SIG-KBS/KBSE研究会

大規模Webアーカイブの時空間分析とその実際

東京大学　生産技術研究所
戦略情報融合国際研究センター
豊田正史

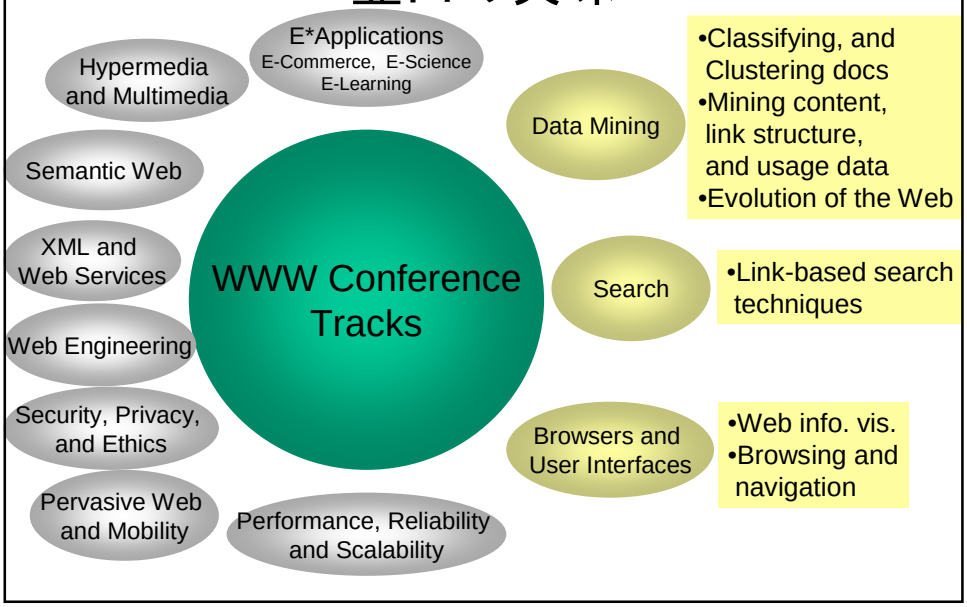
自己紹介

1996年　東京工業大学　情報理工学研究科
　　数理・計算科学専攻修了
1999年　東京工業大学　情報理工学研究科
　　数理・計算科学専攻　博士(理学)号取得
　　ズームを用いた閲覧・編集・検索ユーザインタフェース
同年　　東京大学生産技術研究所ポスドク
2004年　同特任助教授
2006年　同助教授
　　主にWebに関する研究に従事

研究対象としてのWeb

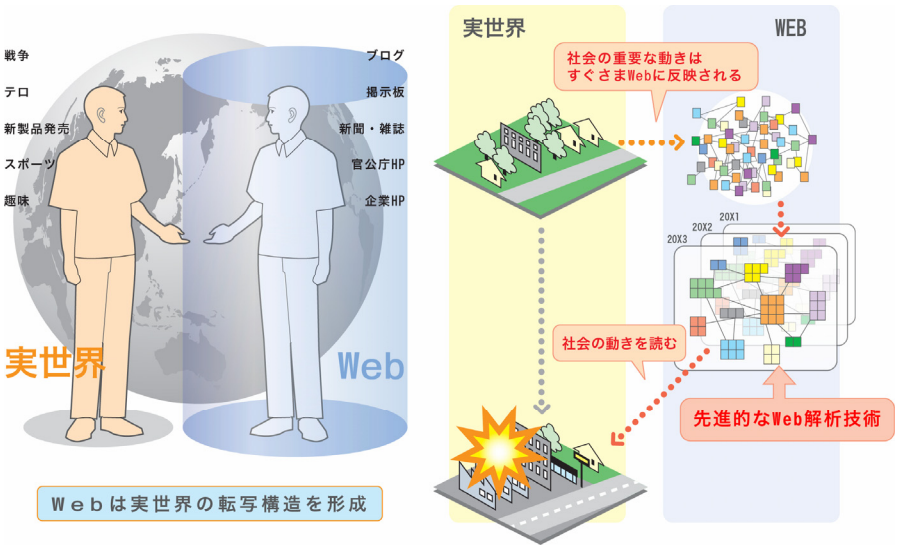
- 膨大な文書集合
  - 200億を超えるテキスト･画像･動画(Yahoo!発表2005/8)
    - 自然言語処理、情報検索、情報抽出、テキストマイニング
- **膨大なグラフ構造**
  - 文書＝ノード、リンク＝エッジの膨大かつ疎な有向グラフ
    - グラフ理論(次数、直径、進化モデル)、情報検索への応用、グラフマイニング
- **動的**
  - 持続的な成長(サーバ数は2000年から年平均36%増加 米Netcraft社)
    無数の著者が日々文書を生成する一方、消滅する文書も多い。
    - 時系列解析(成長率、内容の変化、構造の変化)、社会学
- サービス提供の場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

Webに関する研究領域と豊田の興味



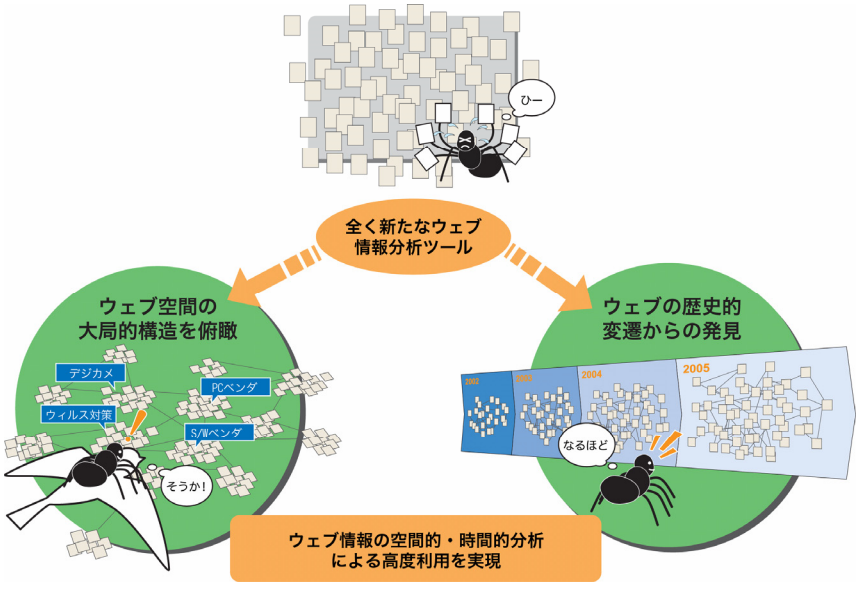
実社会の射影としてのウェブ

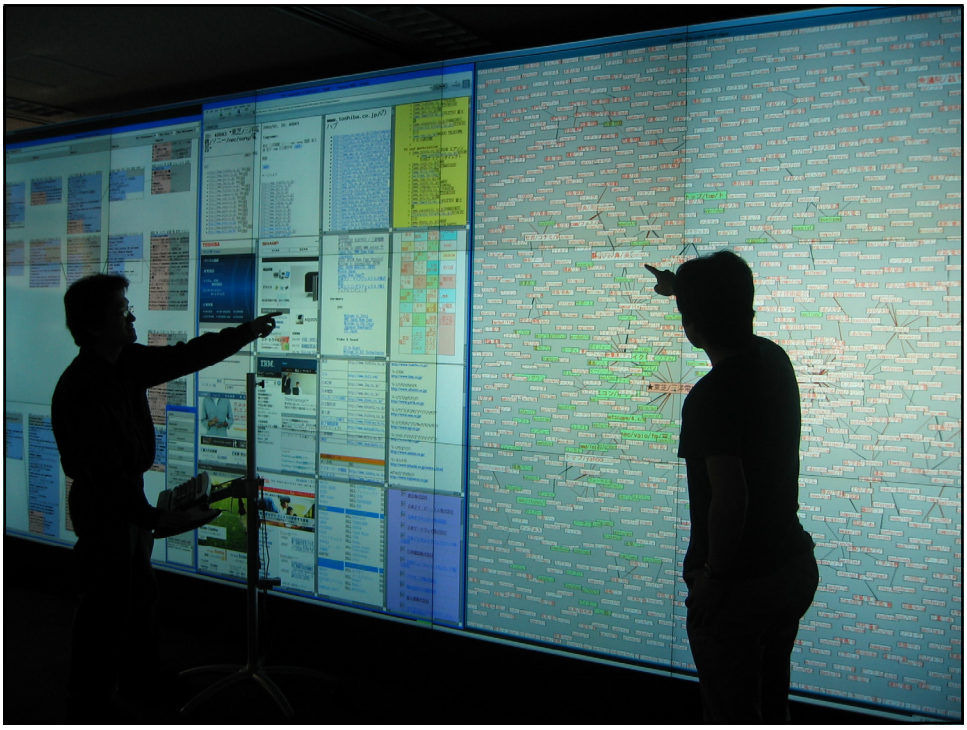
e-Society 文部科学省リーディングプロジェクト



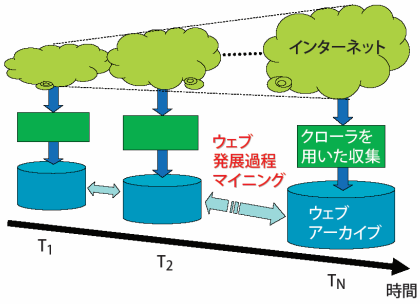
目的：ウェブ情報の高度利用システムの構築（WEBの時空間解析）

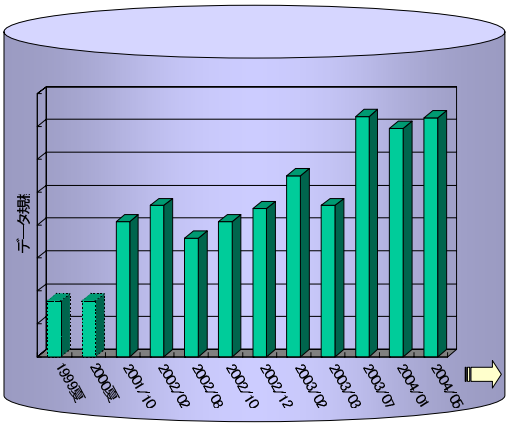
e-Society 文部科学省リーディングプロジェクト





ウェブ時空間解析のための アーカイブ基盤構築

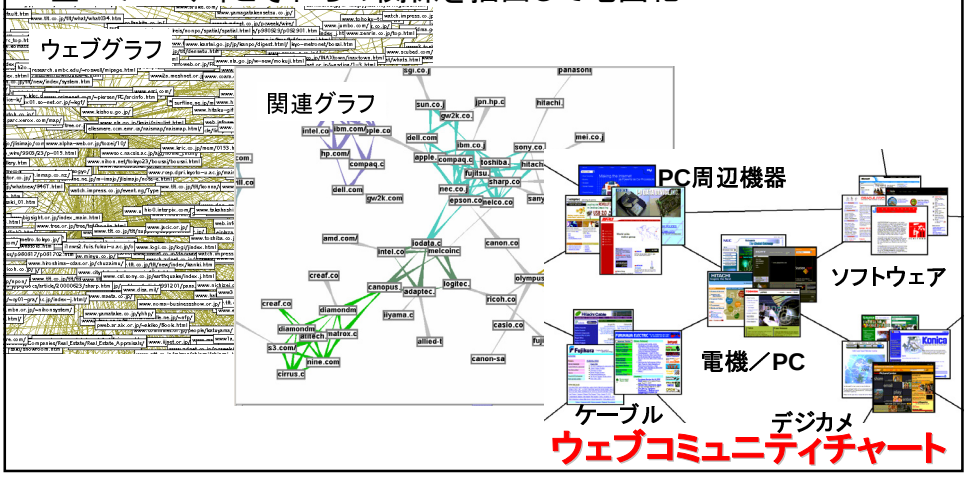




ウェブ空間の構造俯瞰

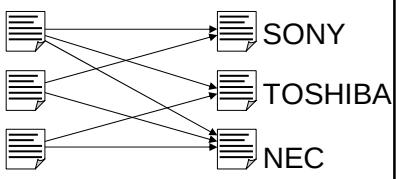
ウェブコミュニティチャート[ACM Hypertext 2001]

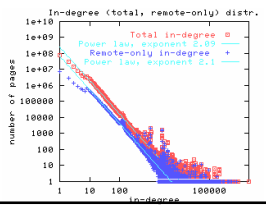
- ある話題に関する有用なページの集合はウェブグラフ上で稠密な構造を持つ(ウェブコミュニティ)
- 全コミュニティとそれらの関係を抽出して地図化

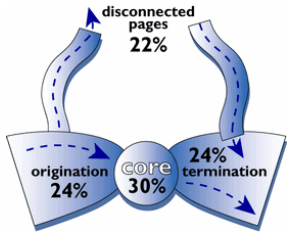


関連研究

- HITS [Kleinberg, 1997]
  - ハブとオーソリティから構成されるウェブコミュニティを抽出
- Trawling [Kumar et al, 1999]
  - 全ウェブコミュニティを列挙、KB化
- Graph Structure in the Web [Broder et al, 2000]
  - ウェブ全体のグラフ構造分析



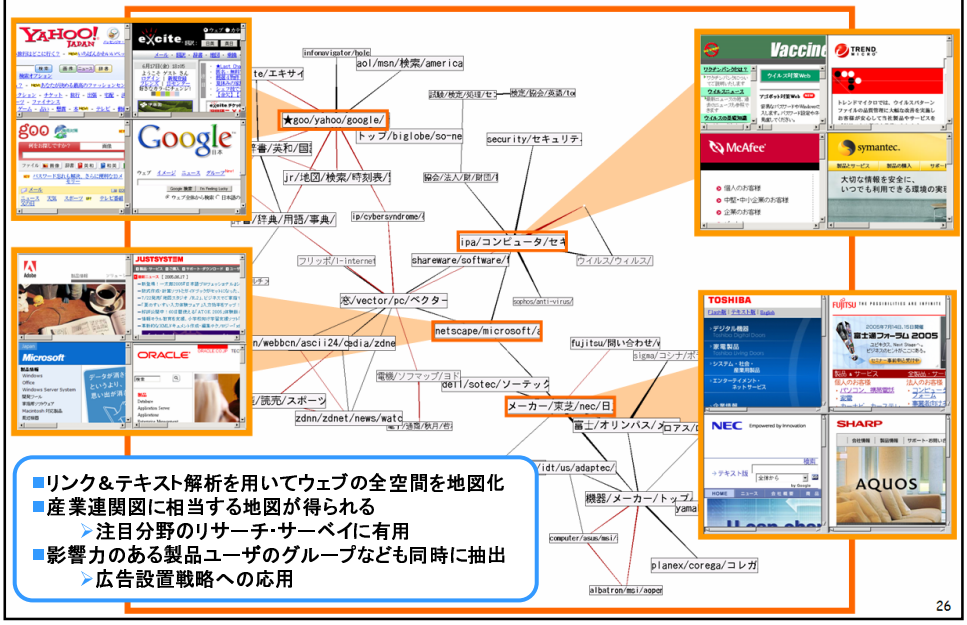




ウェブ空間の構造俯瞰

～コンピューター業界周辺の地図～

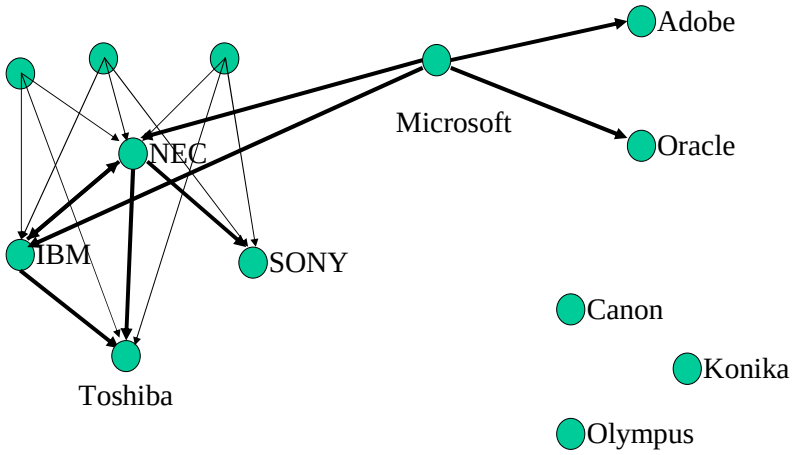
e-Society 文部科学省リーディングプロジェクト



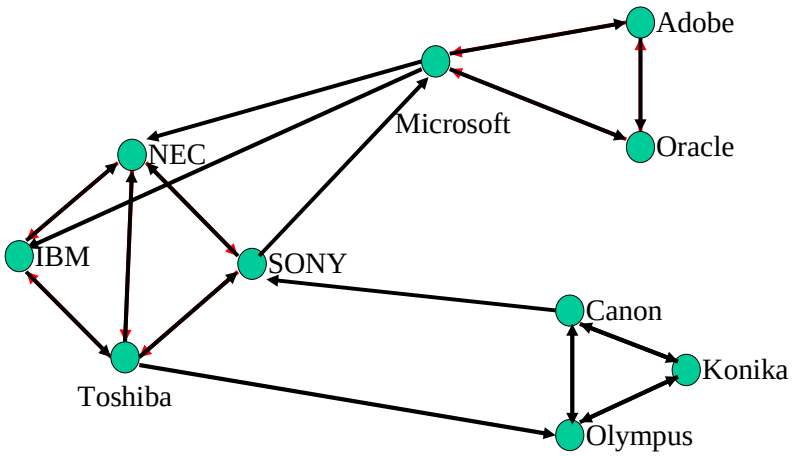
26

Relationship graph

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

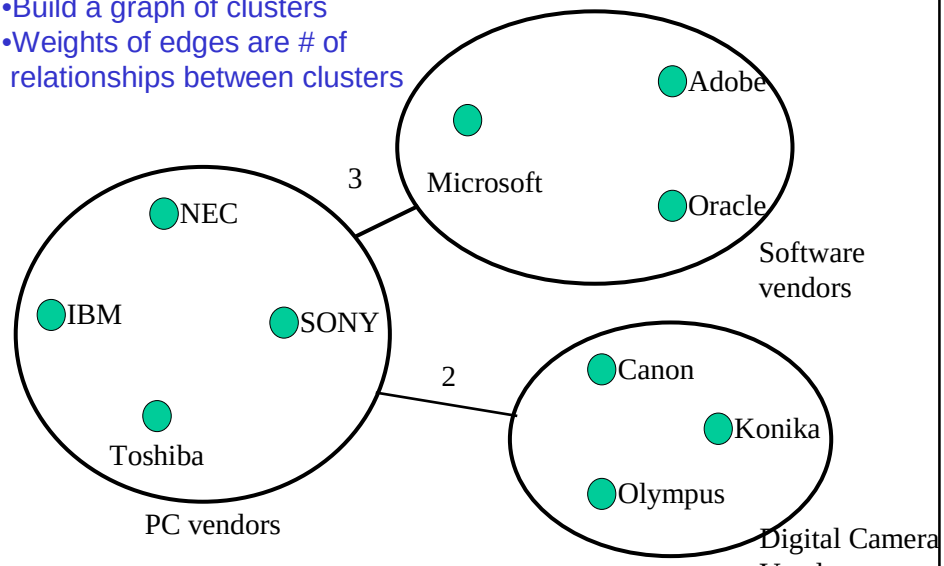


Relationship Graph



Web Community Chart

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

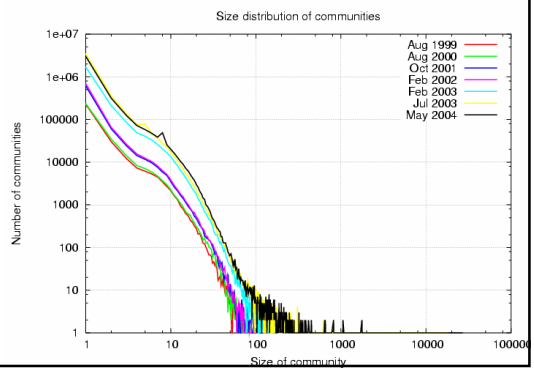


コミュニティチャート生成実験

- 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
- チャート生成
  - 3サイト以上からリンクを受けているページをシードとして使用
  - 上位10位のオーソリティから関連グラフを作成
- サイズの分布
  - べき乗則に従う。SCC,WCCのサイズ分布と同様

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



実験環境

- クローラ： 残念ながら企業秘密
- 解析系： Itanium2サーバ (8 CPU, 128GB Memory)
- 処理時間：
  - スナップショットのスキャンに1日
  - もろもろのデータベース作成に2日～3日
    - リンクDB (URL⇔ID, ID⇒<OutLinks><InLinks>)
      - 2004/05時点では、ID⇒<OutLinks><InLinks>のDBは 2.6Gノード(1999からの延べURL数に相当)、 4.5Gリンクを、 30GB程度に圧縮してオンコアDB化
    - タイトル、アンカーDB (URL⇒Title, URL⇒<AnchorTexts>)
  - チャート作成に半日

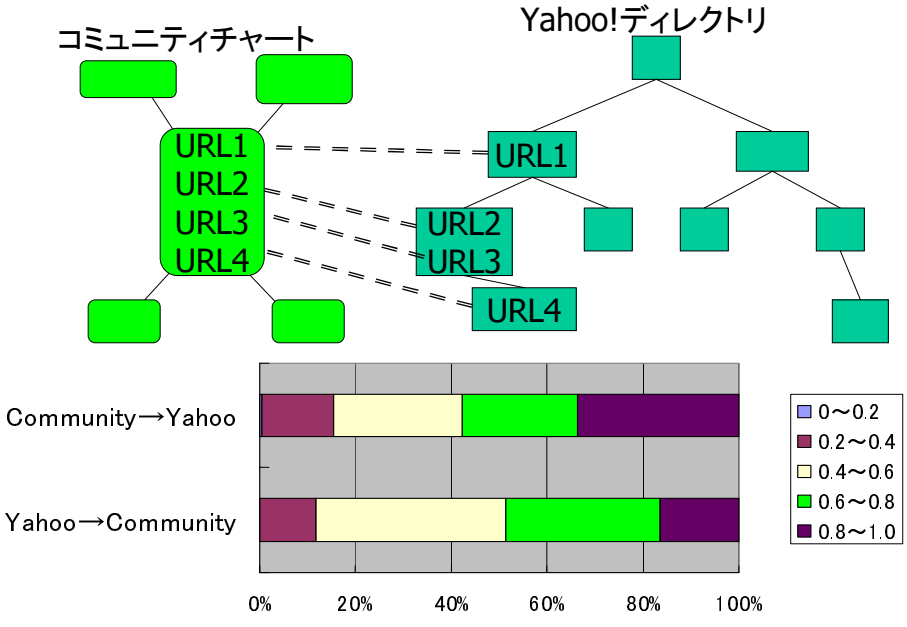
コミュニティチャートとYahoo!の比較 [吉田 2003]

- 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

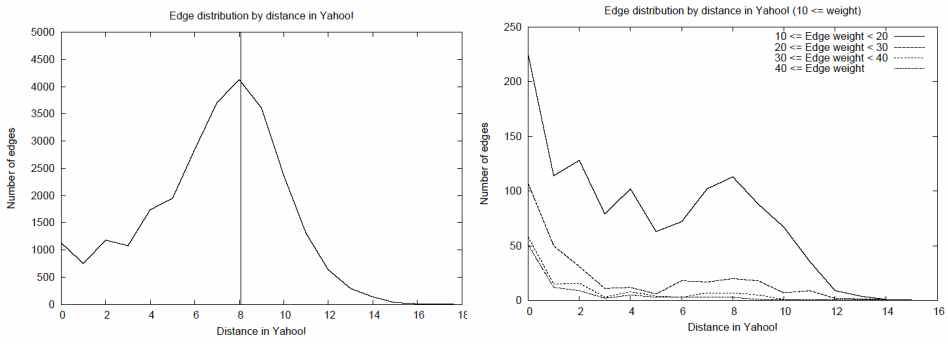
- 比較対象とするコミュニティとディレクトリ
  - 共有部分内においてURL数5以上のもの
    - 4079コミュニティ(33930URL 平均8.13)
    - 4965ディレクトリ(63757URL 平均12.84)

チャートとYahoo!の類似度



コミュニティ間の関連度

- 関連度の高い辺は近いカテゴリを結んでいる



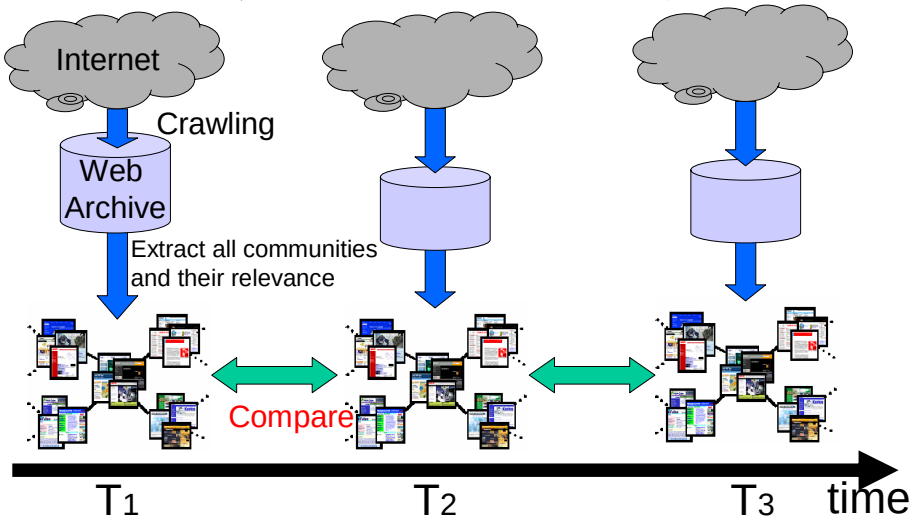
応用研究

- 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

ウェブの時系列分析

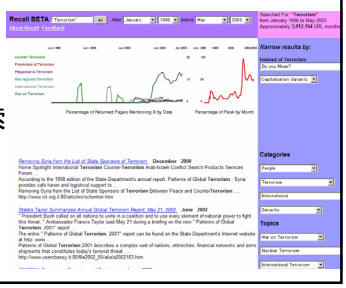
[ACM Hypertext 03]

- 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



関連研究

- A Large-Scale Study of the Evolution of Web Pages [Fetterly et al 2003]
  - 1.5億ページを毎週1回11週間収集
  - 各ページの変化度合いを測定
- What's New on the Web? [Ntoulas et al 2004]
  - 著名な150サイトを取り尽す収集を週1回、1年間実施
  - 新規出現ページ・コンテンツの測定
- Recall (Internet Archive) [Patterson
  - RecallはInternet Archiveの全文検索エンジン（現在停止中）
  - '96～'03までに収集した110億ページを索
  - 関連キーワードの出現頻度グラフを表示

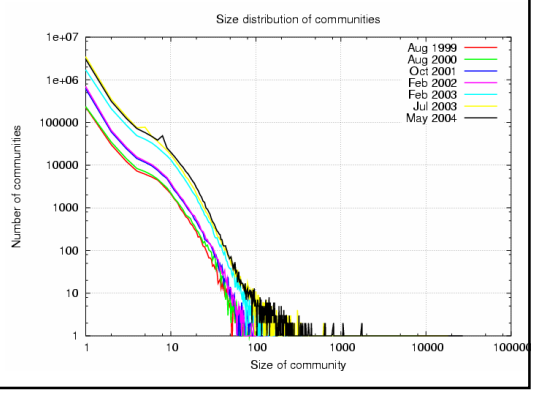


データセット

- 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

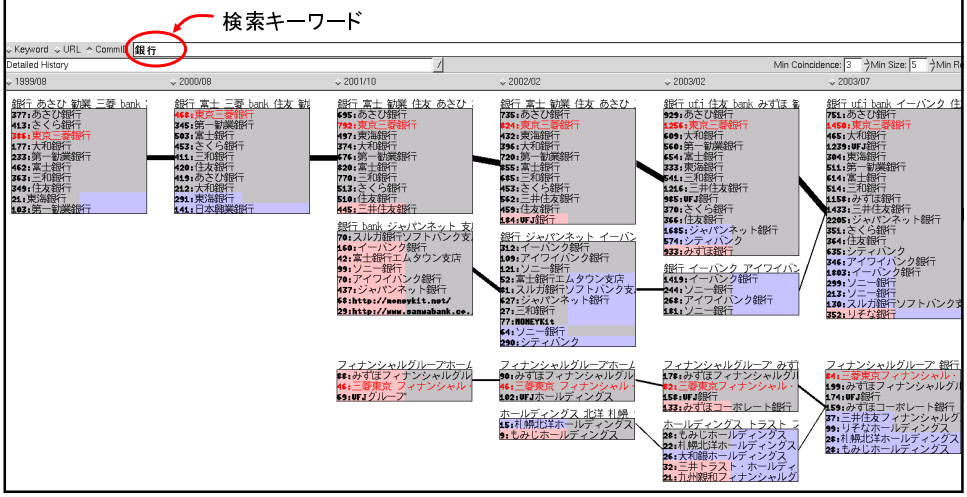
データセット詳細

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



銀行業界の変遷

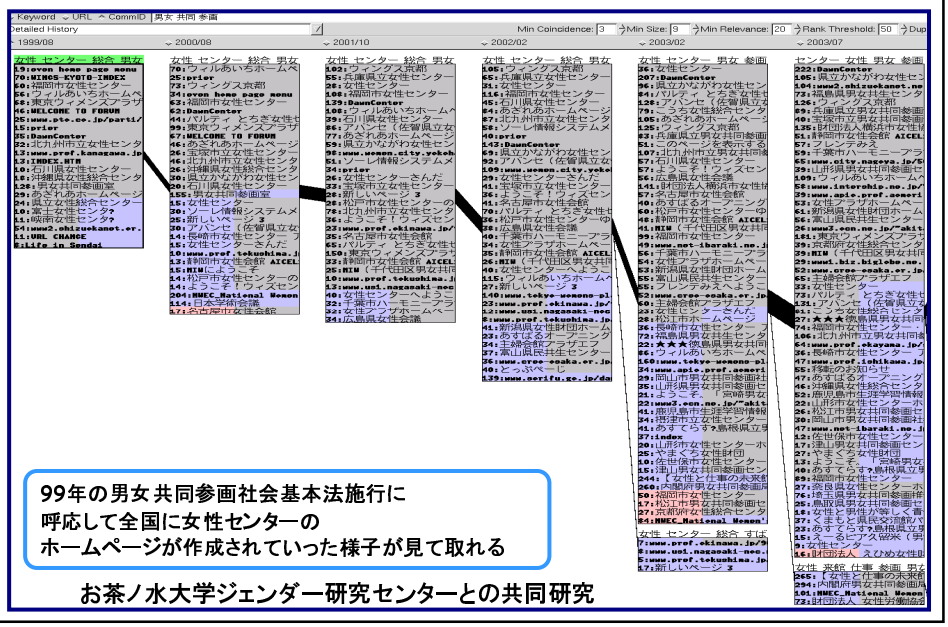
- インターネット銀行の出現と世間への浸透
- 合併した銀行の出現：三井住友、UFJ、みずほ、りそな



社会現象による話題の爆発的発生

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

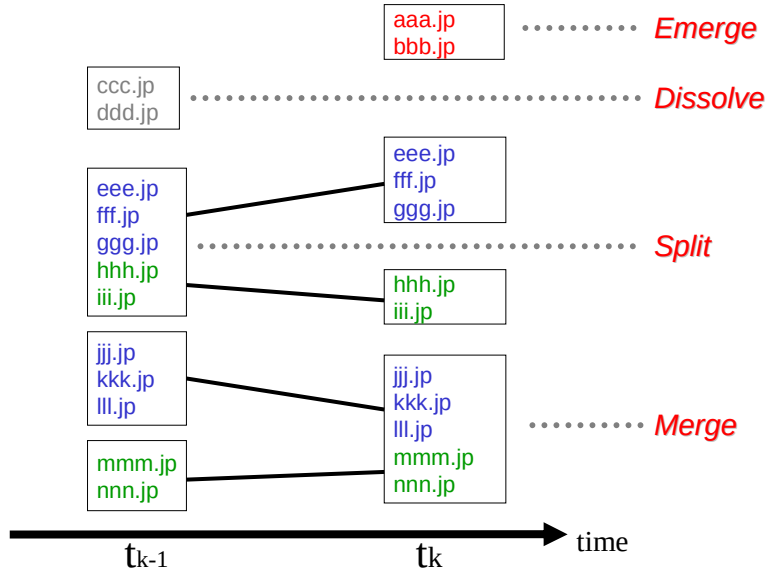


99年の男女共同参画社会基本法施行に
呼応して全国に女性センターの
ホームページが作成されていった様子が見て取れる

お茶ノ水大学ジェンダー研究センターとの共同研究

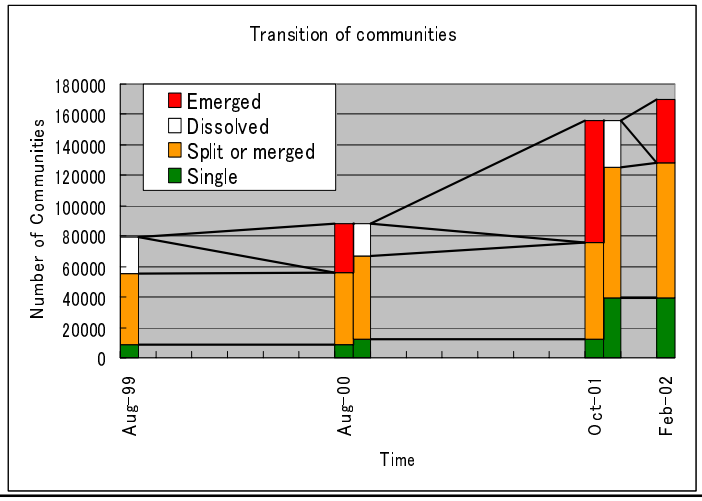
Types of Changes

Changes are detected by comparing neighboring charts



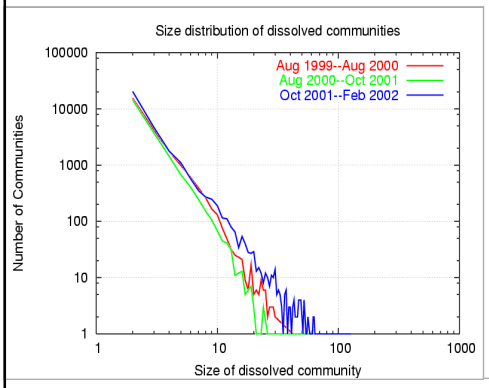
Types of Changes

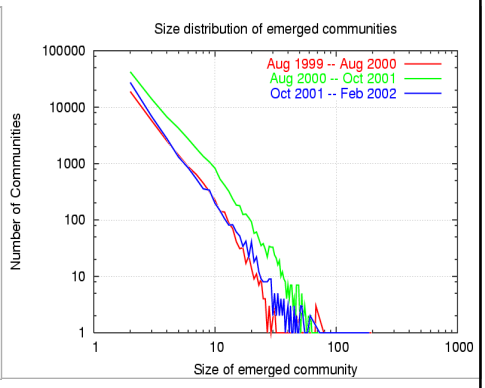
- Structure of communities changes dynamically
- How the size distribution is kept unchanged?



Emerged and Dissolved Communities

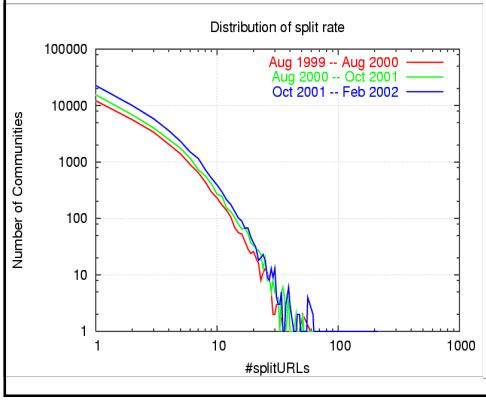
- Both size distributions follow the power-law
- Both exponents are greater than ones in size distribution of all communities

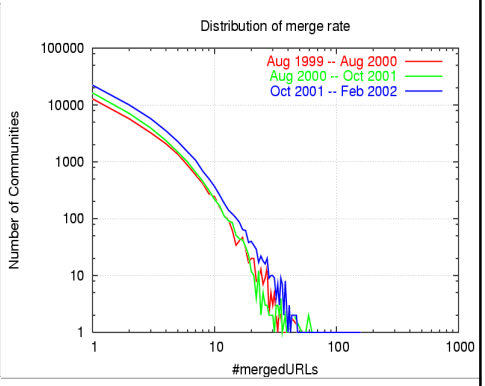




Split and Merged Communities

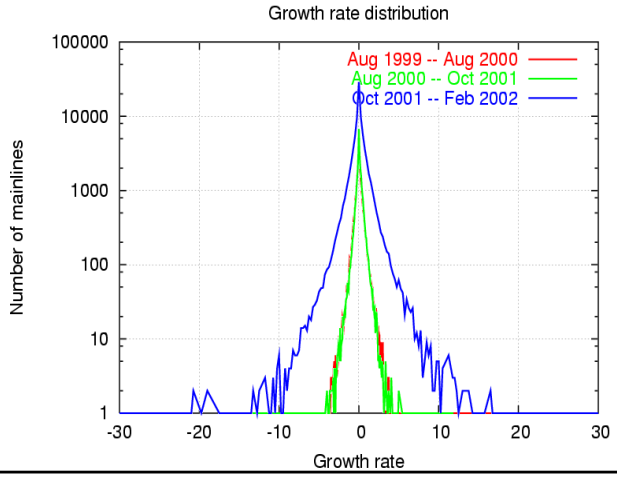
- # of split and merged URLs also follow the power-law, and have clear symmetry





Grown and Shrunken Communities

- Growth rate have clear y-axis symmetry

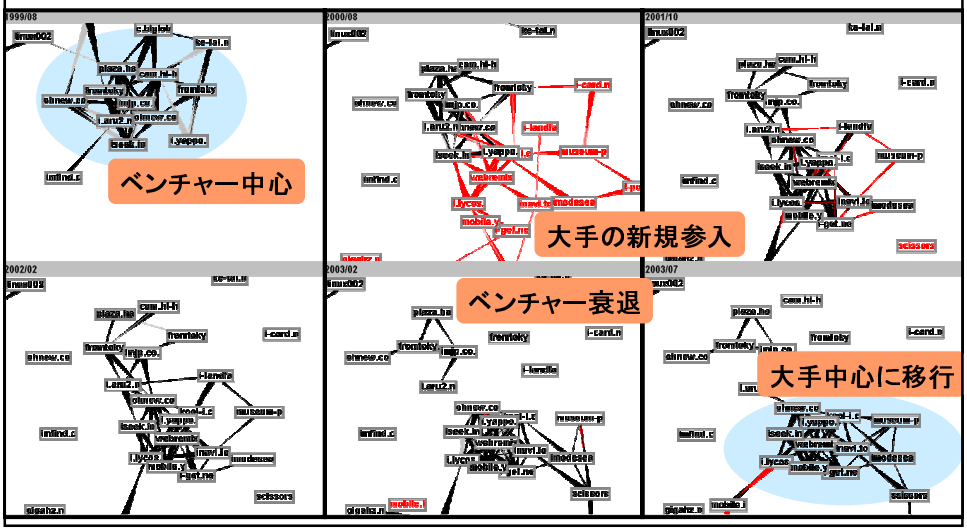


ウェブの時空間分析 [ACM Hypertext 05]

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）

時間



最近の成果と今後の展開

- 検索エンジンスパムの分析
  - 人工的な稠密構造による検索エンジン騙しの傾向分析[小野. DEWS2006]
- 不安定なアーカイブからの新規ページ抽出法
  - [Toyoda. WWW2006]

- 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
- 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
  - 電通・専修大経営学部との協力開始