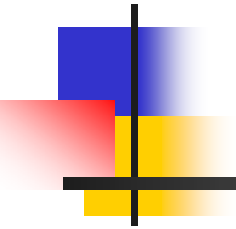
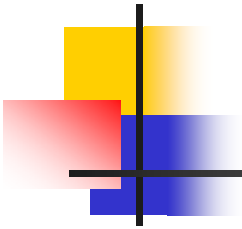


Webダイナミクスを利用した情報探索支援

NTT未来ねっと研究所

風間一洋



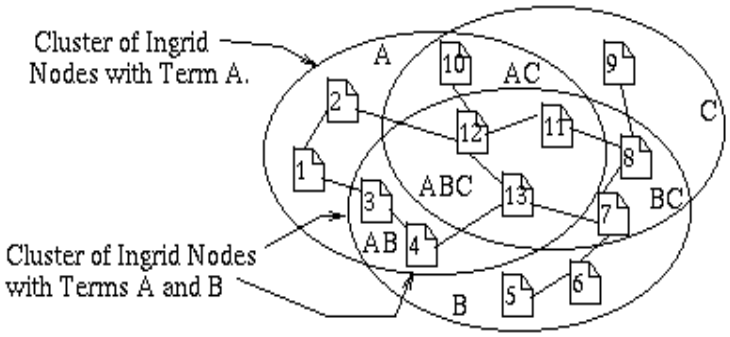
発表概要

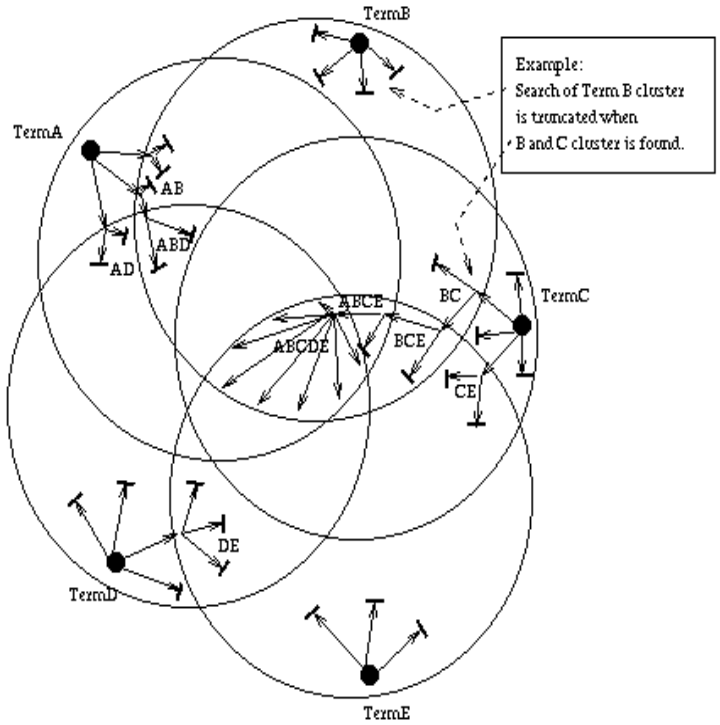
- 分散情報探索インフラストラクチャ
- リンク解析によるランキングの改善
- リンクによる関連ページの発見
- Web空間からの人間関係の発見
- Blogのトラックバックネットワーク解析
- 企業におけるネットワーク科学研究

分散情報探索インフラストラクチャ

- Ingrid (INformaton GRID)
  - [Paul et al. 1995]
  - 特徴語によって結び付けられたWebページのできる限り疎なネットワーク構造(=Ingridトポロジ)を作成
  - Ingridトポロジ上を分散探索
- 情報量に対してリンク数は比較的早く飽和する?

Ingridトポロジと情報探索



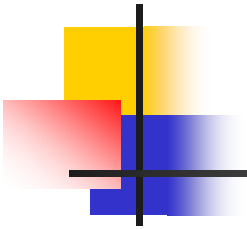


リンク解析によるランキングの改善

- 昔「サーチエンジンは使い物にならない」
→今「とりあえずググれ!」
- 地位向上の理由
  - リンクの非参照関係→内容の信頼度
  - アンカーテキスト＝アノテーション

ランキングの課題

- 膨大な検索結果→上位しか見ない/見れない
  - 精度は重要
    - 検索誤りが多いと目的の情報を探せない
  - **再現率(検索漏れ)はあまり重要ではない**
    - 信頼度が低い類似情報は欠落してかまわない
- 検索者は素人→うまく絞込めない
  - AltaVista: 2.35語 [Silverstein 1998]
  - ODIN: 1.42語 [風間 et al. 2000]

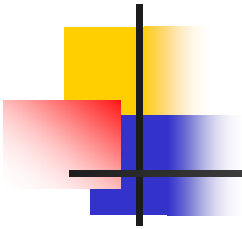


リンク構造の利用

- リンク解析
  - Link Popularity
  - PageRank
- アンカーテキスト

Link Popularity

- あるWebページの近傍のリンク構造の解析
- 被リンク数に応じてスコアをブースト
  - ページ単位
  - サーバ単位
- **検索語とは無関係に計算**

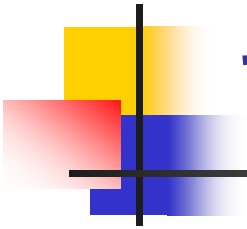


PageRank

- [Brin & Page 1998]
- Web空間全体のリンク構造を解析
- Random Surfer Model
  - 「多くの良質なWebページから参照されているWebページほど良質」
  - Web全体を行動する利用者の閲覧確率
  - 出口のないページの存在
    - 15%の確率でジャンプすることで回避
- 検索語とは無関係に事前に計算

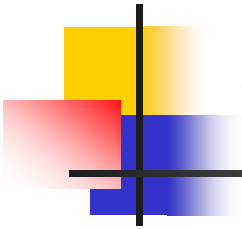
トピックドリフト

- [Baharat & Henzinger 1998]
- 元のトピックとの関連性が低いWebページ群が得られること
- 原因
  - トピックの一般化
    - 検索語「カローラ」→各メーカーのホームページ
    - 検索語と無関係に計算されるために，被参照数に影響されやすい
  - 自動生成されたリンクの影響
    - ナビゲーションメニュー
    - Webサイト設計が影響
  - 相互リンク、(リンクベースの)スパムの影響



アンカーテキスト

- アンカーテキスト=HTMLアンカー部のテキスト
- World Wide Web Worm [McBryan 1994]
- 文書を指すリンクのアンカーテキストも索引付け
  - 未収集ページの検索
  - 非テキスト情報(画像など)の検索
  - 多彩な表現・スペルミス対応
- トピック依存のLink Popularity
  - 高い適合度(内容の的確・簡潔な要約)
  - 検索語数が少ない場合に特に有効



Google

- [Brin & Page 1998]
- アンカーテキスト
  - “I’m Feeling Lucky”ボタン
    - 望む情報が1位である確率が高い
- PageRank
  - 一般的なリンクの信頼度を副次的に反映
  - スパム対策に有効

ODIN(1)

- 公開実験(1999～2002)
- アンカーテキスト
  - サーバ内部のリンク→重み小
    - 文書構造を反映
  - サーバ外部からのリンク→重み大
    - 他人からの推薦を反映
- 機械的なリンク、リンクスパム対策
  - 上限値の設定
- オフィシャルサイトの平均順位 1.42 [風間 et al. 2000]

ODIN(2)

チーズ の検索結果: 810件 / 330 グループ (1 - 10 を表示)

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [前の10件] [次の10件]

1 日本のナチュラルチーズ [畜産振興事業団] (スコア: 3.698)
ナチュラルチーズ入門 ナチュラルチーズの種類 ナチュラルチーズの製造方法 ナチュラルチーズの歴史 チーズとワインのマリアージュ チーズデータあれこれ 日本のナチュラルチーズ・マップ 中央酪農会議 ホームページ ミルククラブ ホームページ スタンプラリー ホームページ ご意見・ご感想・ご質問は、
products@jdc.lin.go.jp （社）中央酪農会議 〒100-0004 東京都千代田区大手町1-6-3
http://jdc.lin.go.jp/report/index.htm (最終更新: 1999/03/08)
1. Report of an Investigation http://jdc.lin.go.jp/report/
2. Report of an Investigation http://jdc.lin.go.jp/report/index.html
3. Report of an Investigation http://jdc.lin.go.jp/report/index2.html
グループをすべて表示 (65件)

2 月報（国内編） [畜産振興事業団] (スコア: 3.449)
ALIC Monthly 月報「畜産の情報」（国内編） 1998年1０月 目 次 【今月の話題】 食品の安全性確保と

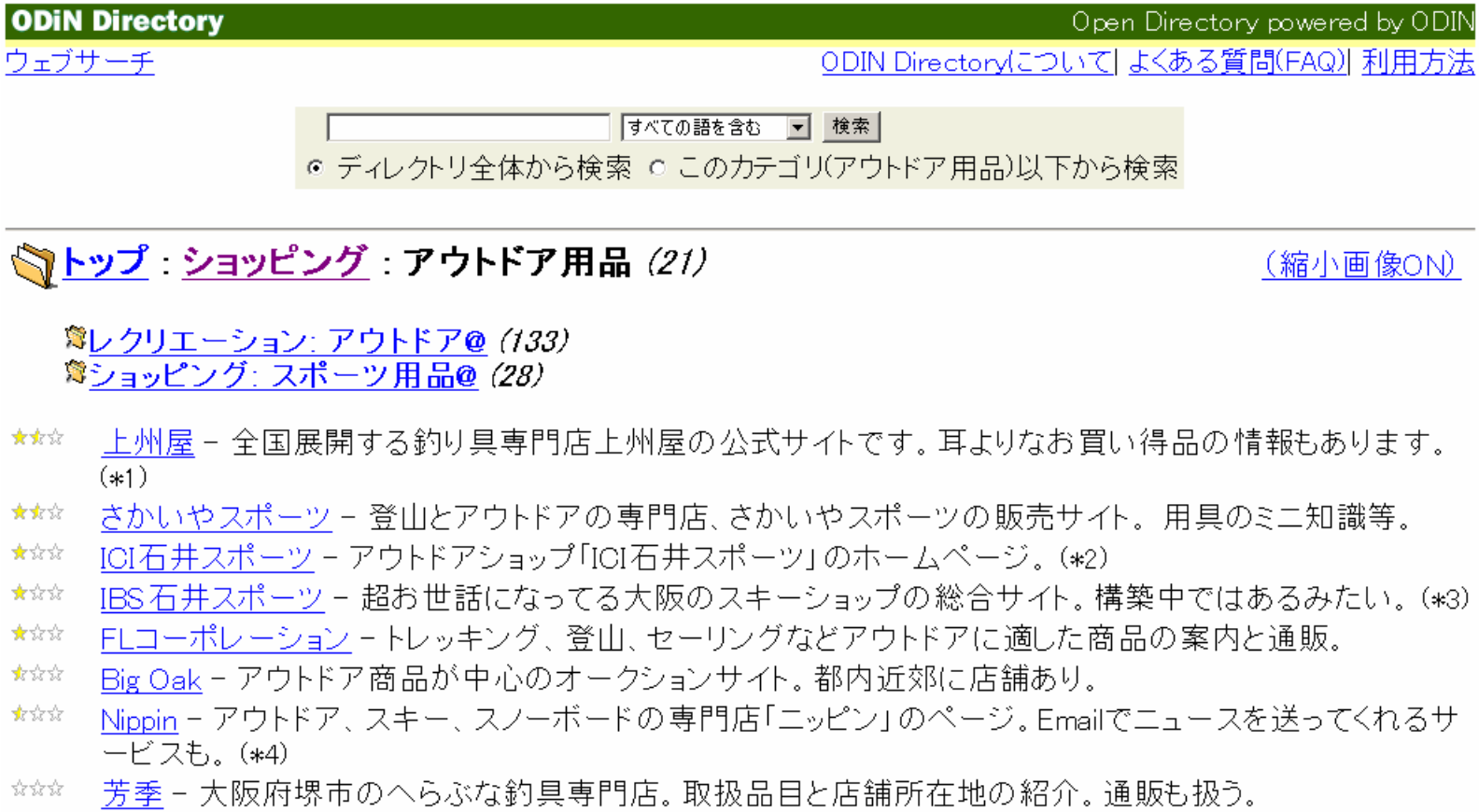
リンクによる関連ページの発見

- **リンクの参照関係→内容の類似性**
- Companionアルゴリズム[Dean & Henzinger 1998]
  - HITSアルゴリズム
- Cocitationアルゴリズム[Dean & Henzinger 1998]
  - 共参照(cocitation)関係の抽出
    - 共通のリンク元に存在するURL
    - ページ内のリンクの位置が一定距離以内
- **MultiCocitationアルゴリズム**[原田 et al. 2000]
  - 補正項の追加→トピックドリフトを抑制
    - 一つの突出したサイトの影響を弱める
    - 多くのサイトとの関連性を持つ場合を高く評価
  - 複数シード化

ODIN Directory(1)

- [風間 et al. 2004]
- 公開実験(2001/4～2002/4)
- Open Directory Projectのデータを使用
- 特徴
  - Webディレクトリのカテゴリを豊富化
    - 各カテゴリのサイト群から関連サイトを導出
    - MultiCocitationアルゴリズムを使用
  - **説明文の自動抽出**

ODIN Directory(2)



ODiN Directory — Open Directory powered by ODIN

ウェブサーチ

ODIN Directoryについて | よくある質問(FAQ) | 利用方法

すべての語を含む 検索

ディレクトリ全体から検索 このカテゴリ(アウトドア用品)以下から検索

トップ : ショッピング : アウトドア用品 (21)　(縮小画像ON)

レクリエーション: アウトドア@ (133)
ショッピング: スポーツ用品@ (28)

★★☆ 上州屋 - 全国展開する釣り具専門店上州屋の公式サイトです。耳よりなお買い得品の情報もあります。(*1)
★★☆ さかいやスポーツ - 登山とアウトドアの専門店、さかいやスポーツの販売サイト。用具のミニ知識等。
★☆☆ ICI石井スポーツ - アウトドアショップ「ICI石井スポーツ」のホームページ。(*2)
★☆☆ IBS石井スポーツ - 超お世話になってる大阪のスキーショップの総合サイト。構築中ではあるみたい。(*3)
★☆☆ FLコーポレーション - トレッキング、登山、セーリングなどアウトドアに適した商品の案内と通販。
☆☆☆ Big Oak - アウトドア商品が中心のオークションサイト。都内近郊に店舗あり。
☆☆☆ Nippin - アウトドア、スキー、スノーボードの専門店「ニッピン」のページ。Emailでニュースを送ってくれるサービスも。(*4)
☆☆☆ 芳季 - 大阪府堺市のへらぶな釣具専門店。取扱品目と店舗所在地の紹介。通販も扱う。

ODIN Directory(3)

※以下のページから紹介文を引用させていただきました．（→解説）

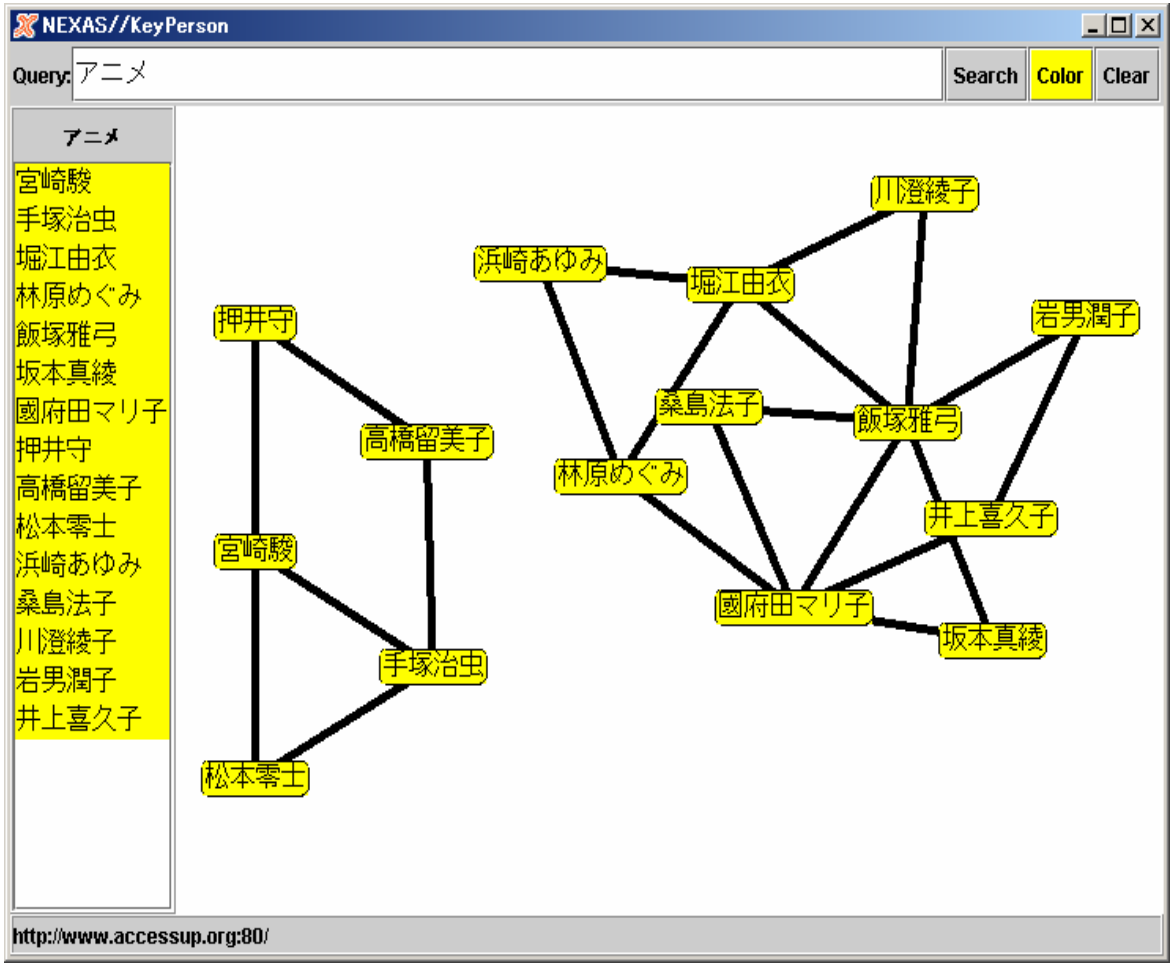
*1 http://www.jet.ne.jp/salon/link2/200105.html
*2 http://www2.wbs.ne.jp/~caribian/link.htm
*3 http://isweb6.infoseek.co.jp/sports/hiro3110/ski-links.html
*4 http://yokohama.cool.ne.jp/maruhide/21.html

人の手による世界最大のウェブ・ディレクトリの編纂にご参加ください。
サイトを追加 - **Open Directory Project** - エディタになろう
このカテゴリにはエディタがいません．

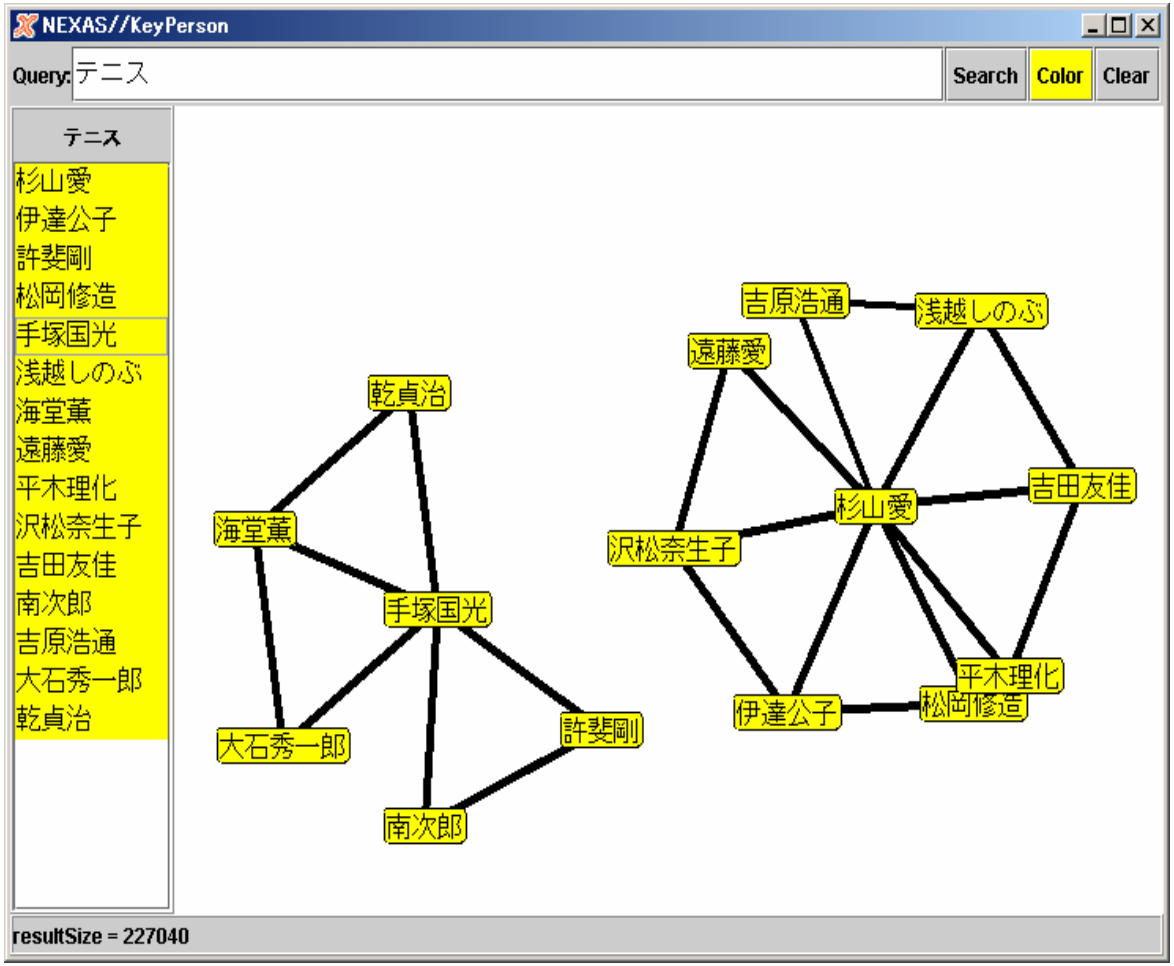
Web空間からの人間関係の発見

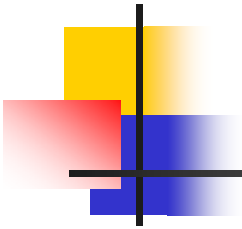
- 実世界指向情報探索
  - 実空間の実体に結び付けられているWeb空間の固有表現を抽出
  - 固有表現の関係を解析して利用
- NEXAS [原田 et al. 2003]
  - Named Entity eXtraction and Association Search
  - プロトタイプ: 固有表現として人名を使用

アニメ（739,160 URL）



テニス（227,040 URL）

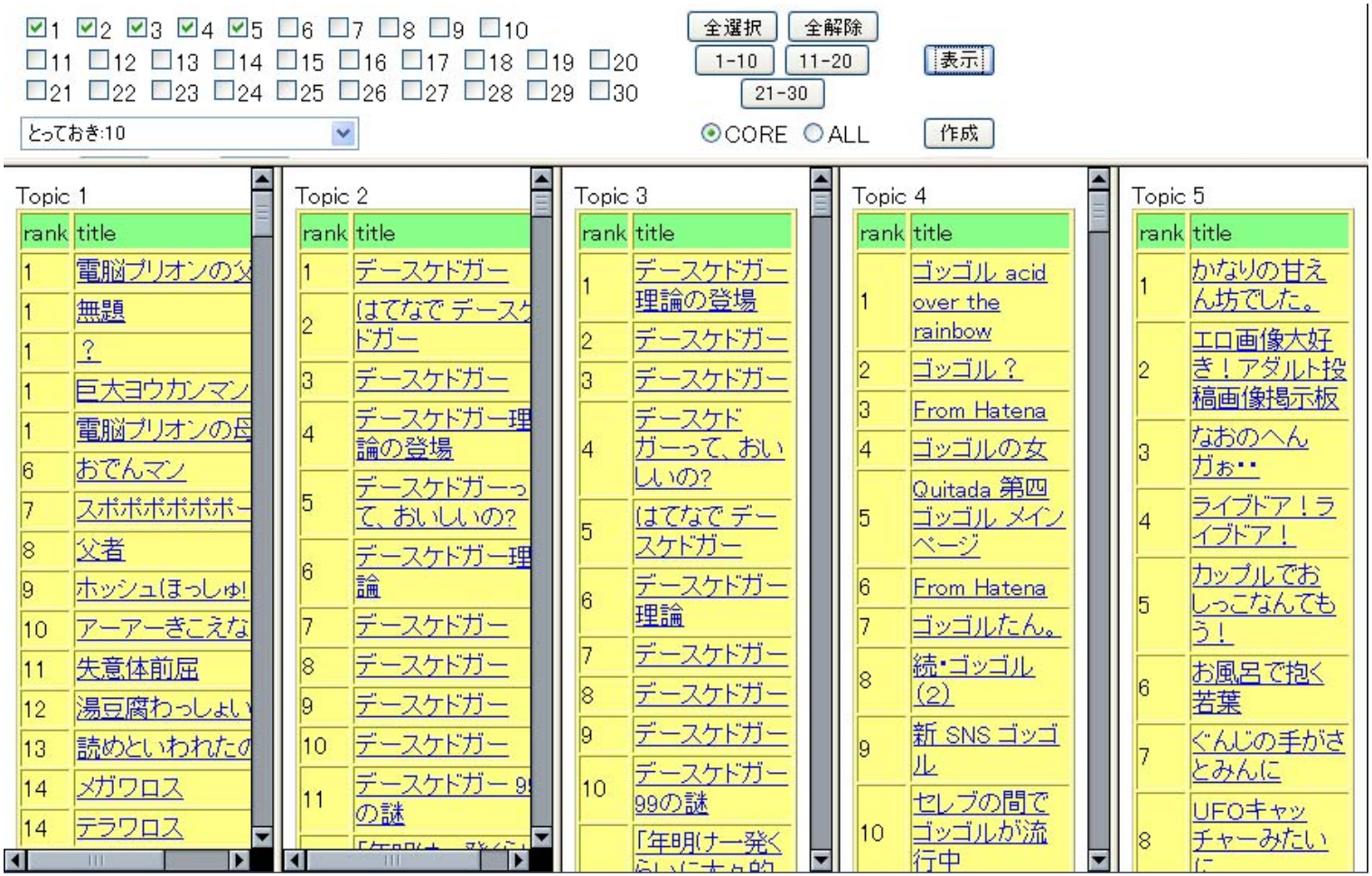




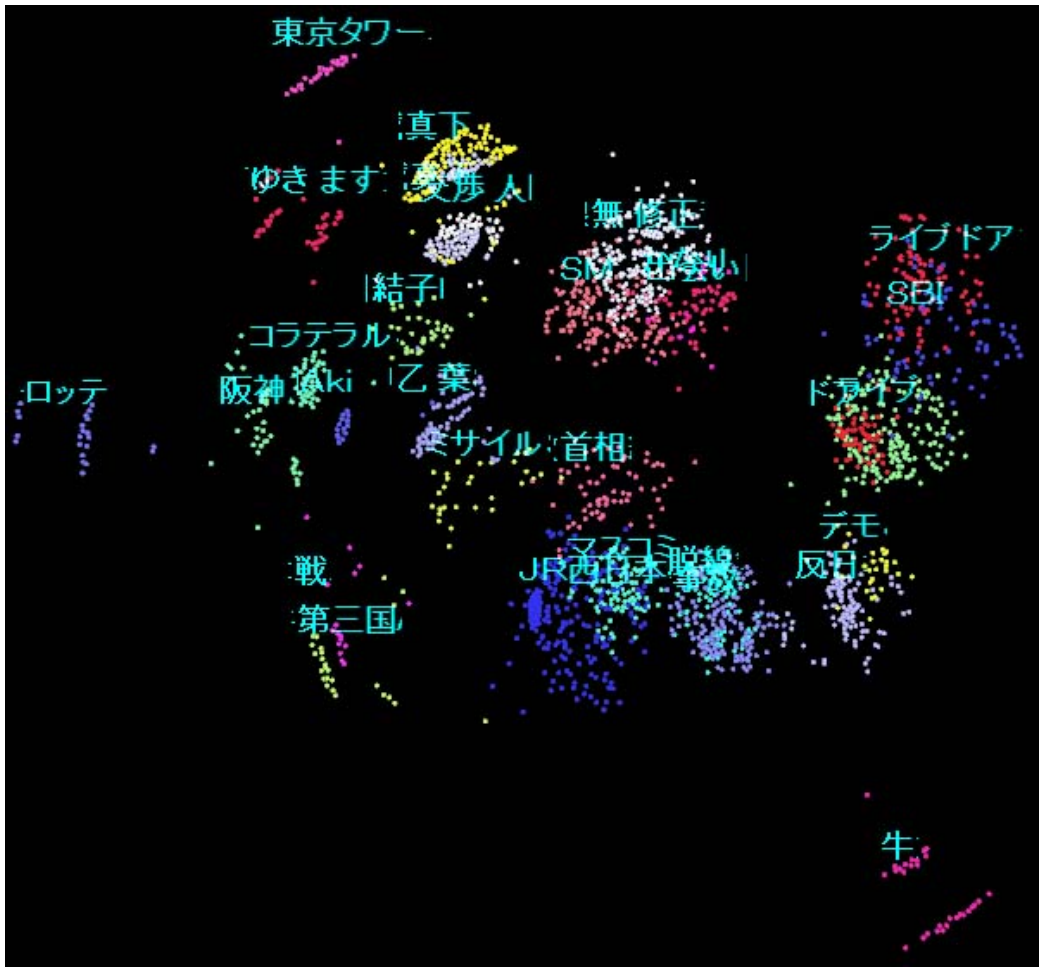
Blogのトラックバックネットワーク解析

- ブログのトラックバック・ネットワークの構造の解析
  - 主要トピックまたはコミュニティを抽出
  - トラックバックスパムの検出
- SR法 [斉藤 et al. 2005]
  - 新しいスペクトラルグラフ分析手法
  - ノード間の結合が比較的密でも分類可能
  - 同一ノードを複数のトピックに分類可能

分類されたトピック


1 2 3 4 5 6 7 8 9 10
11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30
全選択
全解除
1-10
11-20
21-30
表示
とっておき:10
CORE
ALL
作成
Topic 1
rank
title
1 電脳ブリオンの父
1 無題
1 ？
1 巨大ヨウカンマン
1 電脳ブリオンの母
6 おでんマン
7 スポポポポポポー
8 父者
9 ホッシュほっしゅ!
10 アーアーきこえな
11 失意体前屈
12 湯豆腐わっしょい
13 読めといわれたの
14 メガワロス
14 テラワロス
Topic 2
rank
title
1 データケドガー
2 はてなで データケドガー
3 データケドガー
4 データケドガー理論の登場
5 データケドガーって、おいしいの?
6 データケドガー理論
7 データケドガー
8 データケドガー
9 データケドガー
10 データケドガー
11 データケドガー 99の謎
Topic 3
rank
title
1 データケドガー理論の登場
2 データケドガー
3 データケドガー
4 データケドガーって、おいしいの?
5 はてなで データケドガー
6 データケドガー理論
7 データケドガー
8 データケドガー
9 データケドガー
10 データケドガー 99の謎
「年明け一発く
Topic 4
rank
title
1 ゴッゴル acid over the rainbow
2 ゴッゴル？
3 From Hatena
4 ゴッゴルの女
5 Quitada 第四ゴッゴル メインページ
6 From Hatena
7 ゴッゴルたん。
8 続・ゴッゴル (2)
9 新 SNS ゴッゴル
10 セレブの間でゴッゴルが流行中
Topic 5
rank
title
1 かなりの甘えん坊でした。
2 エロ画像大好き！アダルト投稿画像掲示板
3 なおのへんガぉ・・
4 ライブドア！ライブドア！
5 カップルでおしっこなんてもう！
6 お風呂で抱く若葉
7 ぐんじの手がさとみんに
8 UFOキャッチャーみたいに

トピックマップ


東京タワー
真下
ゆきます
無修正
ライブドア
SBI
結子
コラテラル
ロッテ
阪神
乙葉
ドライブ
サイル
首相
デモ
戦
第三国
牛

企業のネットワーク科学研究

- Web関連サービスのIT企業
- 対象となりそうな分野
  - Web情報検索
  - Blogサービス
  - ソーシャルネットワーキングサイト
  - ソーシャルブックマーク
  - 評判検索
- 目的
  - Web情報空間の状況の把握
  - 他サービスとの差別化
  - ユーザ支援

参考資料(1)

- [Paul et al. 1995] Paul Francis, Takashi Kambayashi, Shin-ya Sato and Susumu Shimizu : "Ingrid: A Self-Configuring Information Navigation Infrastructure", 4th International World Wide Web Conference, 1995.
- [風間 et. al. 2000] 風間 一洋, 原田 昌紀, 佐藤 進也 : サーチエンジンの検索結果のマルチレベルグルーピングの評価, 日本ソフトウェア科学会コンピュータソフトウェア, Vol.17, No.4, pp. 58--69, 2000.
- [風間 et al. 2004] 風間 一洋, 原田 昌紀, 佐藤 進也: Webディレクトリ拡張の自動化手法, 情報処理学会論文誌:データベース, Vol. 45, No. SIG 7 (TOD22), pp.218--229, 2004.
- [原田 et al. 2003] 原田 昌紀, 佐藤 進也, 風間 一洋: Web上のキーパーソンの発見と関係の可視化, 情報処理学会研究会報告 DBS-130-03/FI-71-03, pp. 17--24, 2003.
- [斉藤 et al. 2005] 斉藤 和己, 木村 昌弘, 風間 一洋, 佐藤 進也: ブログ空間の主要トピック抽出, 人工知能学会研究会資料 SIG-KBS-A501-02, pp. 5--10, 2005.

参考資料(2)

- [Silverstein 1998] Craig Silverstein, Monika R. Henzinger, Hanns Marais and Moricz, Analysis of a Very Large AltaVista Query Log, Digital SRC, 1998.
- [McBryan 1994] Oliver A. McBryan: GENVL and WWWW: Tools for taming the Web, The 1st International Conference on the World Wide Web, 1994.
- [Brin & Page 1998] The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, Vol. 30, pp. 107—117, 1998.
- [Baharat & Henzinger 1998] Krishna Bharat and Monika R. Henzinger: Improved algorithms for topic distillation in a hyperlinked environment, Proceedings of SIGIR-98, pp. 104—111, 1998.
- [Dean & Henzinger 1998] Jeffrey Dean and Monika R. Henzinger: Finding related pages in the World Wide Web, Computer Networks, Vol. 31, pp. 1467—1479, 1999.