

<http://www.tkl.iis.u-tokyo.ac.jp/~toyoda/lecture/sougou2010.html>

2010年度総合科目
「情報エレクトロニクスの最先端と夢」

10年間にわたるWebアーカイブを 用いた社会分析

2010年4月7日

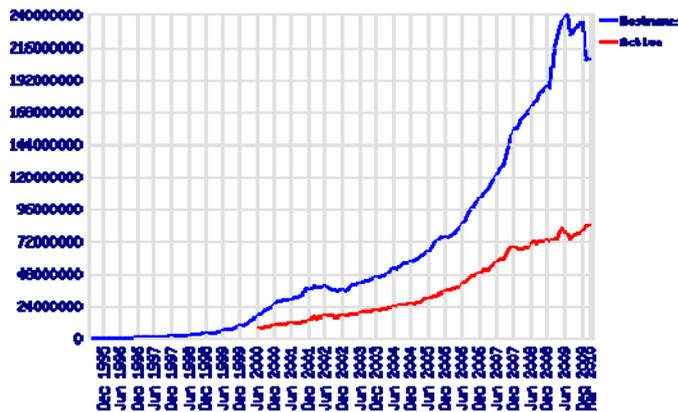
生産技術研究所
豊田正史

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)
- **膨大なネットワーク構造**
 - 文書とハイパーリンクからなる膨大な文書のネットワーク
- **動的な変化**
 - 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。
- **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

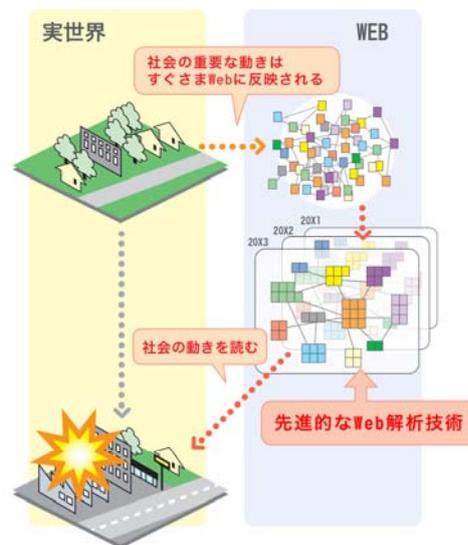
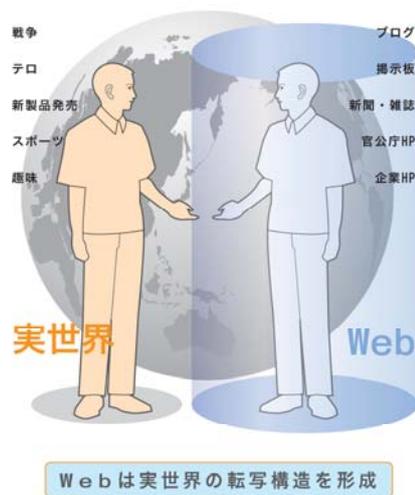
Webの継続的な成長傾向

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



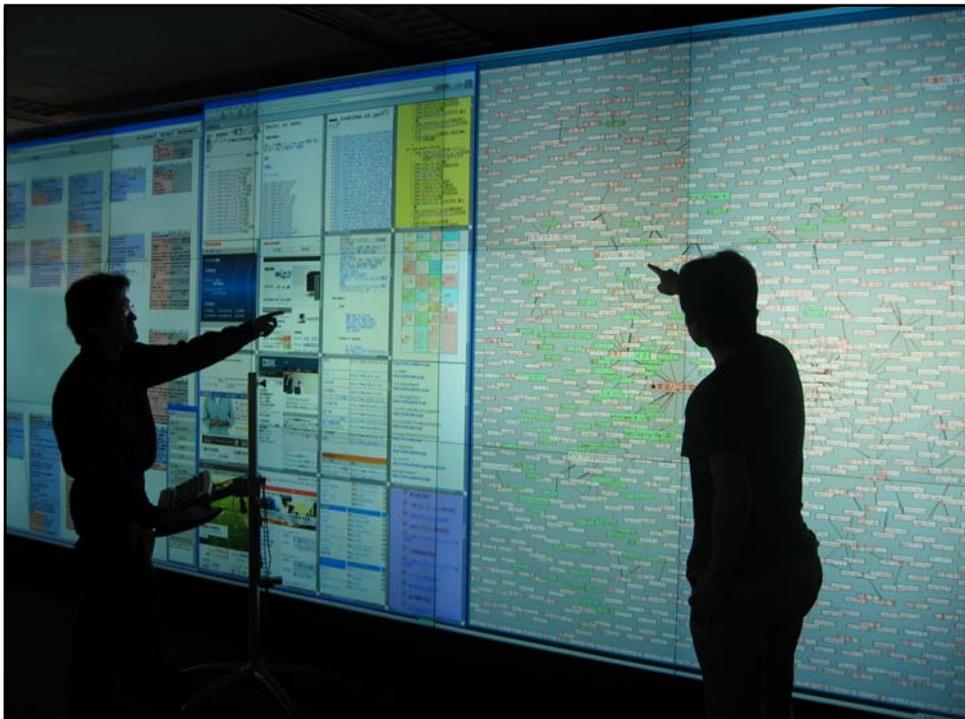
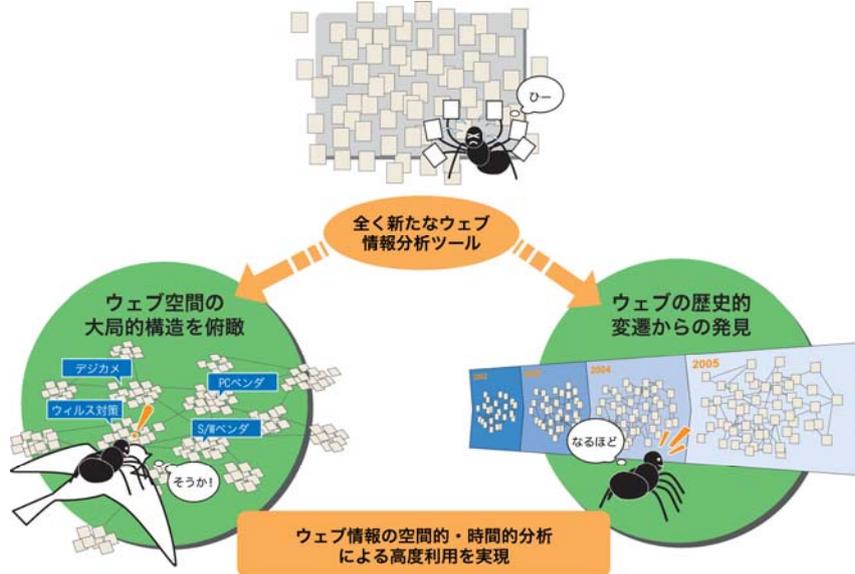
実社会の射影としてのウェブ

e-Society
文部科学省リーディングプロジェクト



目的: ウェブ情報の高度利用システムの構築 (Socio-Sense)

e-Society
文部科学省リーディングプロジェクト



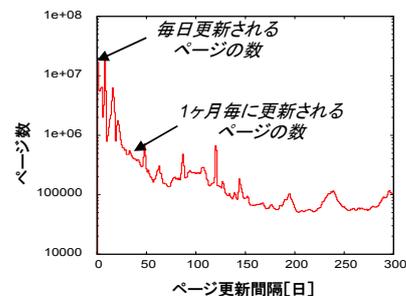
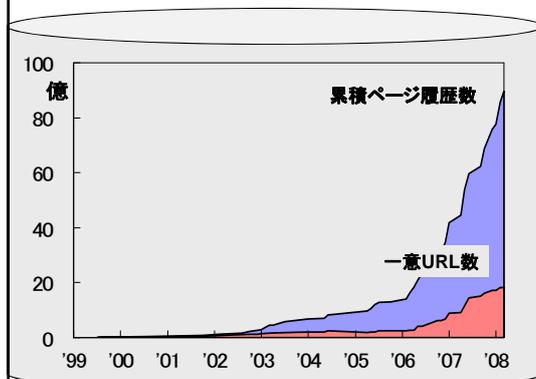
あらまし

- Webアーカイブ基盤
- Web空間の構造俯瞰
- Webの時系列分析
- Webを用いた自然言語処理
- Webスパムの分析

日本語ウェブアーカイブの構築

e-Society
文部科学省リサーチングプロジェクト

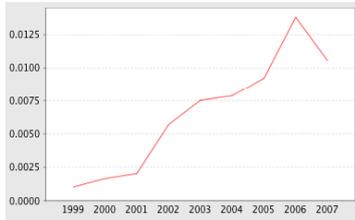
- 10年間にわたり150億ページ規模の日本語ウェブページを集積し、継続期間および規模において**アジア圏最大級**のウェブアーカイブを構築
- 各URLの更新頻度に応じた収集技術を開発し、1日～1年の**可変周期収集**を実現



8

Webアーカイブ検索エンジン

ヤフーの検索結果 (187秒)



1999年の検索結果 (156件)

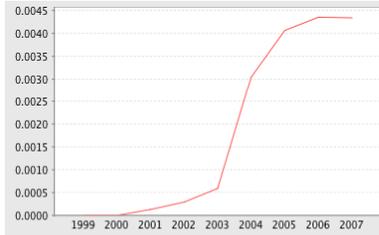
ヤフーは墓穴を掘って

ヤフーは墓穴を掘っていませんか？ 先ずお断りしておきます。このページが書
きで、ヤフー(株)に対する批判や内政干渉を意図したものではありません。セ
フトに登録申請しても、ヤフーの審査が極めて厳しく、登録し直せる可能性
どう評価すべきか、皆さんと一緒に考えたいと思います。タイトルの「ヤフーは墓
ははや...」他意はありません。ヤフーの強みは、何と言っても、カテゴリが
コメントも簡潔で分かり易く使い勝手が良いことです。ヤフーでは担当サーファ
のサイトは掲載する価値があるかを判断し、登録するカテゴリとコメントを決
手の良い検索エンジンになっています。ヤフーの危うさは、審査が厳しすぎて、
登録申請があっても、最近ではほとんど登録していないという問題です。ヤフ
簡単に理解して貰えると思います。ヤフーは店舗は立派でも、新刊誌ほど
であり、新刊誌ほど面白い
<http://www2.biglobe.ne.jp/~hakuzou/Link-X5.htm> - キヤン

ヤフーに登録できず困

ヤフーに登録できず困っている方へ朗報あなたも、もしかして、Yahoo! JAF
困惑していませんか？ 私はホームページを初めて作成し、直ちにヤフーに登録
審査、検索エンジンベスト10を獲ると、ヤフーは墓穴を掘っていませんか？
ました。更に平成11年4月には3つ目として、「トヨタの新車をWWWで買う功

グーグルの検索結果 (186秒)



1999年の検索結果 (0件)

2000年の検索結果 (0件)

2001年の検索結果 (46件)

ITトレンド

次世代ネット技術の米グーグル、見たいサイト一発で検索グーグル 共同創業者の
とエキサイトを生んだスタンフォード大学から昨秋、インターネット検索サービ
企業が誕生した。次世代ネット検索技術の事業化を目指すグーグルは、大学院生
(26)とサーゲイ・ブリン氏(26)が休学して起業した。全世界で5億に上るヤフーな
のグーグル(「ホワイハウス」)を入力すると、ポルノサイトが出てくる話も有名。今や
5億に上ると見られており、この中から必要な情報を見付けるのは至難の業だ。論文が
グーグルの技術者達は、同じような目的の検索の要、同一検索結果を返す

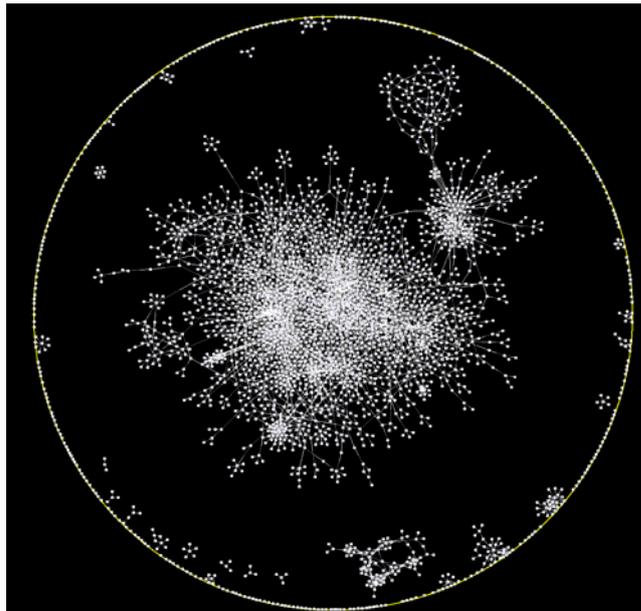
Webページの履歴閲覧

あらまし

- Webアーカイブ基盤
- Web空間の構造俯瞰
- Webの時系列分析
- Webを用いた自然言語処理
- Webスパムの分析

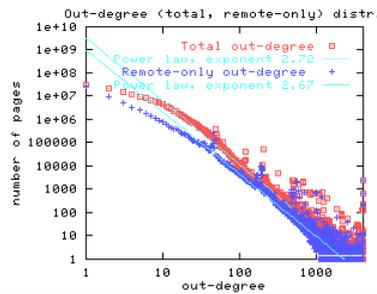
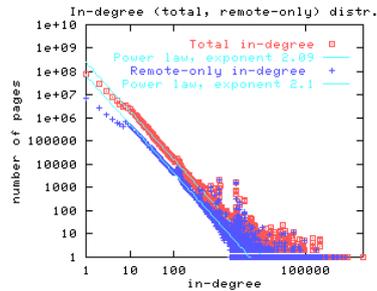
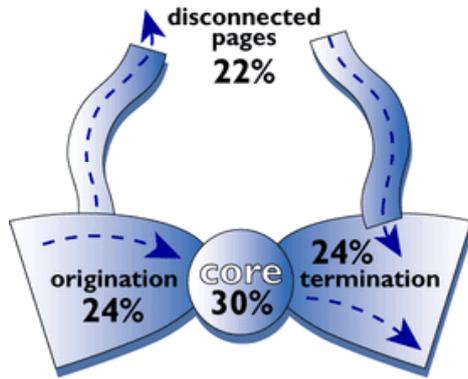
Web Graph

- Graph
 - 複数の点と、それらを結ぶ辺がなす構造
- Web Graph
 - 点: ページ
 - 辺: リンク



Graph Structure in the Web [Broder et al. 2000]

- ウェブ全体のグラフ構造分析
 - 次数分布 / SCCを中心とした蝶ネクタイ構造 / ウェブの直径



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

The network graph shows connections between various web pages and services, including Yahoo!, Google, Microsoft, Oracle, and others. The text box explains the methodology and applications:

- 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

ウェブ空間の構造俯瞰

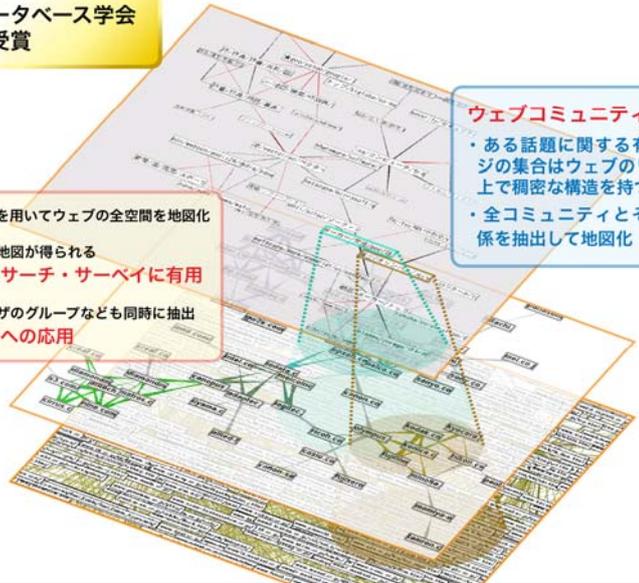
～ウェブ全空間の地図化～



日本データベース学会
論文賞受賞

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

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



15

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メーカーのコミュニティ



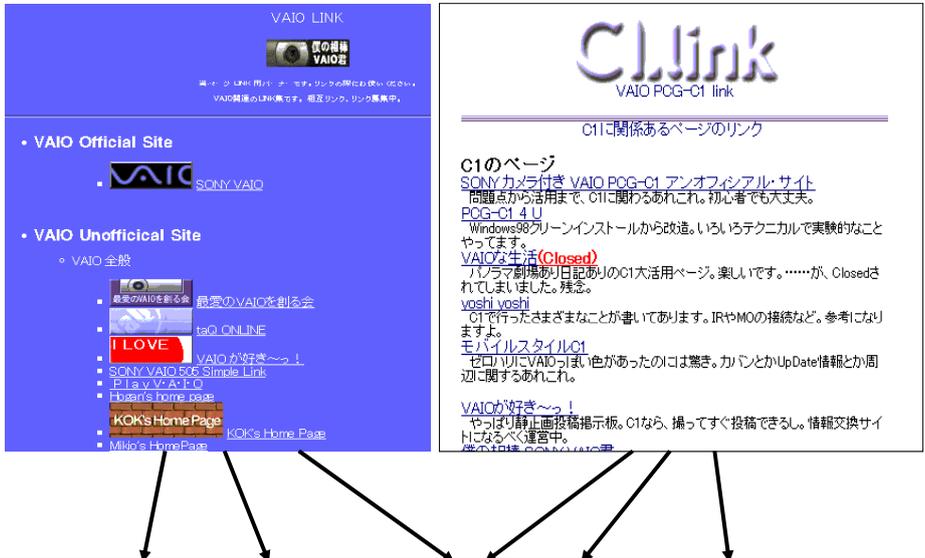
HITS [Kleinberg '97]

以下の観測を基にコミュニティを発見する

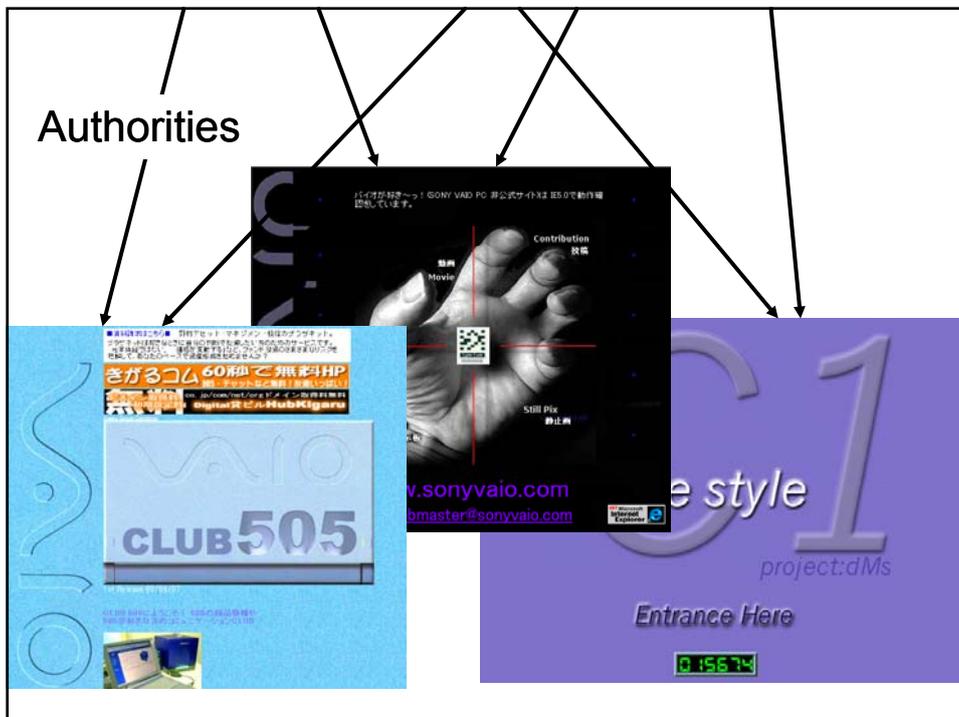
- ハイパーリンクはリンク先のページを推薦する
 - お勧めしないページはわざわざリンクしない
- 色んな人が同種類のハイパーリンクを分類してまとめている
 - ブックマーク、お気に入り、リンク集、ポータルサイト、相互リンク、お友達リンク、etc.

HubとAuthorityの例

Hubs

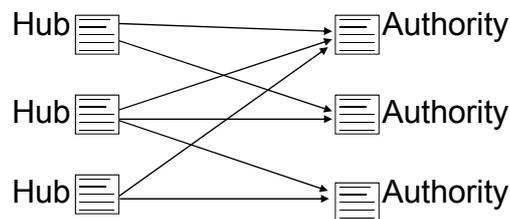


Authorities



HubとAuthority

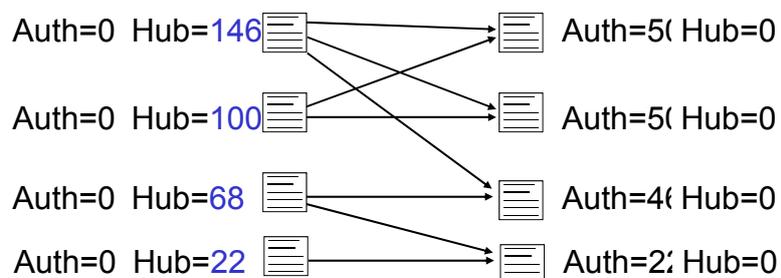
- 適当なウェブの部分グラフから良いhubとauthorityを抽出する
 - Hub: 多くの良いauthorityを指しているリンク集
 - Authority: 多くの良いhubから指されているページ



良いAuthorityとHubの集まりをコミュニティと呼ぶ

Hub, 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
 $auth(n) = \sum hub(m)$, for all $m \rightarrow n$
 $hub(n) = \sum auth(m)$, for all $m \leftarrow n$



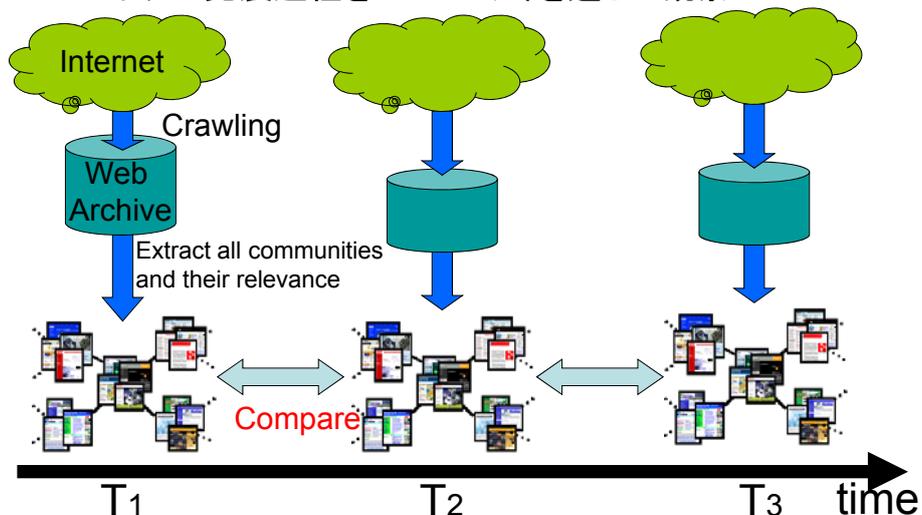
あらまし

- Webアーカイブ基盤
- Web空間の構造俯瞰
- **Webの時系列分析**
- Webを用いた自然言語処理
- Webスパムの分析

ウェブの時系列分析

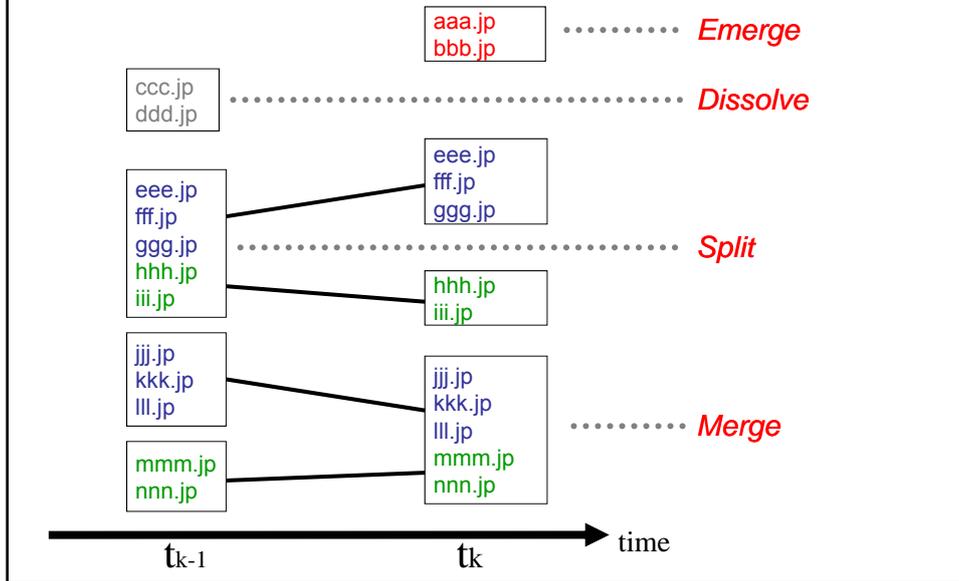
[ACM Hypertext 03]

- 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



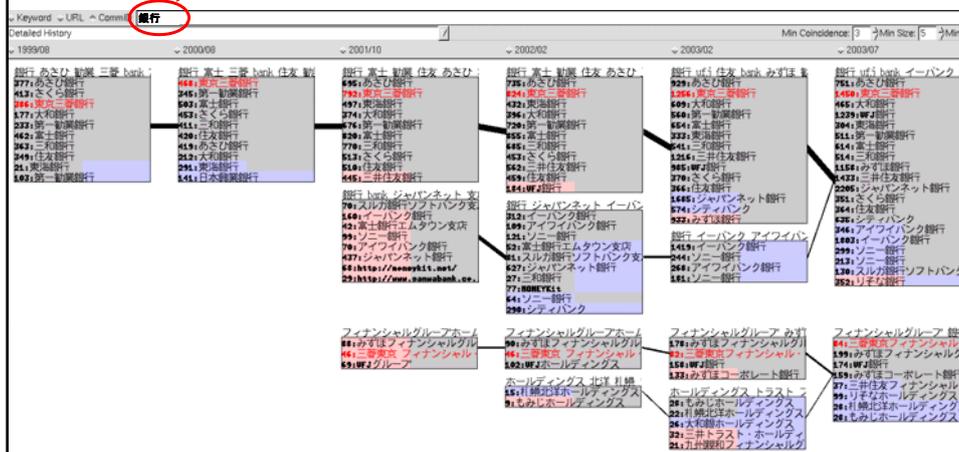
成果②ウェブの時系列分析

～銀行業界の変遷～



- 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

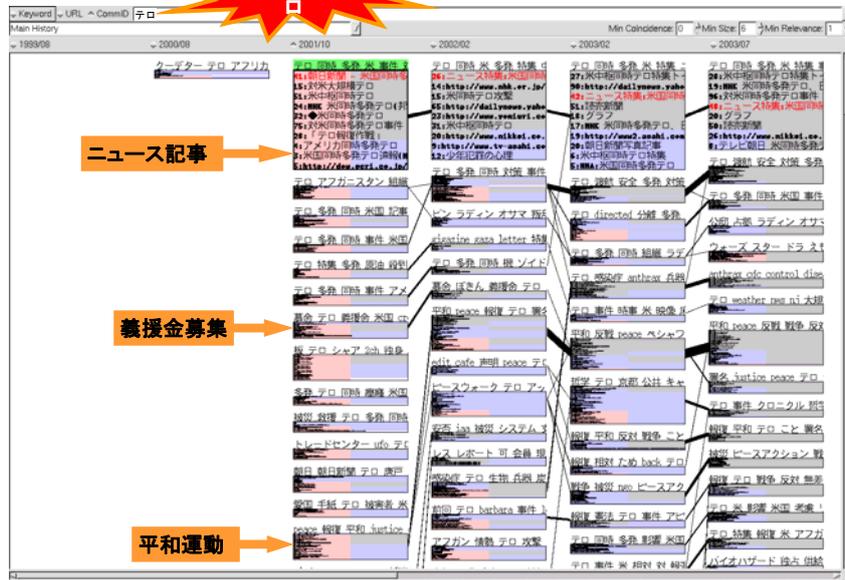
検索キーワード



成果②ウェブの時系列分析

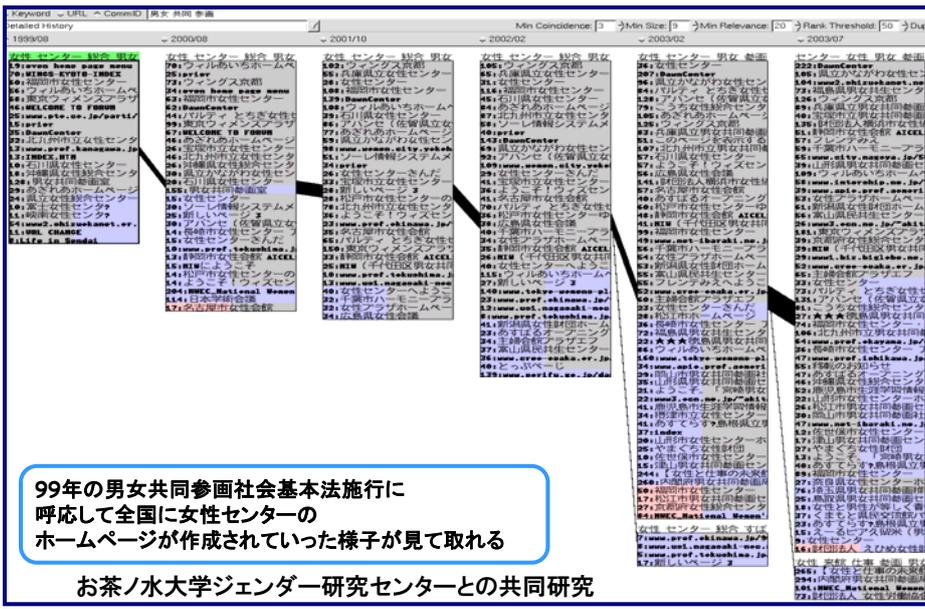
～社会現象による話題の爆発的発生～

同時多発テロ



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～

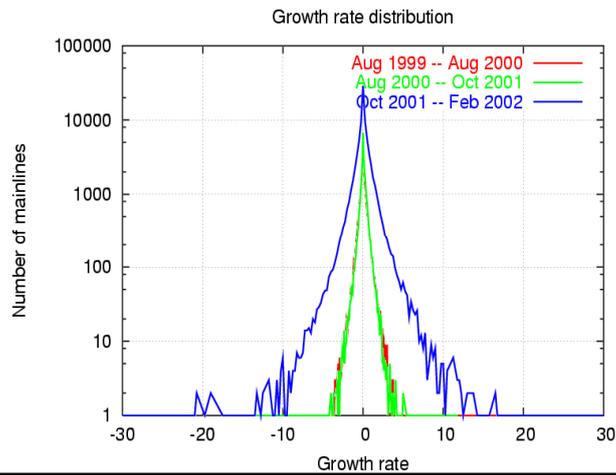


99年の男女共同参画社会基本法施行に呼応して全国に女性センターのホームページが作成されていった様子が見取れる

お茶ノ水大学ジェンダー研究センターとの共同研究

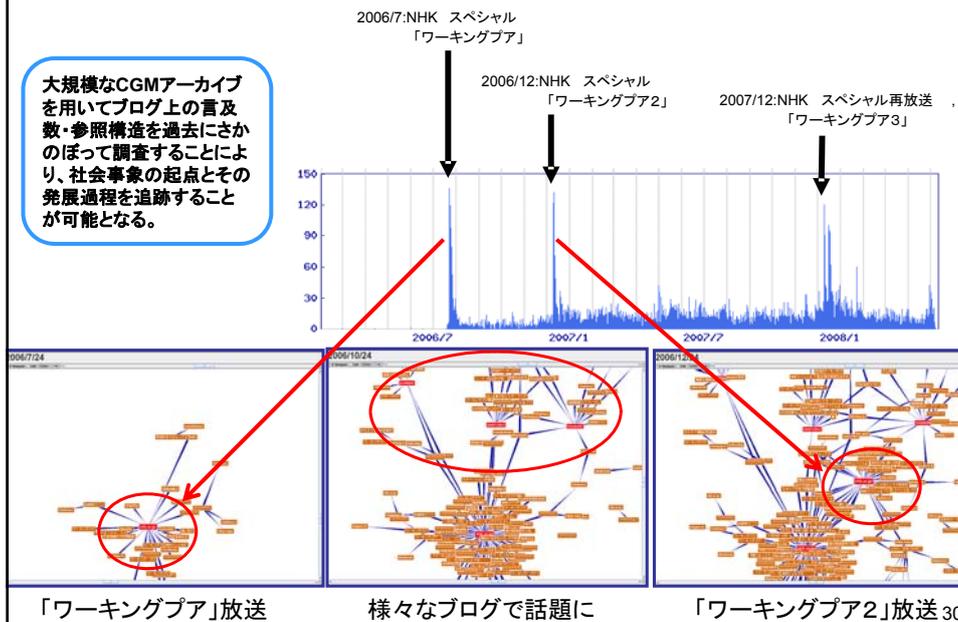
Grown and Shrunken Communities

- Growth rate have clear y-axis symmetry



ウェブの時空間分析 — ワーキングペア問題

大規模なCGMアーカイブを用いてブログ上の言及数・参照構造を過去にさかのぼって調査することにより、社会事象の起点とその発展過程を追跡することが可能となる。

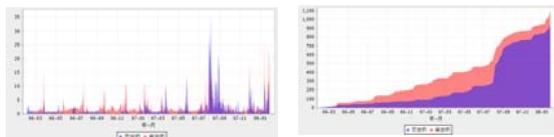


あらまし

- Webアーカイブ基盤
- Web空間の構造俯瞰
- Webの時系列分析
- **Webを用いた自然言語処理**
- Webスパムの分析

評判情報抽出による世論の分析 ～朝青龍の例～

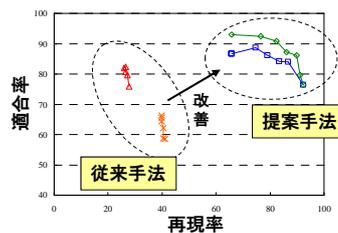
好不評書き込み数の通時的変遷



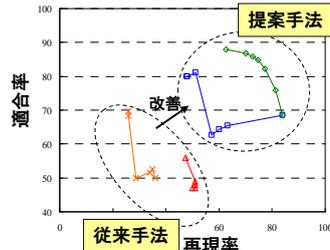
抽出された
書き込み

朝青龍張り手強 朝青龍好き 朝青龍品が無い
朝青龍相撲良い 朝青龍立ち合い悪すぎる
朝青龍良い 朝青龍稽子が良い 朝青龍
巻き替え速い 朝青龍速い 朝青龍強い 朝
青龍時天空速い 朝青龍映像まし 朝青龍
横綱面白くない
朝青龍八百長問題うやむや 朝青龍振る舞
い悪い 朝青龍強い 朝青龍つらい 朝青龍
良い 朝青龍精神年齢低すぎる 朝青龍まず
い 朝青龍何故
朝青龍品が無い 朝青龍粘り強い
朝青龍強い 朝青龍好き 朝青龍ショック
が大きい 朝青龍態度不愉快 朝青龍寂しい
朝青龍ホップ 朝青龍石舂やばい

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

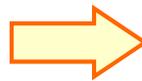
評価文コーパス

〈好評〉機種が多く、接写能力が高い。
 〈不評〉販売価格が高くなりがちだ。
 〈不評〉ソフトの価格が高かった。

 〈好評〉丈夫でちっとも壊れない。
 〈好評〉ドゥカティ製で壊れにくい。
 〈不評〉壊れやすそうな気がする。

評価表現辞書

極性値	評価表現
2.99	能力が高い
-3.07	価格が高い
2.58	壊れない
1.55	壊れにくい
-3.71	壊れやすい
...	...



言語解析
+
統計処理



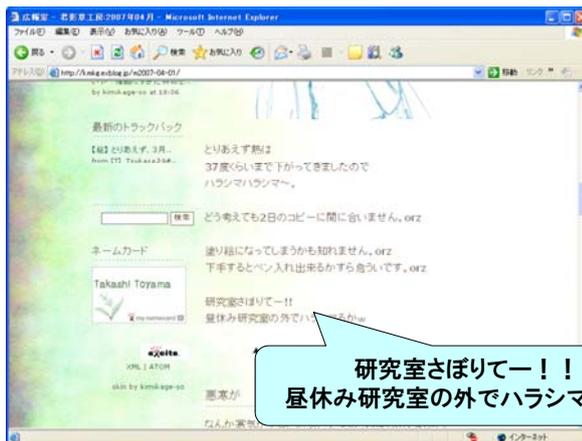
自動抽出

大規模ウェブアーカイブ



33

新語抽出



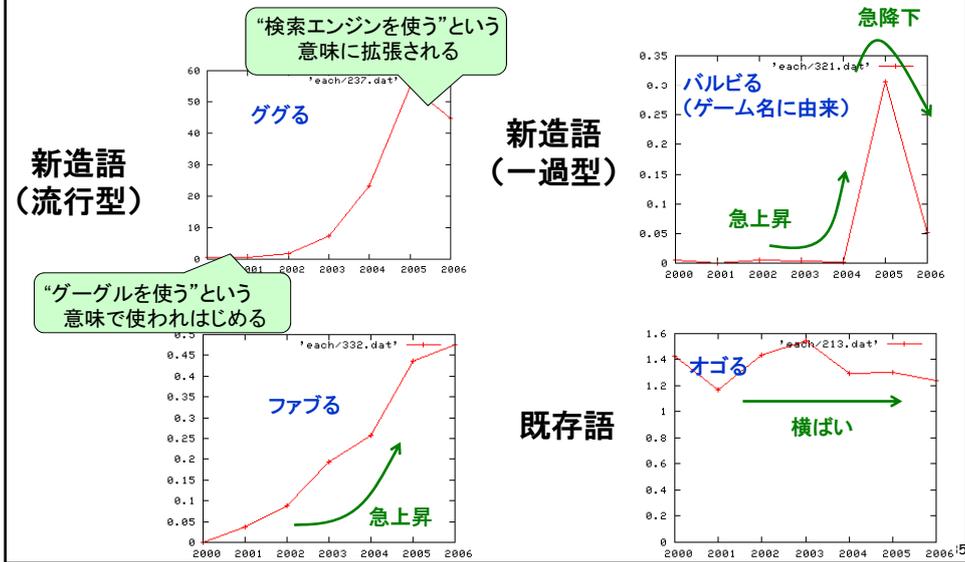
はらしま・る【ハラシマる】（動詞-五段ラ行）

- (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→原稿→はらしま」という誤読.
 一連休は腰を据えてハラシマるぞ.

34

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



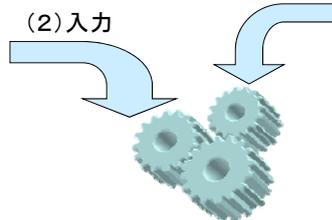
ウェブからの新造語獲得手法 機械学習を用いたカタカナ用言抽出

「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
.....

(1) 学習データとして与える

(2) 入力



(3) 品詞を出力

動詞

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	
.....

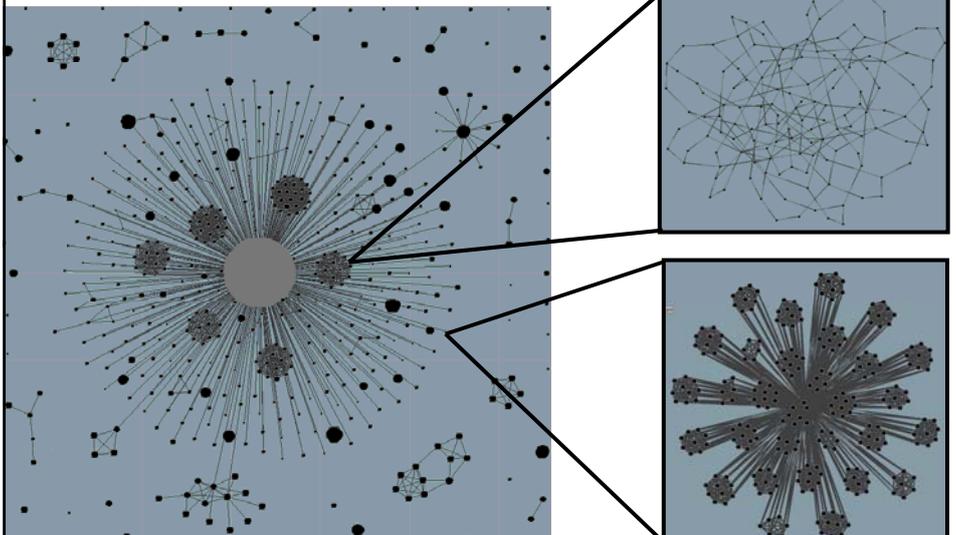
る	300
らない	120
ない	3
った	450
が	0
を	3
.....

あらまし

- Webアーカイブ基盤
- Web空間の構造俯瞰
- Webの時系列分析
- Webを用いた自然言語処理
- **Webスパムの分析**

検索エンジンスパムの構造分析

強連結成分分解、極大クリーク列挙 等のグラフアルゴリズム
を用いて大局的な**リンクスパム構造**を分析



『日本広報学会第12回研究発表大会』発表資料

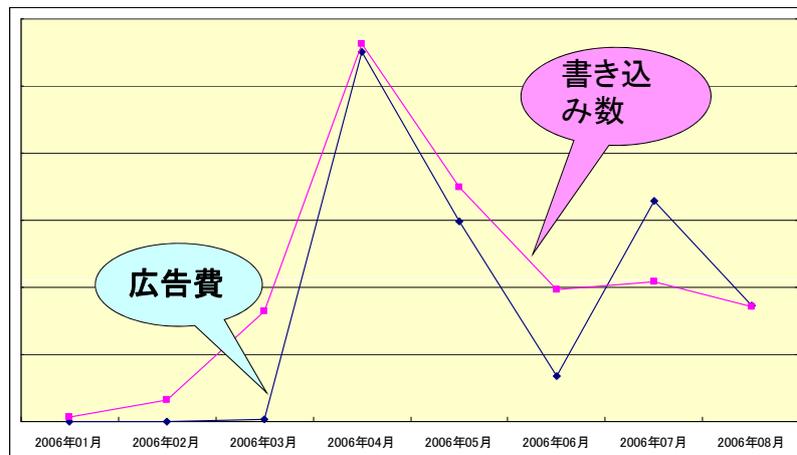
ブログからレピュテーション分析の可能性を 探る

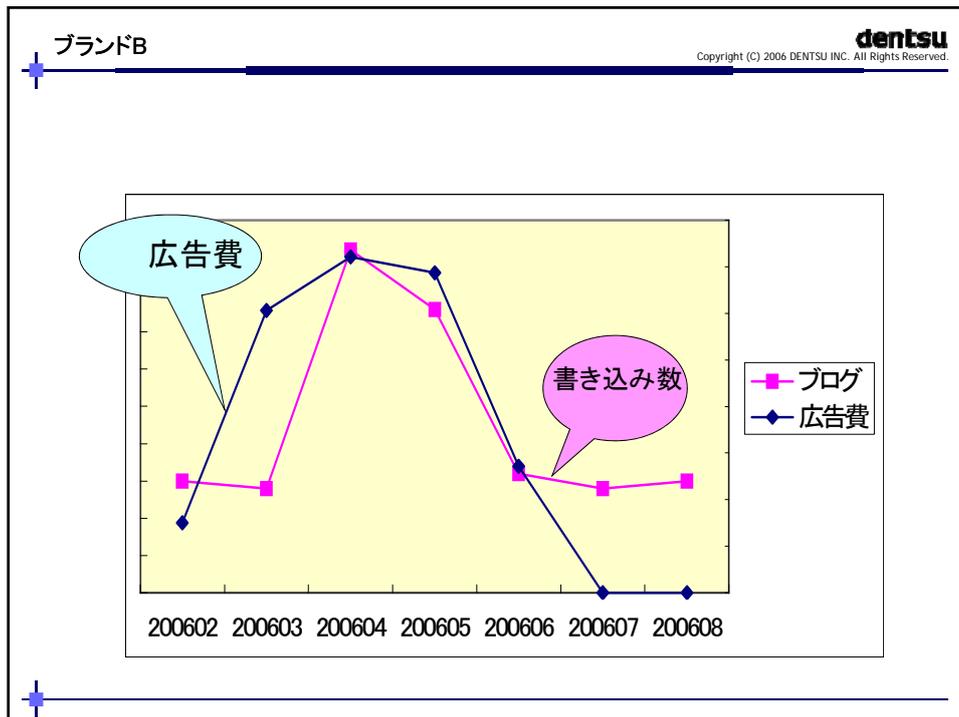
__Web2.0時代の新たな方法論へのトライアル__

2006年11月19日

株式会社電通 馬渡一浩
株式会社電通 富田英裕
専修大学経営学部 新井 範子
東京大学生産技術研究所 豊田 正史
東京大学生産技術研究所 鍛冶 伸裕
東京大学生産技術研究所 喜連川 優

ブランドA





<http://www.tkl.iis.u-tokyo.ac.jp/~toyoda/lecture/sougou2010.html>

まとめ

- 検索エンジンとSocio-Senseとの違い
 - 検索エンジン: **今のウェブ**をリアルタイムに検索することに集中(twitter検索も同様の流れ)
 - Socio-Sense: **過去から現在にいたるウェブの変遷**から価値ある情報を見つけ出す
- Socio-Sense
 - ウェブアーカイブ基盤
 - 大規模なウェブの構造俯瞰
 - ウェブ構造の時系列変化の追跡
 - 大規模自然言語処理(評判抽出、新語抽出…)
 - ウェブスパムの分析