筑波大学　研究談話会

大規模なウェブの構造とその変遷を探る

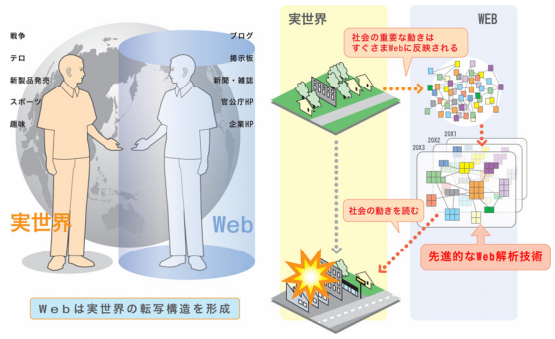
2007年11月13日

生産技術研究所
豊田正史

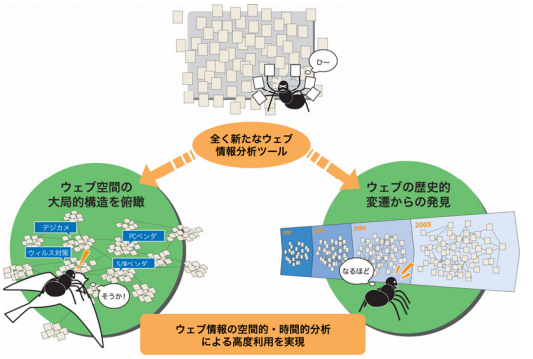
研究対象としてのWeb

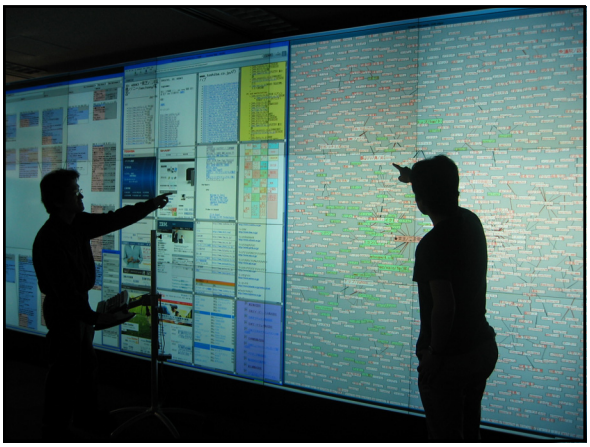
- 膨大な文書集合
  - 200億を超えるテキスト・画像・動画(Yahoo!発表2005/8)
  - 自然言語処理、情報検索、情報抽出、テキストマイニング
- **膨大なグラフ構造**
  - 文書＝ノード、リンク＝エッジの膨大かつ疎な有向グラフ
  - グラフ理論(次数、直径、進化モデル)、情報検索への応用、グラフマイニング
- **動的**
  - 持続的な成長(サーバ数は2000年から年平均36%増加 米Netcraft社)
    無数の著者が日々文書を生成する一方、消滅する文書も多い。
  - 時系列解析(成長率、内容の変化、構造の変化)、社会学
- サービス提供の場
  - 広告、通信販売、メール、ブログ、写真共有、企業間取引
  - XML、Webサービス、セキュリティ、経済学

実社会の射影としてのウェブ



目的：ウェブ情報の高度利用システムの構築（WEBの時空間解析）





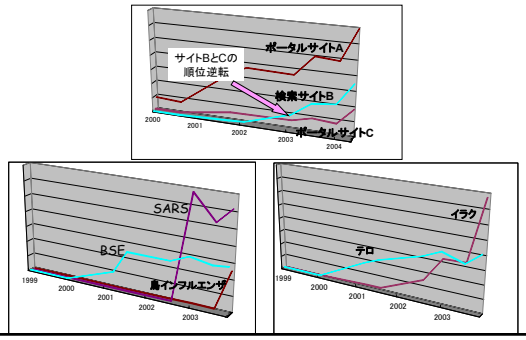
あらまし

- ウェブアーカイブ基盤
  - アーカイブの概要と簡単なアプリケーション
- ウェブ空間の構造俯瞰
  - リンク解析によるウェブの地図
  - Yahoo!との比較
- ウェブの時系列分析
  - ウェブ空間構造の時系列変化を可視化

アーカイブの簡単なアプリケーション

- URLを指定して、ページの編集履歴を見る
- 時系列検索エンジン
  - 検索ヒットページ数の推移
  - ランキングの時系列変化
  - 各時期の新規ページ提示

検索に対するヒットページ数の推移



あらまし

- ウェブアーカイブ基盤
  - アーカイブの概要と簡単なアプリケーション
- ウェブ空間の構造俯瞰
  - リンク解析によるウェブの地図
  - Yahoo!との比較
- ウェブの時系列分析
  - ウェブ空間構造の時系列変化を可視化

ウェブ空間の構造俯瞰
～コミュニティチャート～

ウェブコミュニティとは

同じトピックに関心を持つ人々または組織が作成したウェブページの集まり

例1　千葉ロッテマリーンズファンのコミュニティ

例2　PCメーカのコミュニティ

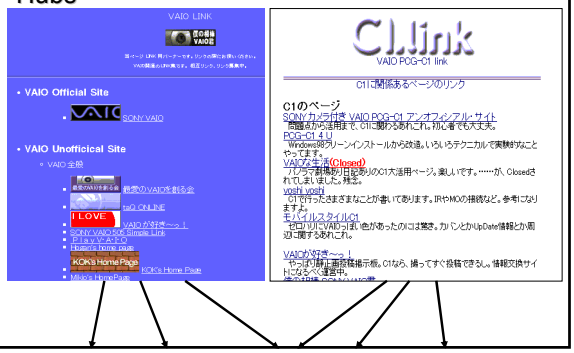
HITS [Kleinberg '97]

以下の観測を基にコミュニティを発見する

- ハイパーリンクはリンク先のページを推薦する
  - お勧めしないページはわざわざリンクしない
- 同種類のハイパーリンクは一箇所にまとめられることがある
  - ブックマーク、お気に入り、リンク集、ポータルサイト、相互リンク、お友達リンク、etc.

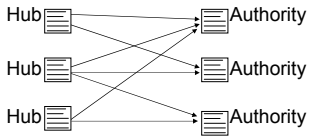
HubとAuthorityの例

Hubs



HubとAuthority

- 適当なウェブの部分グラフから良いhubとauthorityを抽出する
  - Hub: 多くの良いauthorityを指しているリンク集
  - Authority: 多くの良いhubから指されているページ

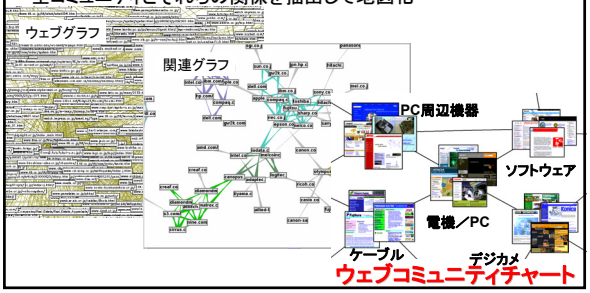


良いAuthorityとHubの集まりをコミュニティと呼ぶ

ウェブ空間の構造俯瞰

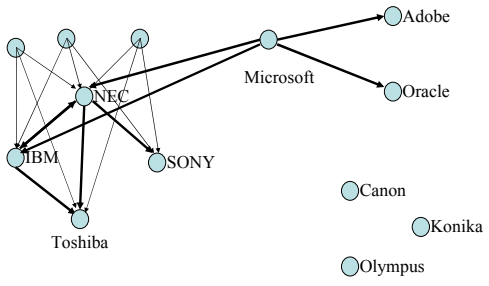
ウェブコミュニティチャート[ACM Hypertext 2001]

- ある話題に関する有用なページの集合はウェブグラフ上で稠密な構造を持つ(ウェブコミュニティ)
- 全コミュニティとそれらの関係を抽出して地図化

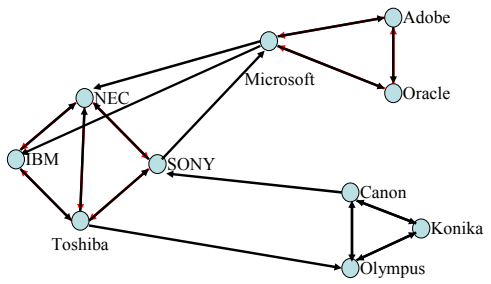


Relationship graph

For each page, find authorities in the neighborhood, and make edges from the page to authorities

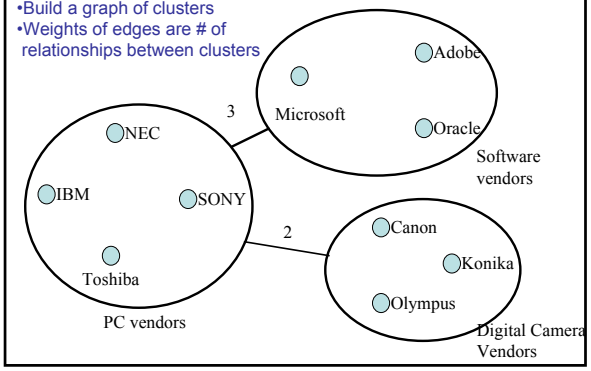


Relationship Graph



Web Community Chart

•Build a graph of clusters
•Weights of edges are # of relationships between clusters



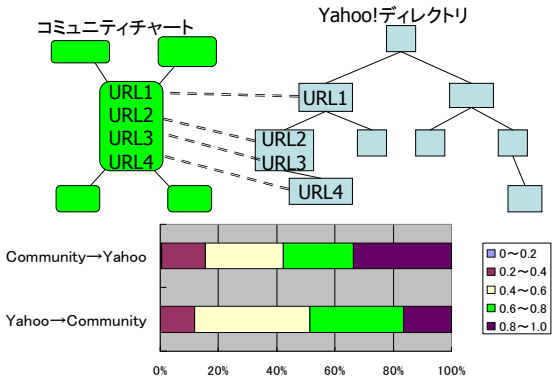
コミュニティチャートとYahoo!の比較 [吉田 2003]

- 共有URL数(2002年のデータを使用)

Yahoo!の重複を取り除いたURL	177,000
ウェブコミュニティチャートのURL	1,000,000
共有URL	81,000

- 比較対象とするコミュニティとディレクトリ
  - 共有部分内においてURL数5以上のもの
    - 4079コミュニティ(33930URL 平均8.13)
    - 4965ディレクトリ(63757URL 平均12.84)

チャートとYahoo!の類似度



応用研究

- 地球環境ポータル構築の試み[菊池(高知大)]
  - DEWS2001
- ジェンダー関連ポータルサイト構築[増永,小山(お茶女)]
  - 重点研究「グローバル化とジェンダー規範」2000～2001
- Web Community Browser [福地(東工大)]
  - DEWS2002, WISS2002, FIT2002
- ウェブディレクトリとの比較[吉田]
  - DEWS2003, TOD22
- 大域ウェブアクセスログ解析[大塚]
  - TOD20, DEXA2004
- リンク解析による全文検索エンジンの精度向上[RICOH]
  - NTCIR3 Web

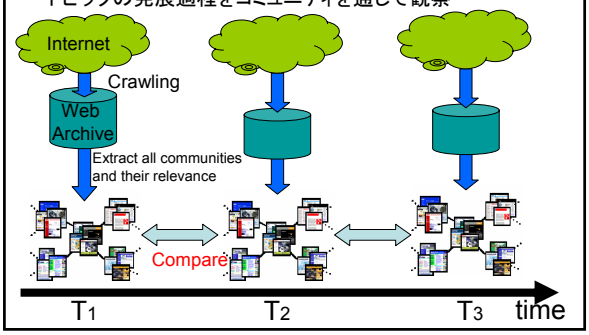
あらまし

- ウェブアーカイブ基盤
  - アーカイブの概要と簡単なアプリケーション
- ウェブ空間の構造俯瞰
  - リンク解析によるウェブの地図
  - Yahoo!との比較
- ウェブの時系列分析
  - ウェブ空間構造の時系列変化を可視化

ウェブの時系列分析

[ACM Hypertext 03]

- 定期的にウェブを大規模収集
- トピックの発展過程をコミュニティを通して観察

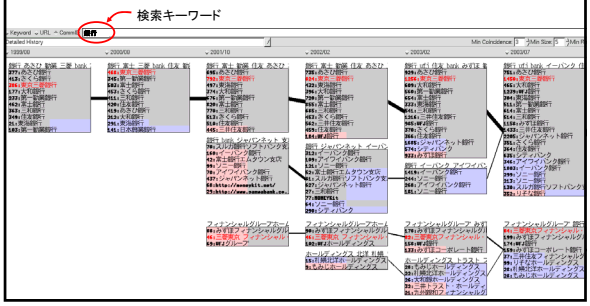


成果②ウェブの時系列分析

～銀行業界の変遷～

e-Society 文部科学省リーディングプロジェクト

- インターネット銀行の出現と世間への浸透
- 合併した銀行の出現：三井住友、UFJ、みずほ、りそな



成果②ウェブの時系列分析

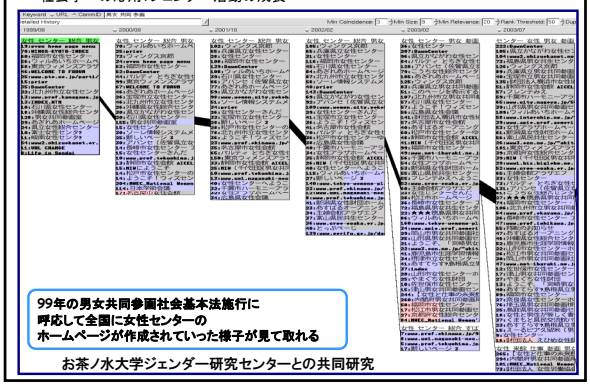
～社会現象による話題の爆発的発生～

e-Society 文部科学省リーディングプロジェクト

ウェブの時系列分析

～社会学への応用：ジェンダー活動の成長～

e-Society 文部科学省リーディングプロジェクト



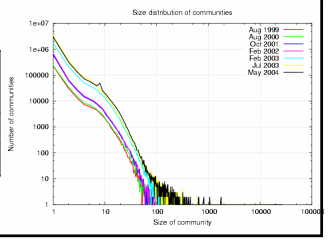
お茶ノ水大学ジェンダー研究センターとの共同研究

データセット

- 日本中のウェブサイトからロボットを用いて収集したアーカイブ 1999～2004の7回分

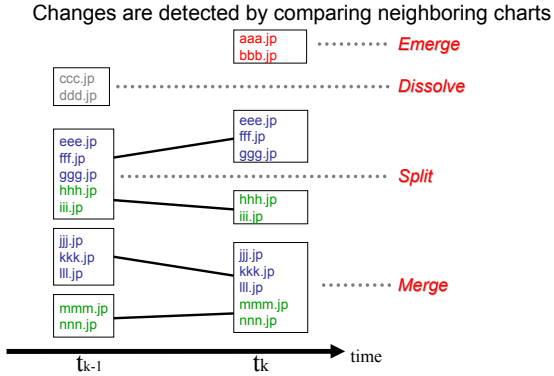
データセット詳細

Year	#Pages	#seeds	#comms
1999/8	17M	671K	83K
2000/8	17M	741K	94K
2001/10	40M	1431K	158K
2002/2	45M	1583K	171K
2003/2	66M	4646K	554K
2003/07	97M	7870K	874K
2004/05	96M	8192K	849K



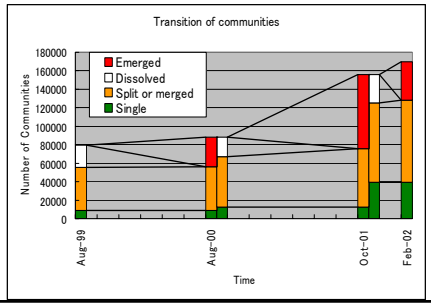
Types of Changes

Changes are detected by comparing neighboring charts



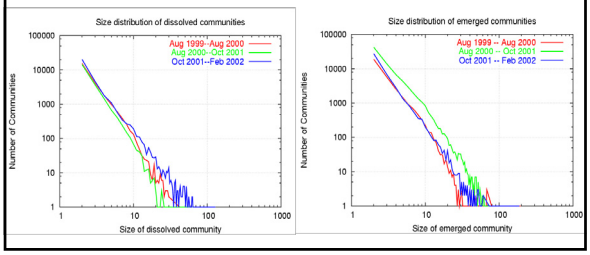
Types of Changes

- Structure of communities changes dynamically
- How the size distribution is kept unchanged?



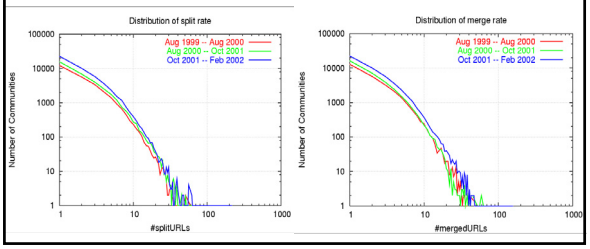
Emerged and Dissolved Communities

- Both size distributions follow the power-law
- Both exponents are greater than ones in size distribution of all communities



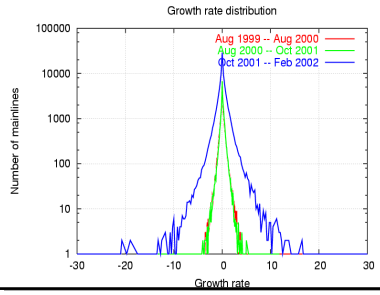
Split and Merged Communities

- # of split and merged URLs also follow the power-law, and have clear symmetry



Grown and Shrunken Communities

- Growth rate have clear y-axis symmetry



Evolution Metrics

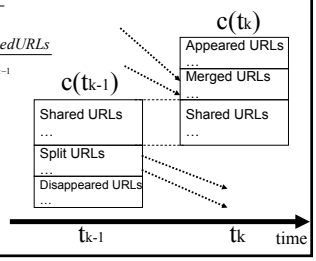
Growth rate: $\frac{\#c(t_k) - \#c(t_{k-1})}{t_k - t_{k-1}}$

Novelty: $\frac{\#appearedURLs}{t_k - t_{k-1}}$

Disappearance rate: $\frac{\#disappearedURLs}{t_k - t_{k-1}}$

Split rate: $\frac{\#splitURLs}{t_k - t_{k-1}}$

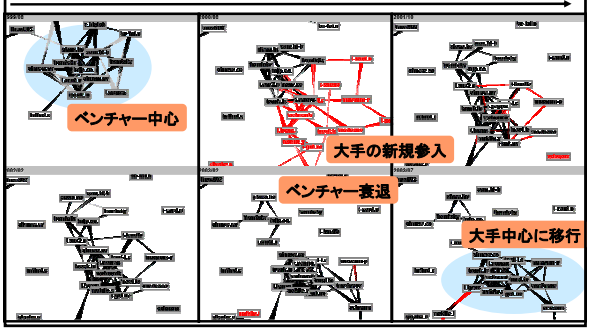
Merge rate: $\frac{\#mergedURLs}{t_k - t_{k-1}}$



ウェブの時空間分析 [ACM Hypertext 05]

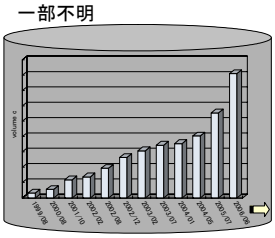
e-Society 文部科学省リーディングプロジェクト

空間＋時間分析：コミュニティの変遷
（例：i-mode検索サイト）

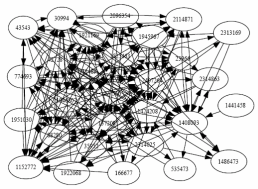


定量評価基盤構築の課題

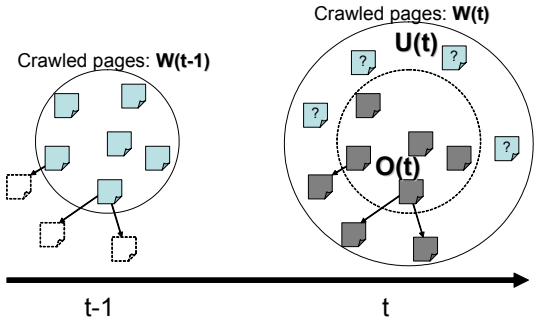
アーカイブの不完全性
・全空間の完全収集は不可能
・ページの作成・消滅時間が一部不明



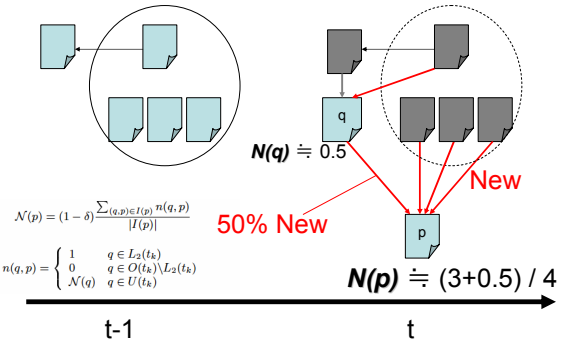
スパム・ミラーサイト
・意図的なテキスト・リンク操作の増加 (全サイトの9%～25%)
・22%～29%のページがミラー



ページの新規性推定手法 [WWW2006]



Novelty Measure



$$\mathcal{N}(p) = (1-\delta)\frac{\sum_{(q,p)\in I(p)} n(q,p)}{|I(p)|}$$

$$n(q,p) = \begin{cases} 1 & q \in L_2(t_k) \\ 0 & q \in O(t_k) \backslash L_2(t_k) \\ \mathcal{N}(q) & q \in U(t_k) \end{cases}$$

「生協の白石さん」

WWW2006: Toyoda,Kitsuregawa

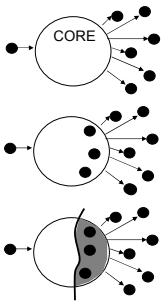
A Large-Scale Study of Link Spam Detection by Graph Algorithms

Hiroo Saito — University of Tokyo. JST, ERATO
Masashi Toyoda — University of Tokyo
Masaru Kitsuregawa — University of Tokyo
Kazuyuki Aihara — University of Tokyo. JST, ERATO

AIRWEB'07, May 8, 2007

Outline

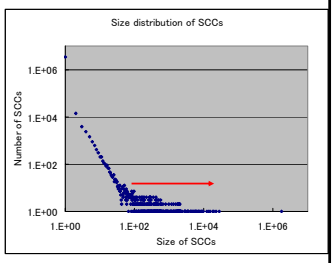
- Propose a link farm detection method using graph algorithms
- Distribution of detected link farms in the Web graph structure



1. SCC decomposition
   Around the largest SCC (CORE), large SCCs are link farms
2. Maximal clique enumeration
   Link farms in CORE can be found as maximal cliques
3. Minimum cut
   Link farms are expanded by min-cut. How many links for cutting them out?

SCC decomposition

- Size distribution follows the power-law ($1 \leqq n \leqq 100$) with a long and thick tail
- Large SCCs are spams ($100 < n$)
  - 552 SCCs, 0.57M sites
  - 550 sample sites

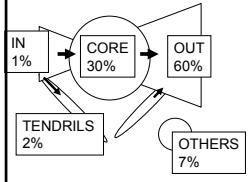


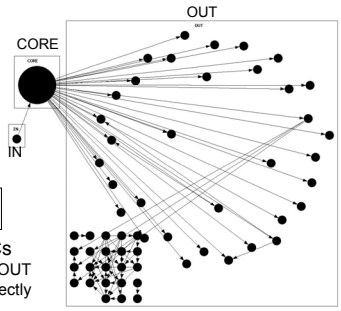
Sampling results

	spam	suspicious	non-spam
#sites	527	23	0
ratio (%)	95.8	4.2	0

Distribution of SCCs in the bow tie

- Bow-tie structure [Broder et al. 2000]





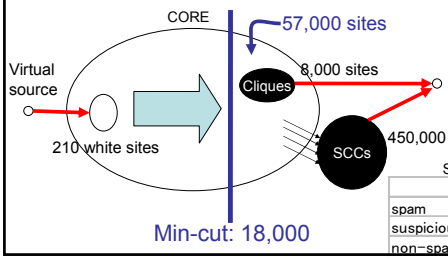
- Distribution of large SCCs
  - 450 / 552 (81%) SCCs in OUT
  - 385 / 450 (85%) SCCs directly connected to CORE
- CORE has many spam sites connecting to them

SCCs whose size are larger than 1,000

Minimum cut

- How many spam sites around large SCCs and cliques?
- How many links for cutting off spam sites?

Apply max-flow / min-cut on the directed site graph



Min-cut: 18,000

Sampling result

	number	ratio (%)
spam	459	94.3
suspicious	27	5.5
non-spam	1	0.2

今後の展開

- より連続的な構造進化の解析
  - 収集間隔の短縮（月、週、毎日程度まで）
- 自然言語処理の導入による各手法の発展
  - コミュニティ抽出の精度向上
  - アーカイブ全文検索を用いたより詳細な話題伝播分析・評判分析
- 社会学、マーケティングへの応用
  - お茶の水女子大ジェンダー研との研究は継続中
  - 他にも色々と進行中