2009年度総合科目
「情報エレクトロニクスの最先端と夢」

大規模Webアーカイブからの社会分析

2009年4月8日

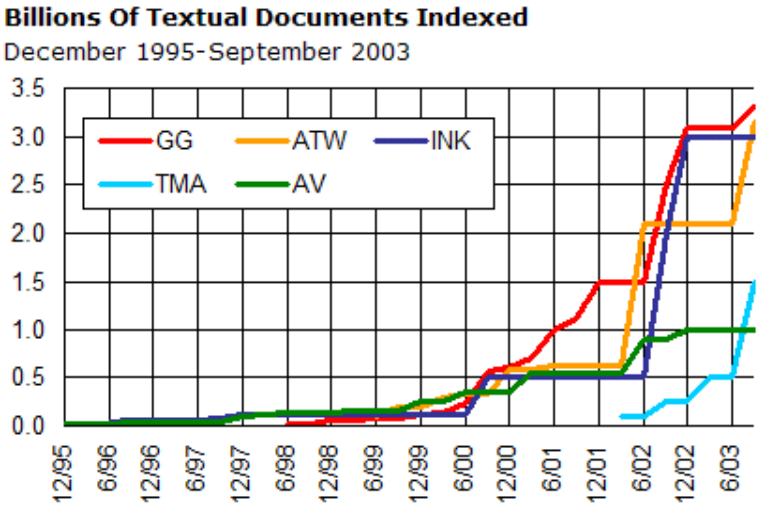
生産技術研究所
豊田正史

研究対象としてのWeb

- 膨大な文書集合
  - 1兆を超えるウェブページ（URL)(Google Official Blog 2008/7)
  - 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

検索エンジンのサイズ

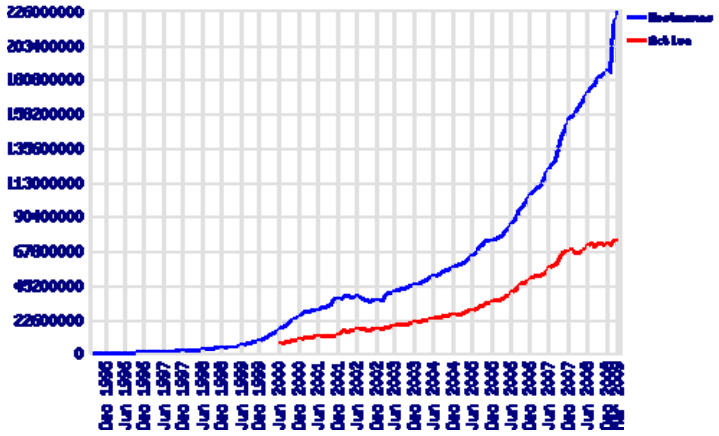
- 2004 GG: 8.1B, 2005 Y!: 20B, 2008 GG: 1T



http://searchenginewatch.com

ウェブの継続的な成長傾向

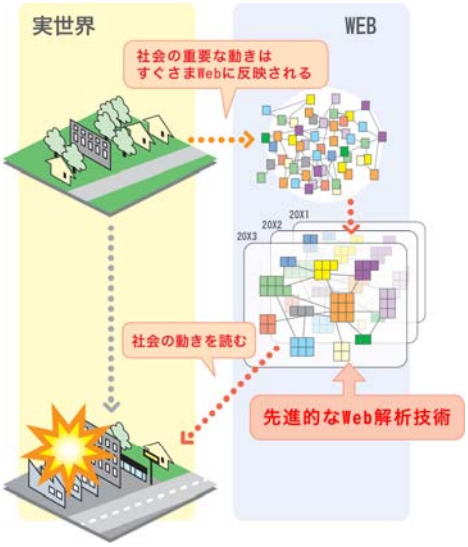
- 米Netcraft社によるウェブサイト数の推移 (news.netcraft.com)



実社会の射影としてのウェブ

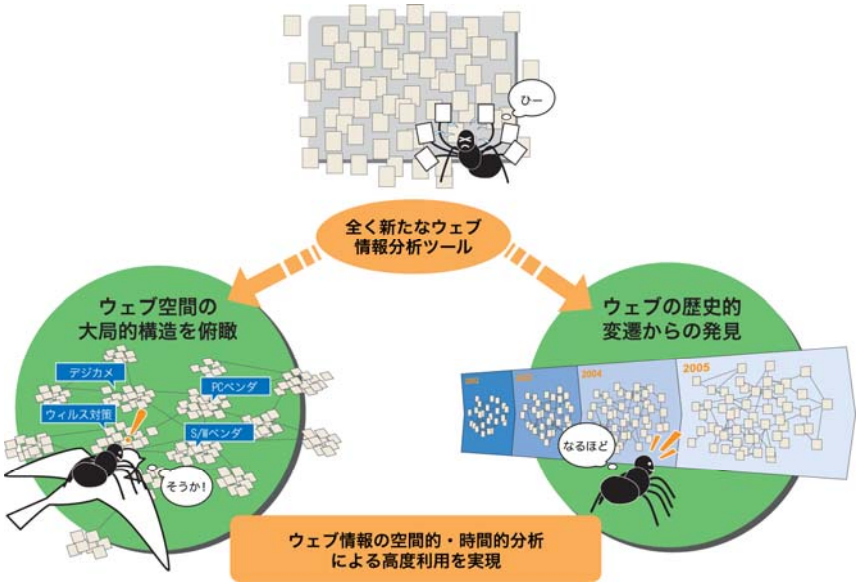
e-Society 文部科学省リーディングプロジェクト

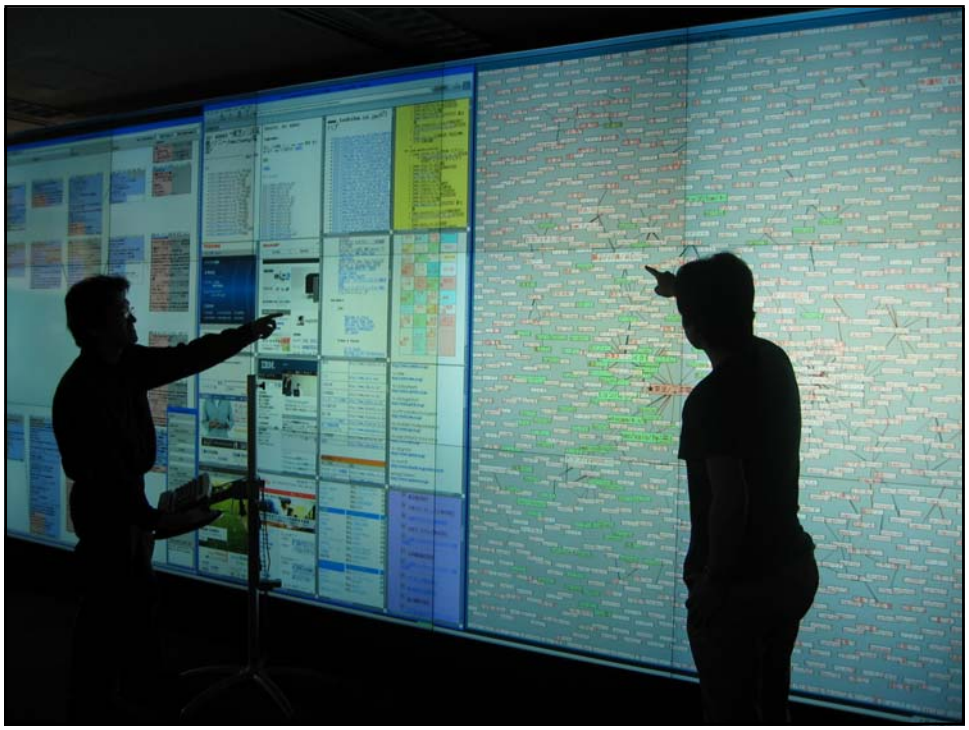
Ｗｅｂは実世界の転写構造を形成



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





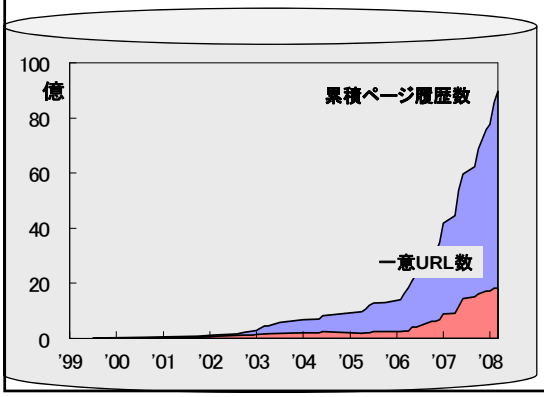
あらまし

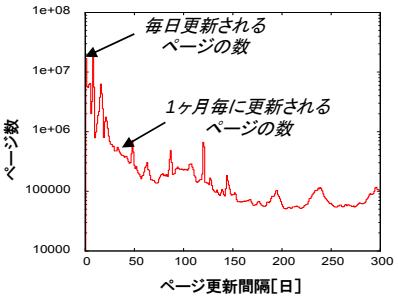
- ウェブアーカイブ基盤
  - アーカイブの概要と簡単なアプリケーション
- ウェブ空間の構造俯瞰
  - リンク解析によるウェブの地図
  - Yahoo!との比較
- ウェブの時系列分析
  - ウェブ空間構造の時系列変化を可視化

日本語ウェブアーカイブの構築

e-Society 文部科学省リーディングプロジェクト

- 9年間にわたり100億ページ規模の日本語ウェブページを集積し、継続期間および規模において**アジア圏最大級**のウェブアーカイブを構築
- 各URLの更新頻度に応じた収集技術を開発し、1日～1年の**可変周期収集**を実現

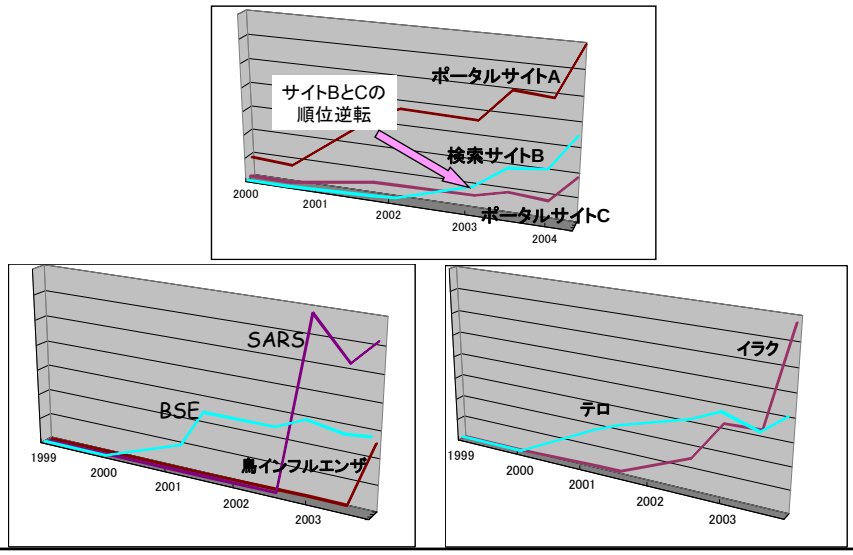




アーカイブの簡単なアプリケーション

- 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

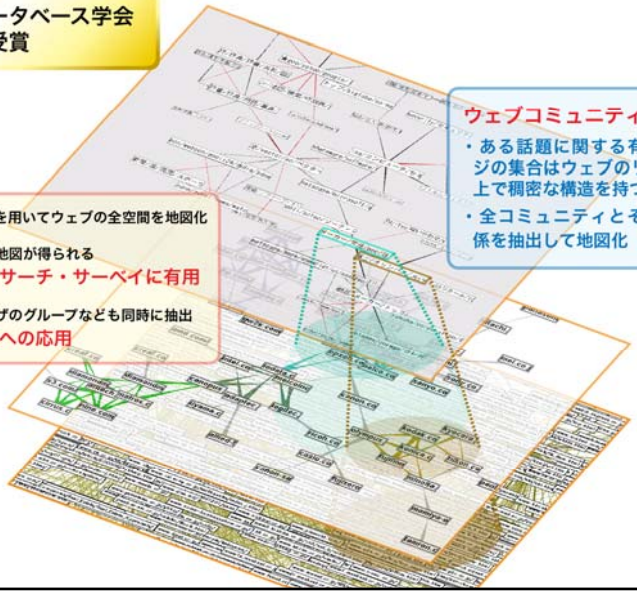
- ウェブアーカイブ基盤
  - アーカイブの概要と簡単なアプリケーション
- ウェブ空間の構造俯瞰
  - リンク解析によるウェブの地図
  - Yahoo!との比較
- ウェブの時系列分析
  - ウェブ空間構造の時系列変化を可視化

ウェブ空間の構造俯瞰

～ウェブ全空間の地図化～

e-Society 文部科学省リーディングプロジェクト

日本データベース学会
論文賞受賞



ウェブコミュニティチャート
- ある話題に関する有用なページの集合はウェブのリンク空間上で稠密な構造を持つ
- 全コミュニティとそれらの関係を抽出して地図化

リンク＆テキスト解析を用いてウェブの全空間を地図化

産業連関図に相当する地図が得られる
➡注目分野のリサーチ・サーベイに有用

影響力のある製品ユーザのグループなども同時に抽出
➡広告設置戦略への応用

13

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

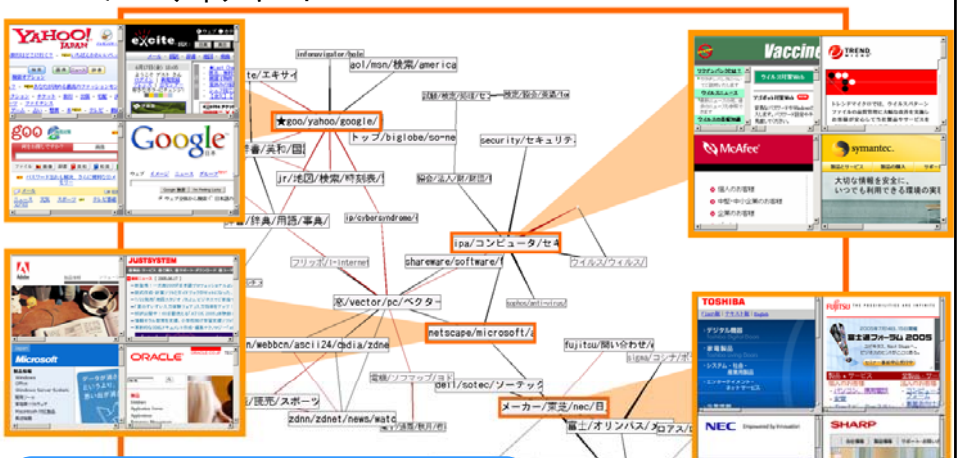
例1　千葉ロッテマリーンズファンのコミュニティ

例2　PCメーカのコミュニティ

ウェブ空間の構造俯瞰

～コミュニティチャート～

e-Society
文部科学省リーディングプロジェクト



- ■リンク＆テキスト解析を用いてウェブの全空間を地図化
- ■産業連関図に相当する地図が得られる
  - ➢注目分野のリサーチ・サーベイに有用
- ■影響力のある製品ユーザのグループなども同時に抽出
  - ➢広告設置戦略への応用

26

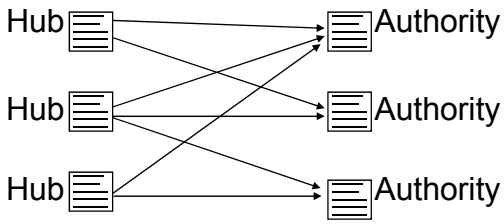
HITS [Kleinberg '97]

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

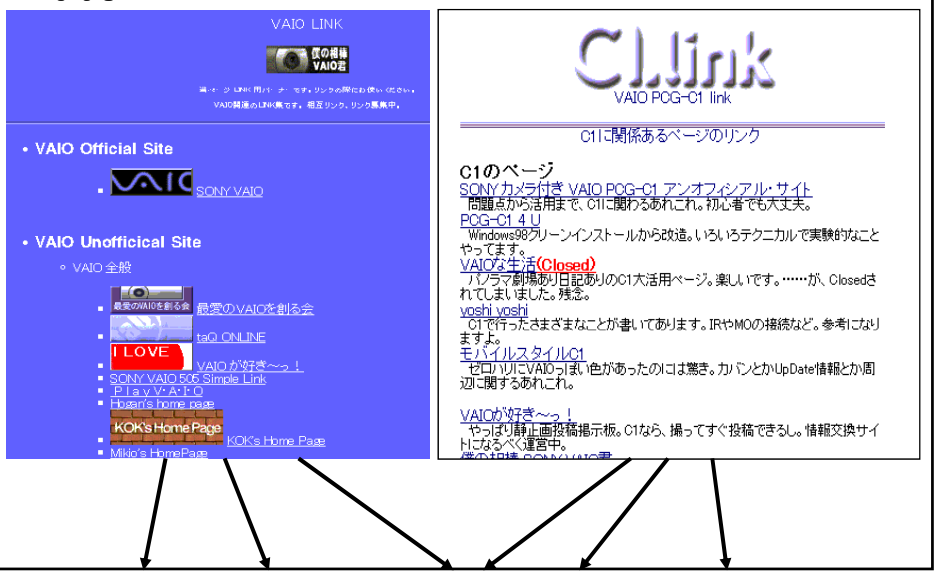
- 適当なウェブの部分グラフから良いhubとauthorityを抽出する
  - Hub: 多くの良いauthorityを指しているリンク集
  - Authority: 多くの良いhubから指されているページ



良いAuthorityとHubの集まりをコミュニティと呼ぶ

HubとAuthorityの例

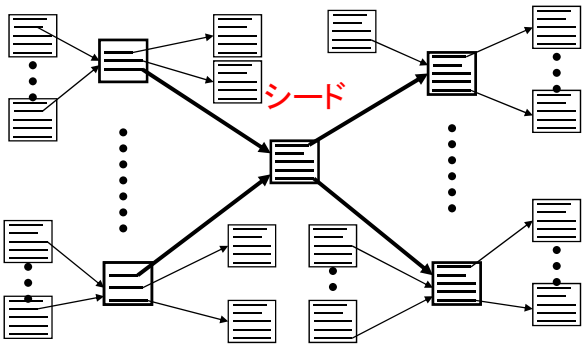
Hubs



Authorities

サブグラフ作成（シードページから）

距離2以内のページを集める

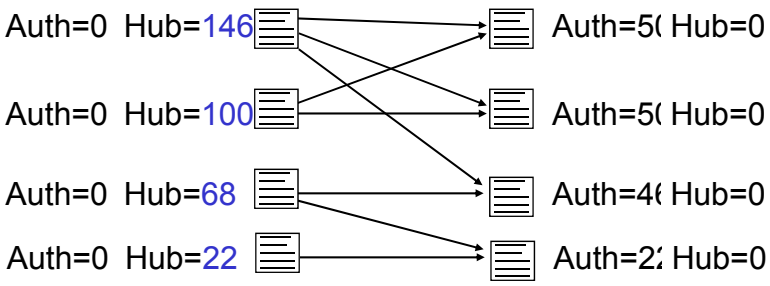


Hub,Authorityスコアの計算

すべてのページのauth(n) = hub(n) = 1
スコアが収束するまで以下を繰り返す

$$auth(n) = \Sigma\ hub(m), \text{ for all } m \rightarrow n$$

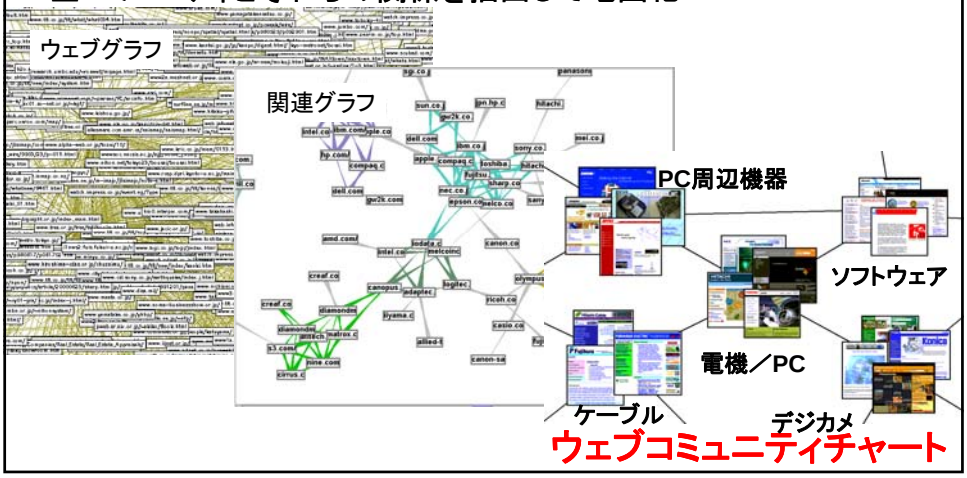
$$hub(n) = \Sigma\ auth(m), \text{ for all } m \leftarrow n$$



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



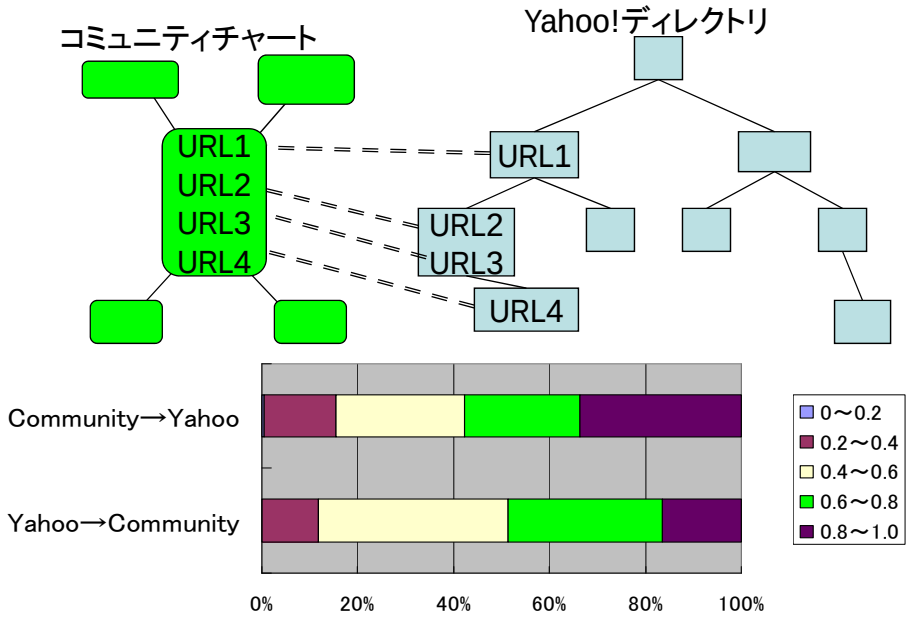
コミュニティチャートとYahoo!の比較 [吉田 2003]

- 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

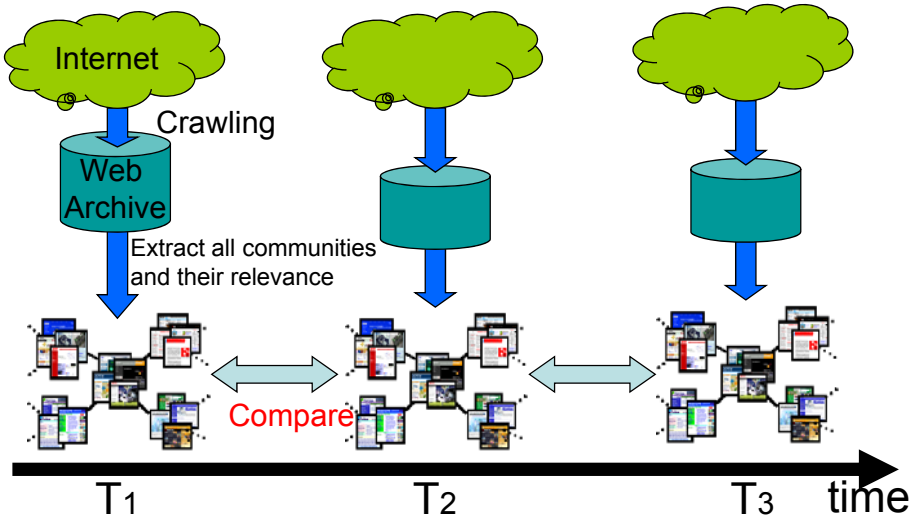
あらまし

- 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

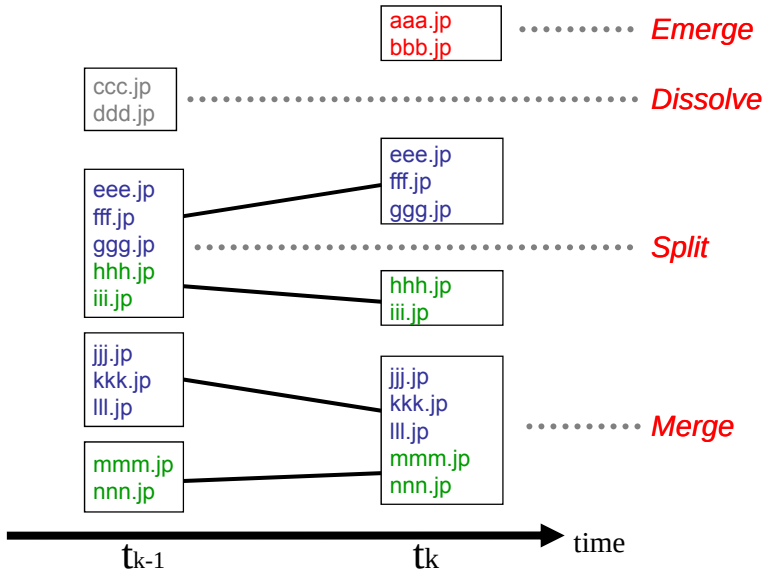
[ACM Hypertext 03]

- 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



Types of Changes

Changes are detected by comparing neighboring charts

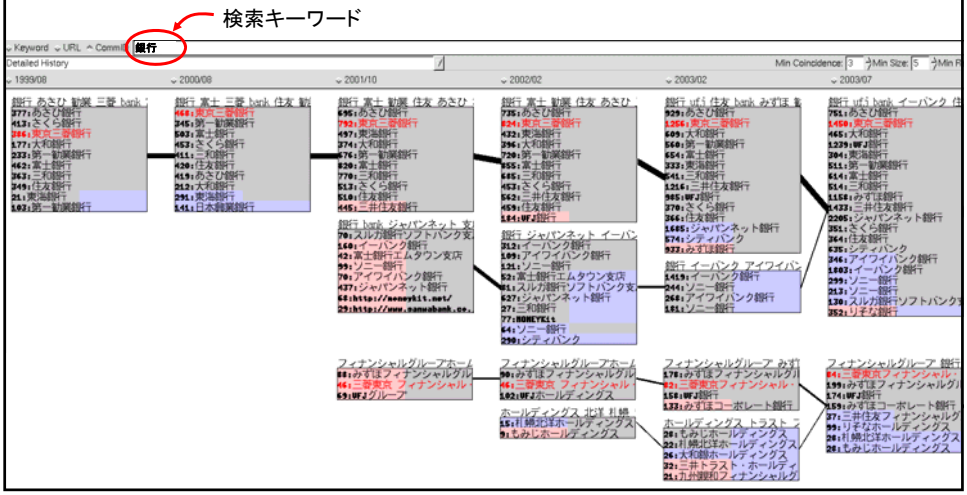


成果②ウェブの時系列分析

～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

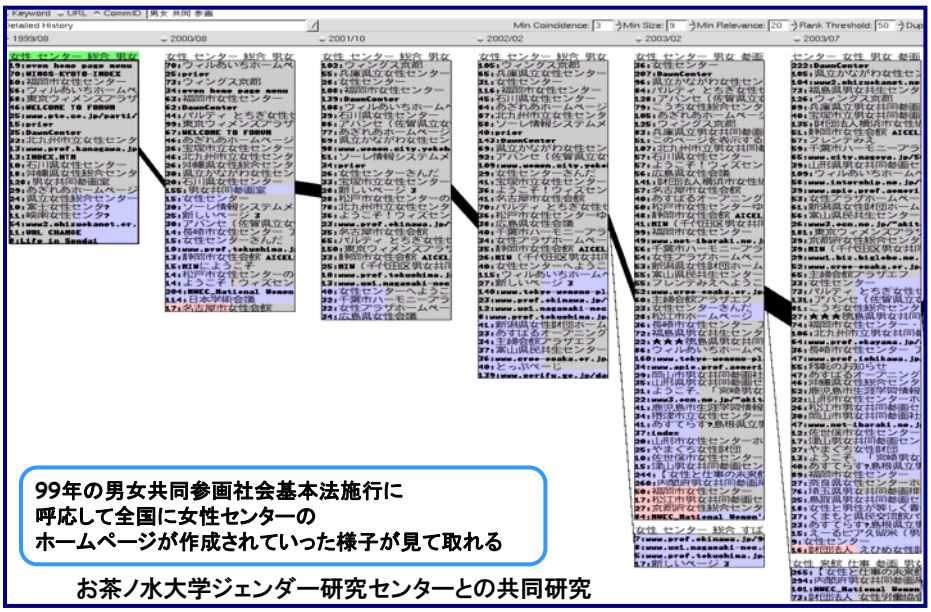


成果②ウェブの時系列分析

～社会現象による話題の爆発的発生～

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～

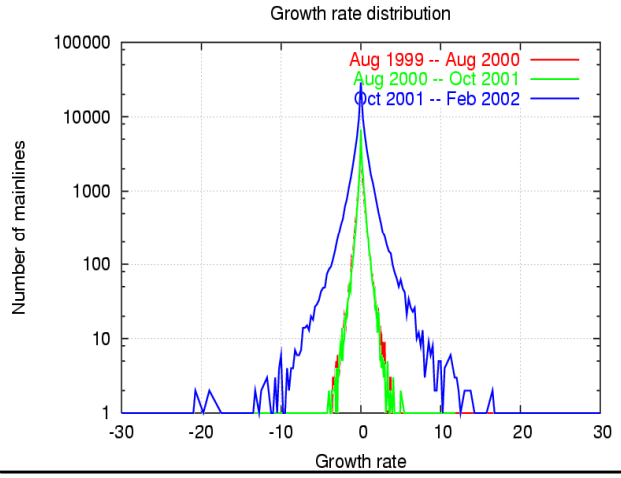


99年の男女共同参画社会基本法施行に
呼応して全国に女性センターの
ホームページが作成されていった様子が見て取れる

お茶ノ水大学ジェンダー研究センターとの共同研究

Grown and Shrunken Communities

- Growth rate have clear y-axis symmetry

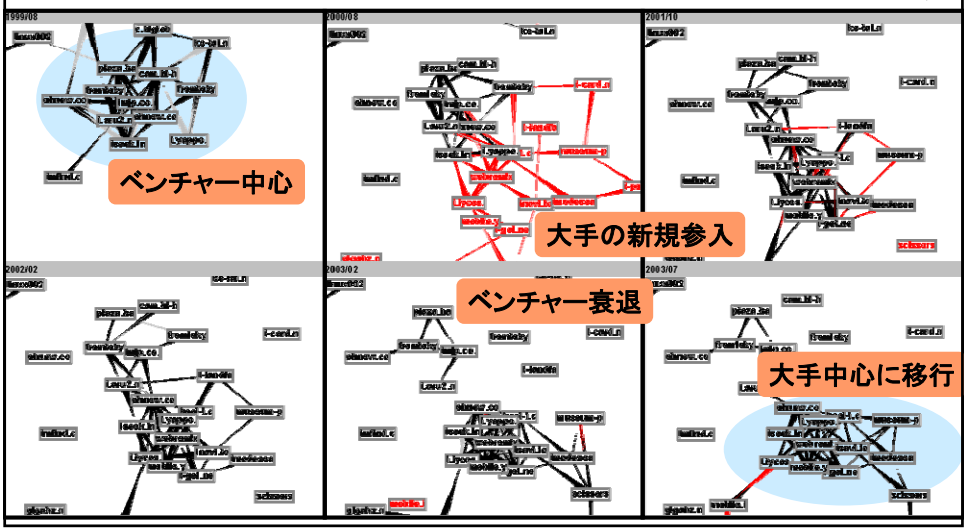


ウェブの時空間分析 [ACM Hypertext 05]

e-Society 文部科学省リーディングプロジェクト

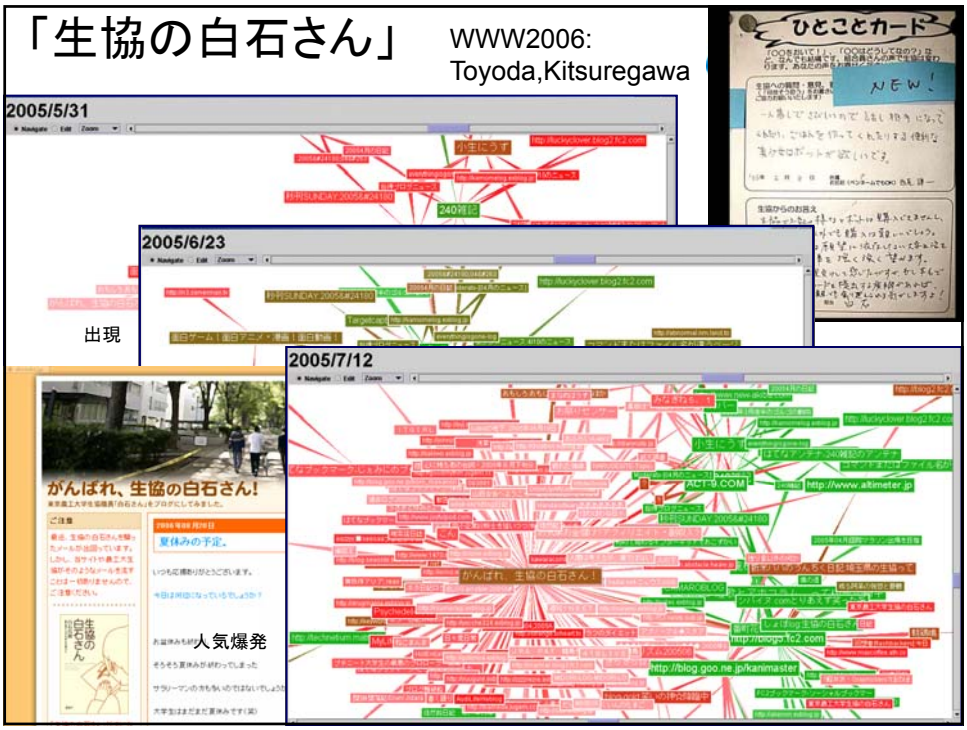
空間＋時間分析：コミュニティの変遷
（例：i-mode検索サイト）

時間



「生協の白石さん」

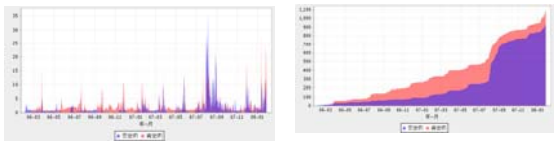
WWW2006:
Toyoda,Kitsuregawa



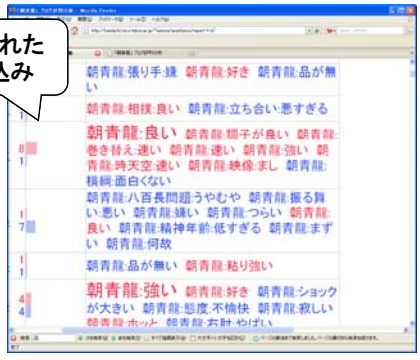
評判情報抽出による世論の分析
～朝青龍の例～

e-Society
文部科学省リーディングプロジェクト

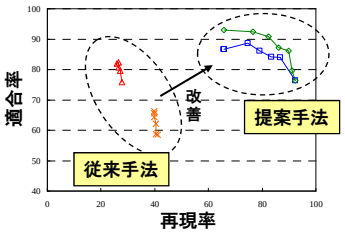
好不評書き込み数の通時的変遷



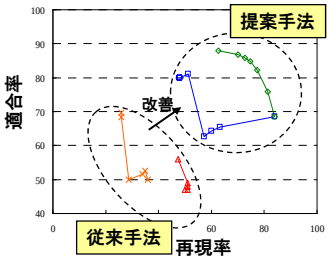
抽出された書き込み



好評表現の抽出精度

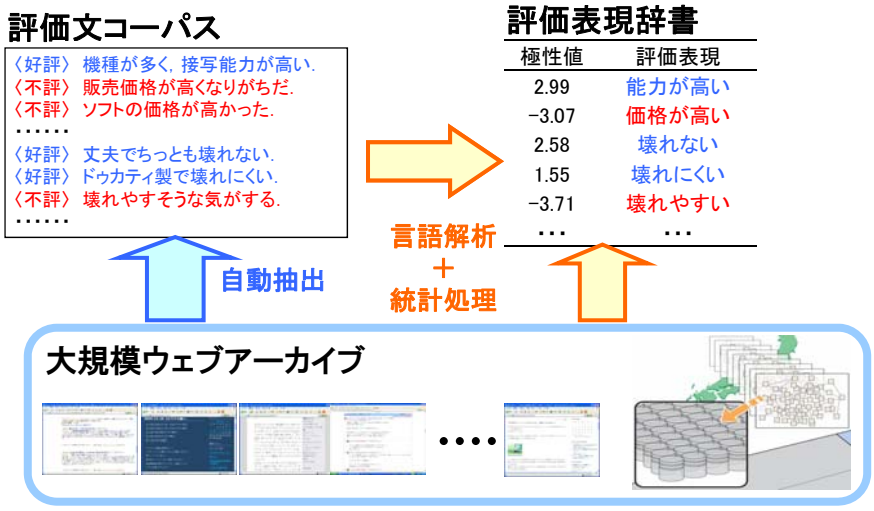


不評表現の抽出精度



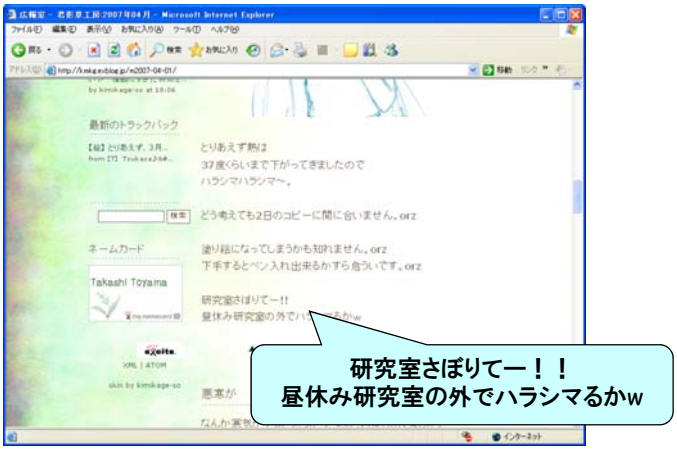
評価表現辞書の自動構築

- 大規模ウェブアーカイブを用いて評価文の自動抽出および評価表現辞書の自動構築を行う



37

新語抽出



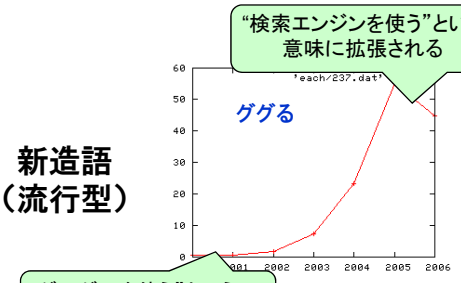
はらしま・る 【ハラシマる】（動詞−五段ラ行）
（1） 原稿を書くこと．同人誌の著者のコミュニティで使われる表現．由来は「原稿→原縞→はらしま」という誤読．
—連休は腰を据えてハラシマるぞ．

38

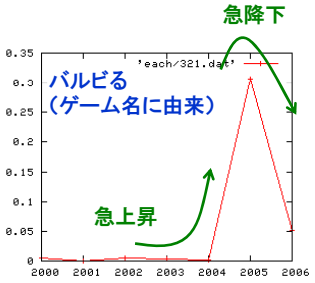
新造語に見られる意味拡張のダイナミズム

- 各種単語の時系列頻度の傾向を分析

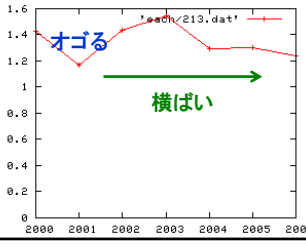
新造語（流行型）



新造語（一過型）



既存語

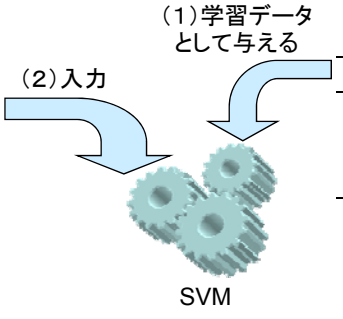


ウェブからの新造語獲得手法

機械学習を用いたカタカナ用言抽出

「テンパ」の後続文字列

る	500
らない	100
ない	3
った	500
た	15
が	1
を	2
……	……



既知の単語の後続文字列

形容詞	動詞	名詞	……
早い	食べる	機械	
美しい	喋る	ご飯	
……	……	……	……

る	300
らない	120
ない	3
った	450
が	0
を	3
……	…

まとめ

- Webアーカイブ
- Web空間の構造俯瞰
- Webの時系列分析
- さまざまな応用
  - 社会学
  - マーケティングへ
  - 言語学