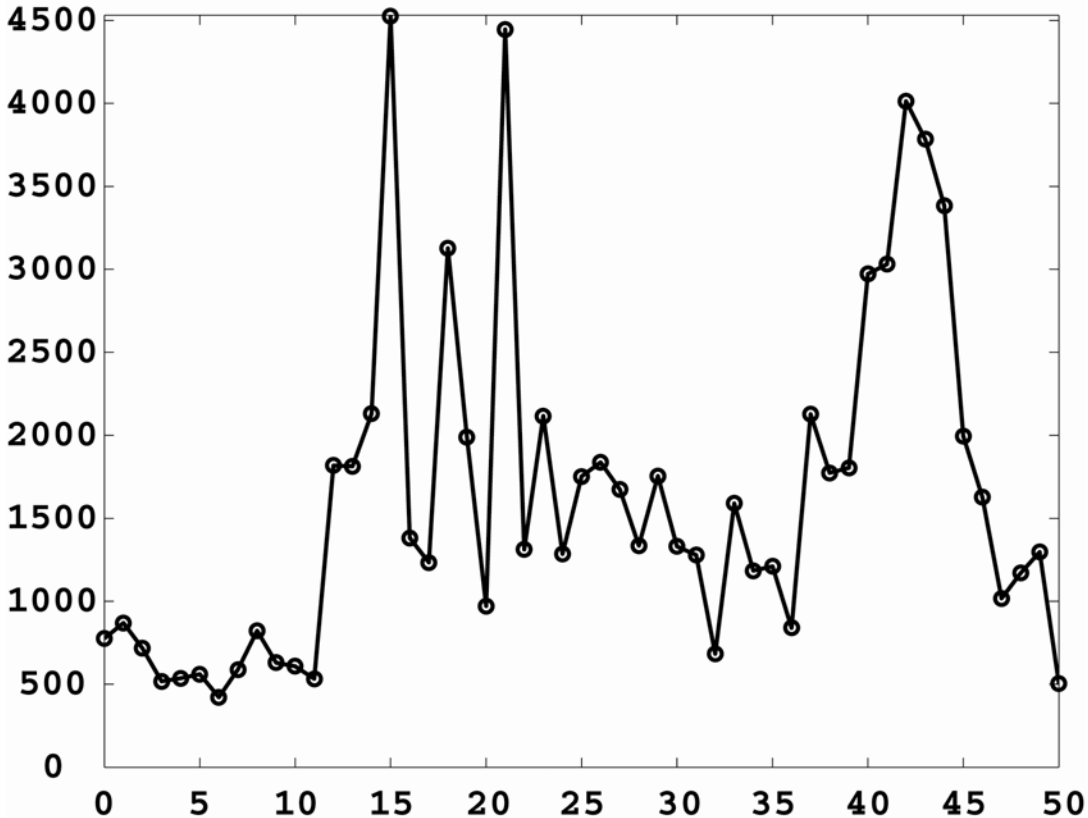
実Webデータによる実験評価

ストリーミングビデオコンテンツを提供する，日本の20Webサイトの利用ログのデータ

NTTが東京ニュース通信社と共同で2002年に実施した「ブロードバンド番組ガイド」の試行サービス実験におけるデータ

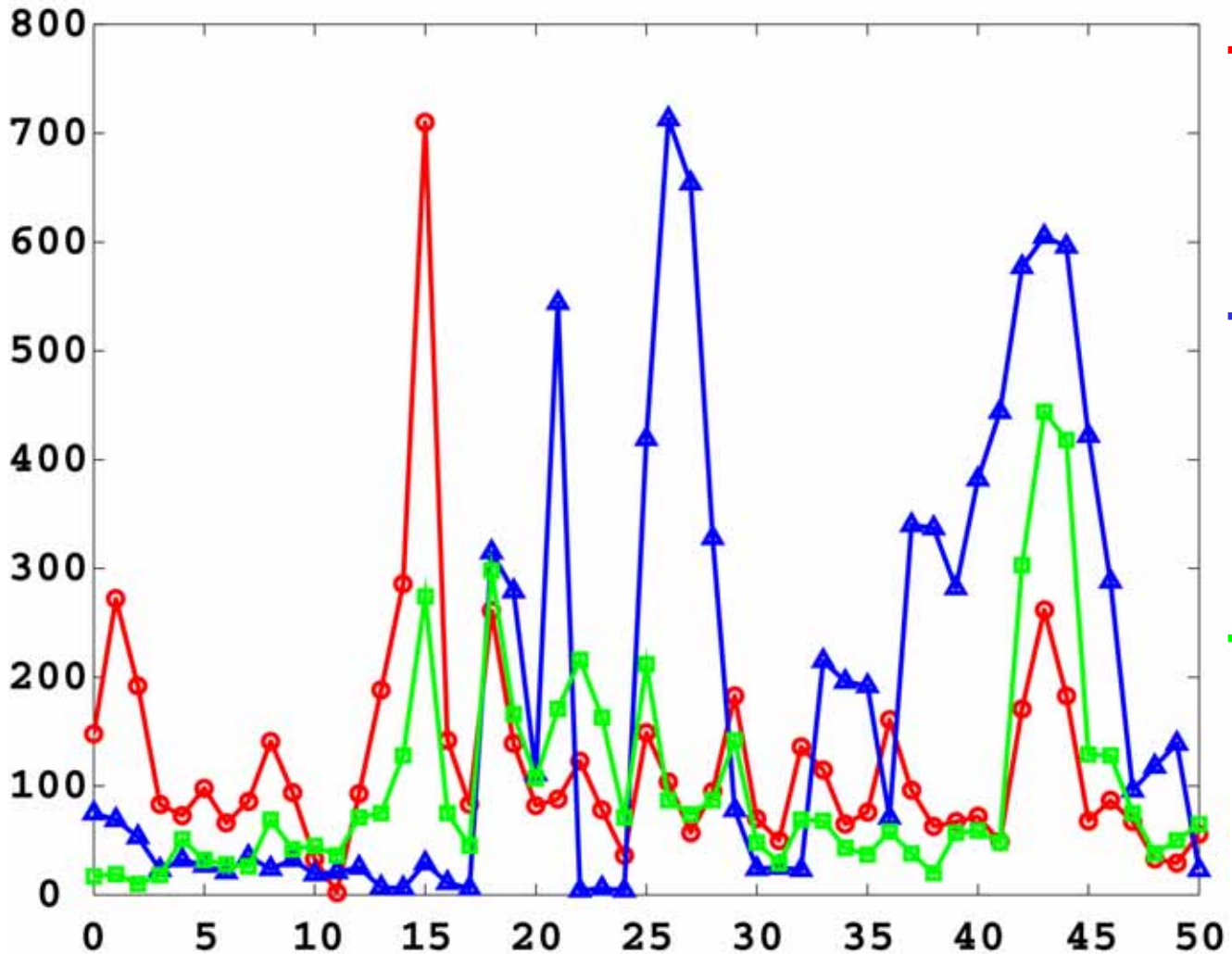
安部，宮原，林，外村，
散策型コンテンツガイドシステム「AssociaGuide」：システムおよび「ブロードバンド番組ガイド」試行サービス実験概要，
映像メディア学会技術報告，
26(81)，1－4，2002.

総訪問者数の変動 $M(t)$



サイト訪問者数の変動例 (実データ)

訪問者数　(人)　　$m_i(t)$



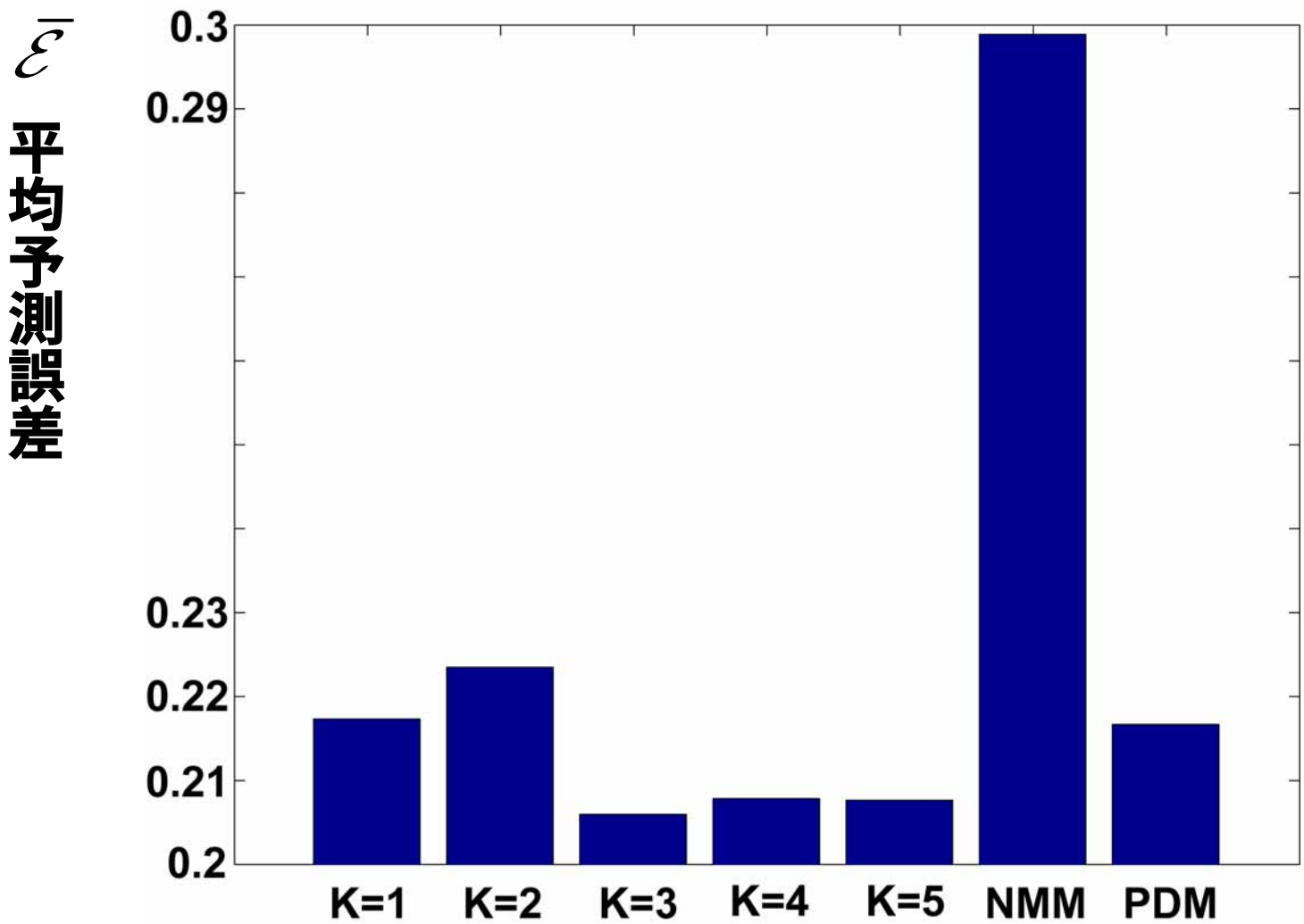
インターネットTV
のサイト
【衛星テレビ局】

(芸術的) ドラマ
のサイト
【古くからの
映画会社】

総合情報サイト
【ITニュース社】

時間ステップ　(1ステップ＝2日)

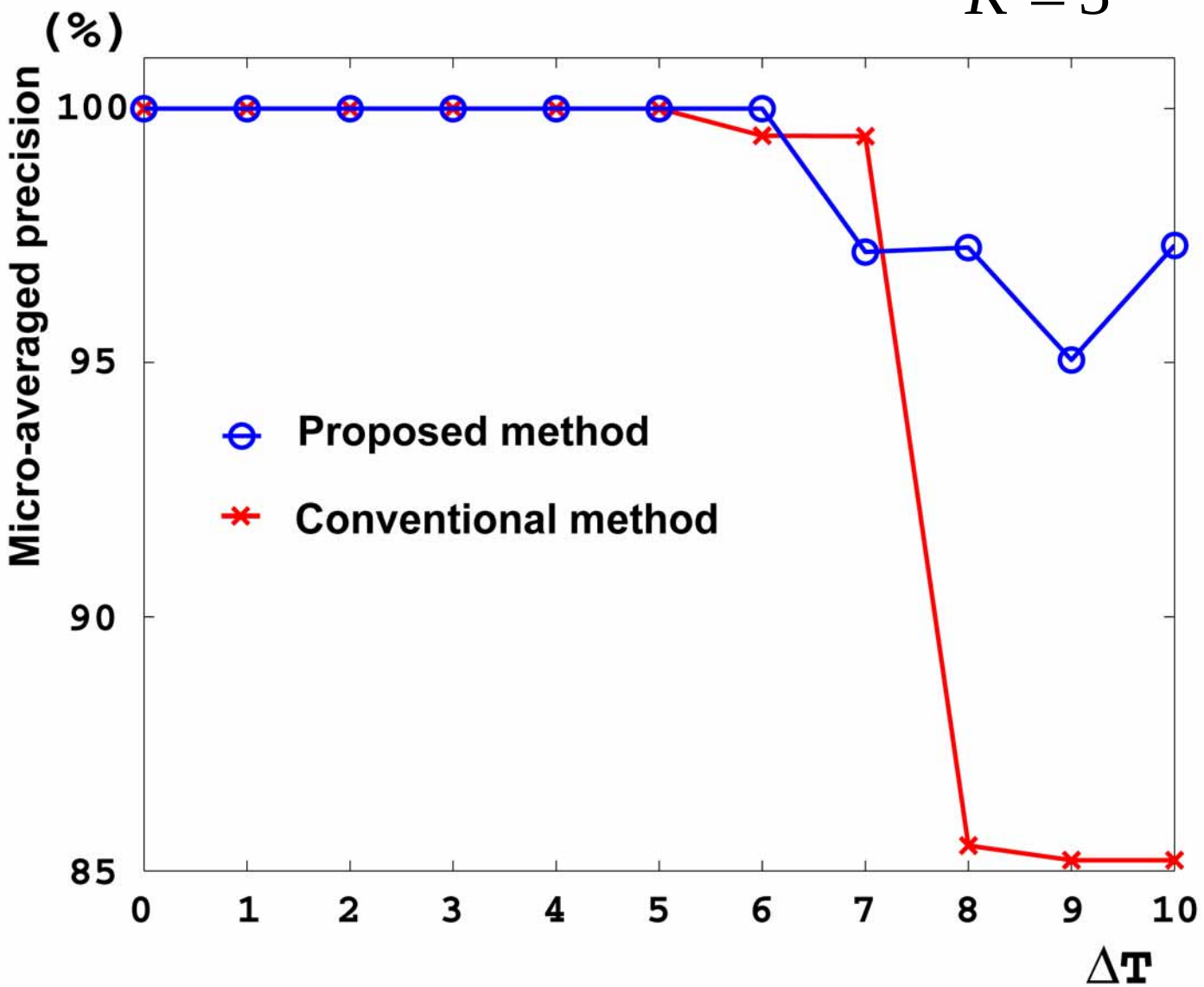
予測性能の比較 (実データ)



分類性能の比較 (実データ)

$K = 3$

$AP(\Delta T)$



注) $\{\boldsymbol{m}(0), \cdots, \boldsymbol{m}(T - \Delta T)\} \rightarrow AP(\Delta T)$

定性的評価　(提案法による分類結果)

(1)　ポピュラーエンターテーメントサイト群

1) ポピュラー音楽ビデオサイト
2) 衛星テレビ局運営のインターネットTVサイト
3) 渋谷ライフスタイル･ショッピング情報サイト
4) イベント情報サイト
5) スポーツエンターテーメントサイト
6) 卓球情報サイト
7) インタラクティブドラマサイト

(2) 芸術的趣味サイト群

8) 映像制作会社運営の次世代スポーツと古典芸能情報サイト
9) エコ商品情報サイト
10) 音楽情報サイト
11) ラジオ局運営のインターネット放送サイト
12) 地方のインターネット放送サイト
13) 古くからの映画会社運営の芸術的ドラマサイト
14) ミステリードラマサイト

(3) 知識提供サイト群

15) 新聞社運営の総合情報サイト
16) ITニュース社運営の総合情報サイト
17) 通信会社運営の総合情報サイト
18) 地方のIT会社が運営の総合情報サイト
19) 子供の教育サイト
20) 育毛情報サイト

まとめ

1．　マーケットを形成するWebサイト集合における

訪問者数の変動を，競合ダイナミックスの観点からモデリング

- 競合構造に基づいたダイナミックスモデルの提案：

多項分布の確率的混合モデル

～ 多項分布のパラメータ値が
レプリケータ方程式で時間発展

- 学習アルゴリズムの導出：

観測時系列データ　→
競合サイトのグループを同定
各グループのレプリケータ方程式を推定

2．　予測モデル（定量的モデル）としての有効性の検証