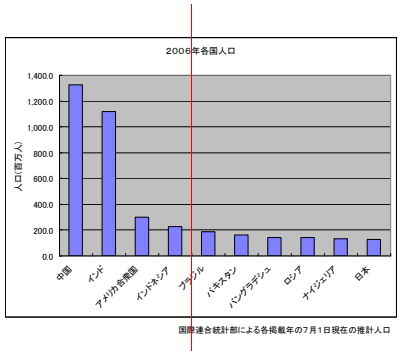
e-Society
文部科学省リーディングプロジェクト 基盤ソフトウェアの総合開発

研究のご紹介

東京大学　生産技術研究所　教授
文部科学官
情報処理学会副会長
文部科学省特定領域研究「情報爆発」領域代表
喜連川優

MySpace

特定領域研究 情報爆発



ワールドビジネスサテライト 2007. 7. 25

e-Society 文部科学省リーディングプロジェクト

Socio-Sense

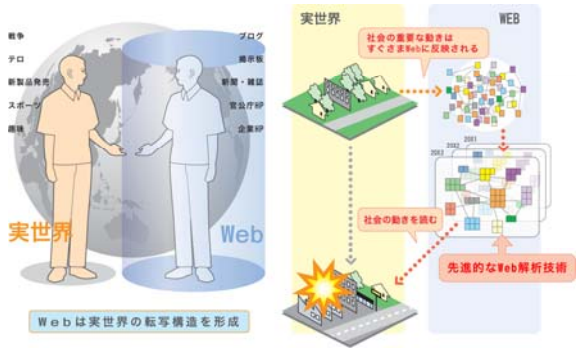
=社会のセンサーとしてのWEB=

研究代表者
喜連川 優
（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）

実社会の射影としてのウェブ

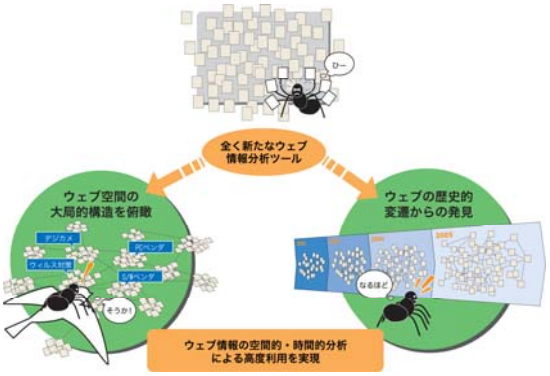
e-Society 文部科学省リーディングプロジェクト

ウェブは社会のセンサー



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

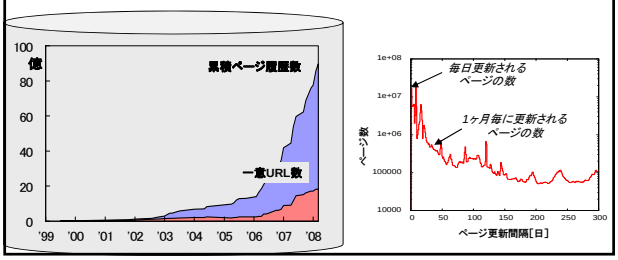


日本語ウェブアーカイブの構築

e-Society 文部科学省リーディングプロジェクト

6

- 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



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

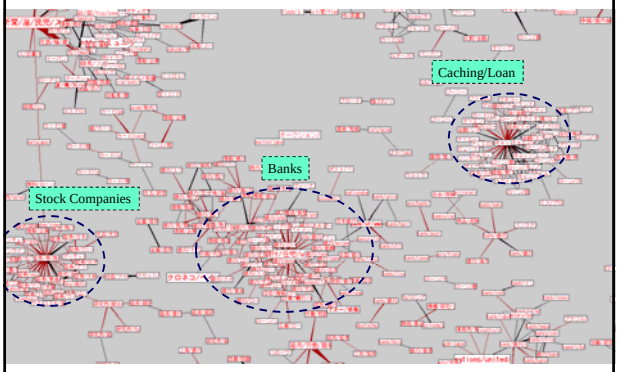
ウェブコミュニティチャート
・ある話題に関する有用なページの集合はウェブのリンク空間上で稠密な構造を持つ
・全コミュニティとそれらの関係を抽出して地図化

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

ウェブ空間の構造俯瞰（serendipity）
～コンピューター業界周辺の地図～

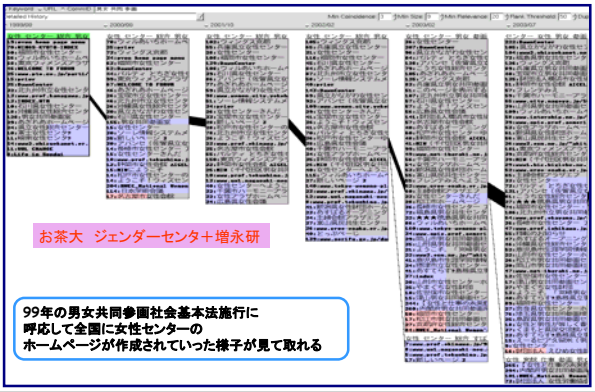
■リンク&テキスト解析を用いてウェブの全空間を地図化
■産業連関図に相当する地図が得られる
　➢注目分野のリサーチ・サーベイに有用
■影響力のある製品ユーザのグループなども同時に抽出
　➢広告設置戦略への応用

Cyber Map for Financial Business



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～

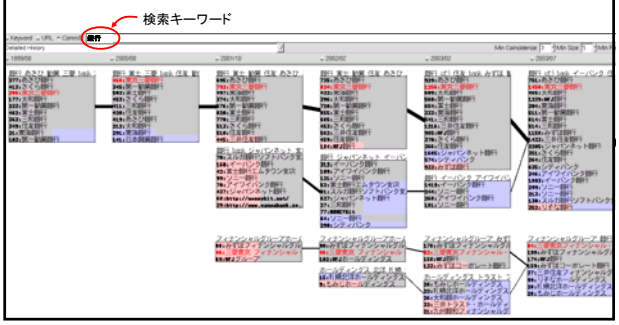


お茶大　ジェンダーセンター＋増永研

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

ウェブの時系列分析
～銀行業界の変遷～

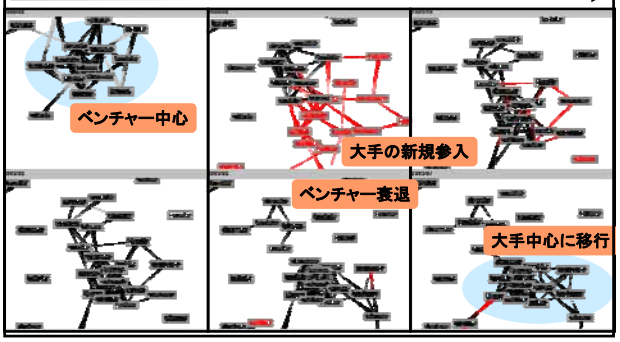
- インターネット銀行の出現と世間への浸透
- 合併した銀行の出現：三井住友、UFJ、みずほ、りそな



ウェブの時空間分析

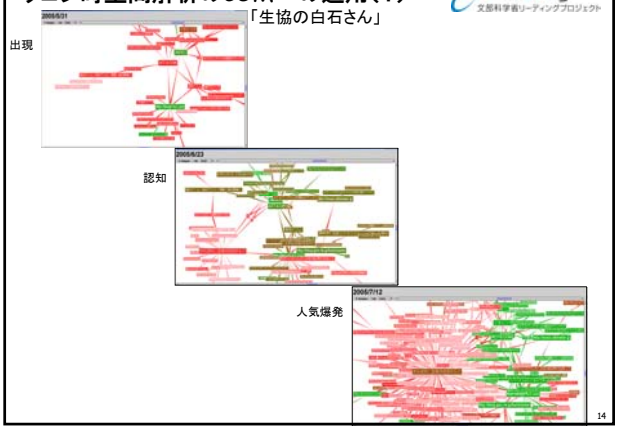
空間＋時間分析：コミュニティの変遷
（例：i-mode検索サイト）

時間



ウェブ時空間解析のCGMへの適用(1)

「生協の白石さん」



14

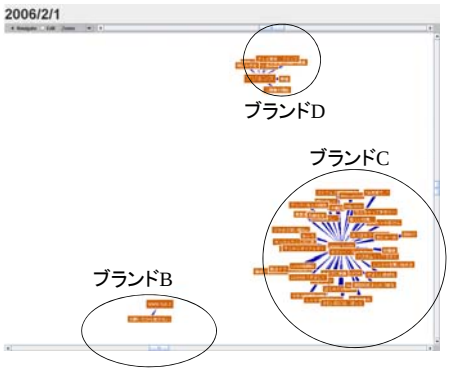
Definition of Novelty Measure

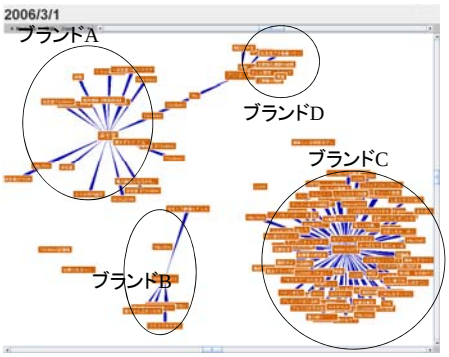
$$\mathcal{N}(p) = (1-\delta)\frac{\sum_{(q,p)\in I(p)} n(q,p)}{|I(p)|}$$

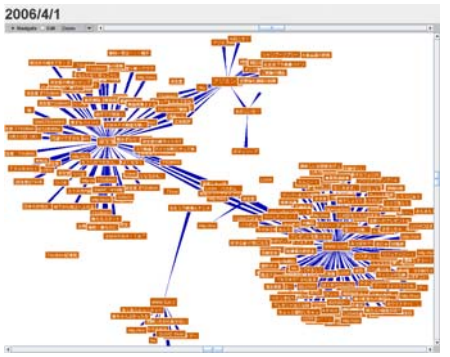
$$n(q,p) = \begin{cases} 1 & q \in L_2(t_k) \\ 0 & q \in O(t_k)\backslash L_2(t_k) \\ \mathcal{N}(q) & q \in U(t_k) \end{cases} \qquad (1)$$

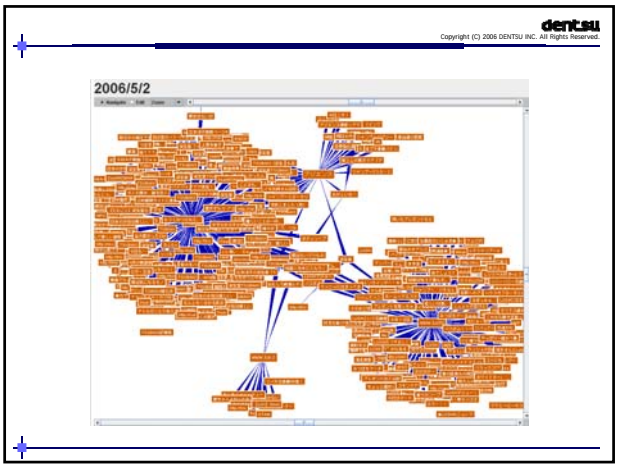
- δ: damping factor
  - probability that there were links to ***p*** before t-1

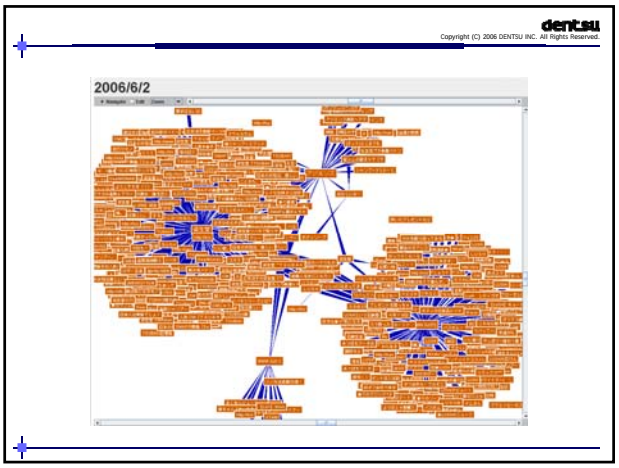
リンクの成長プロセス

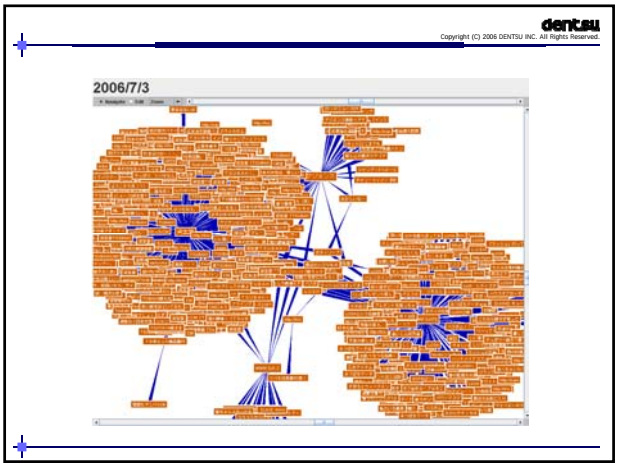


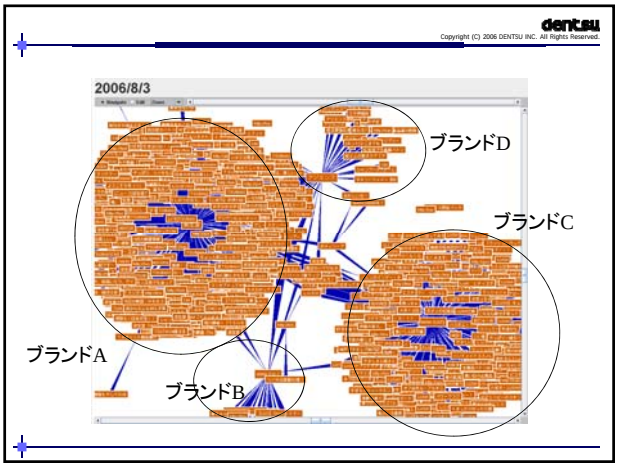


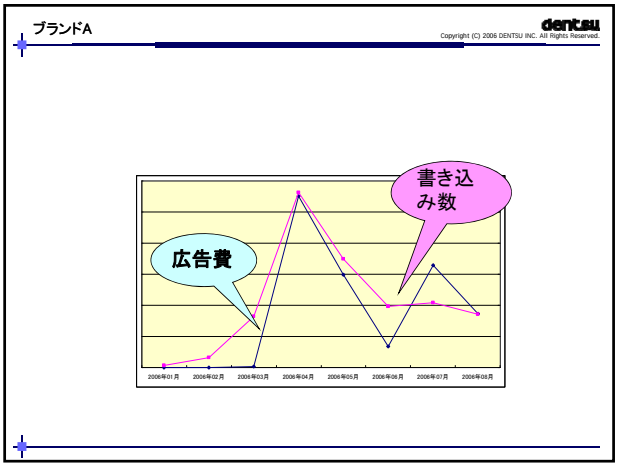


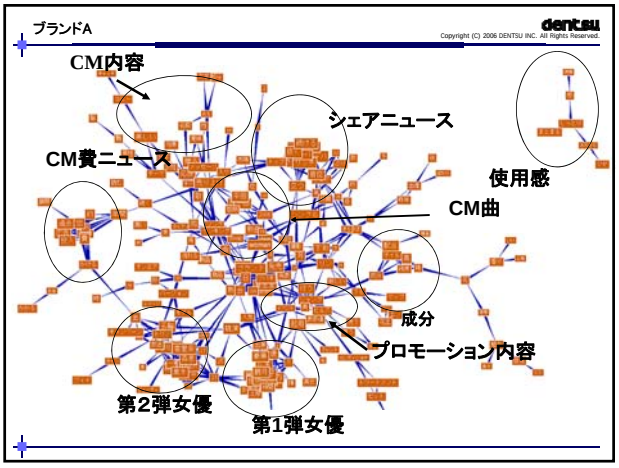

2006/5/2


2006/6/2


2006/7/3

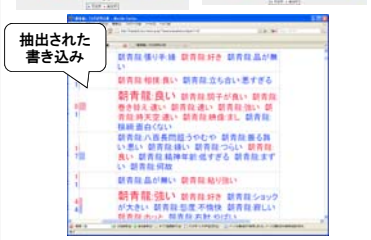

2006/8/3
ブランドD
ブランドC
ブランドA
ブランドB


ブランドA
書き込み数
広告費

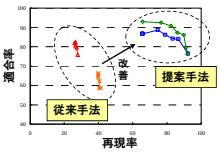

ブランドA
CM内容
シェアニュース
CM費ニュース
使用感
CM曲
成分
プロモーション内容
第2弾女優
第1弾女優

評判情報抽出による世論の分析
～朝青龍の例～

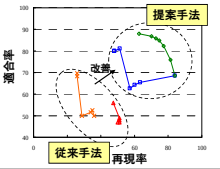
好不評書き込み数の通時的変遷



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〉 ドッカティ製で壊れにくい。
〈不評〉 壊れやすそうな気がする。
……

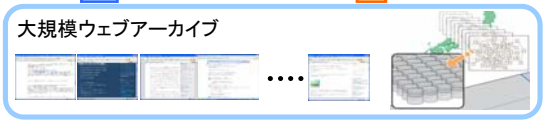
自動抽出

言語解析
＋
統計処理

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
…	…

大規模ウェブアーカイブ



WEBからの未知語（カタカナ用言辞書）の獲得

五段ラ行	一段	五段カ行	形容詞	形容動詞
ググる	モエる	トキメく	チャチい	サイコーだ
パモる	トロケる	ホザく	イナタい	オトコマエだ
ハラシマる	イケテる	ドツく	ヤヴァい	フルーティだ
ヒキコモる	シャガレる	シバく	カバエい	エイジレスだ
バザる	コジャレる	ムカツく	スゲい	メタメタだ
ストロベリる	ハッチャケる	イタダく	マンドクサい	フニャフニャだ
ガンガる	デケる	ワメく	ヘコい	レピッシュだ

動詞 5段 ラ行
はらしま・る
ハラシマる

動詞 5段 ラ行
ばも・る
パモる

動詞 5段 ラ行
もふ・る
モフる

「はらしまる」について

特定領域研究 情報爆発

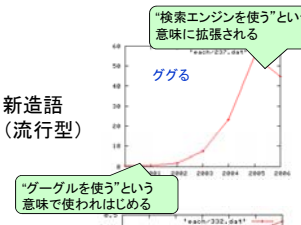
原稿
縞

- 業界でしか使わなかった言葉が発信されるようになりつつある。

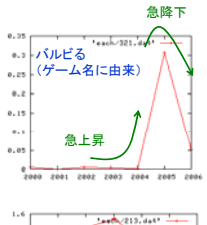
新造語に見られる意味拡張のダイナミズム

- 各種単語の時系列頻度の傾向を分析

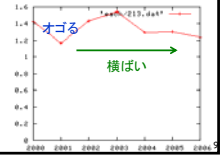
新造語（流行型）



新造語（一過型）



既存語



ウェブマイニング実証システム

- ディスプレイ壁上で分析ツール群を統合したプロトタイプを開発

New IT Infrastructure for the Information Explosion Era
Research Project funded by MEXT Grant-in-Aid for Scientific Research in Priority Areas (FY 2005－2010)

info-plosion

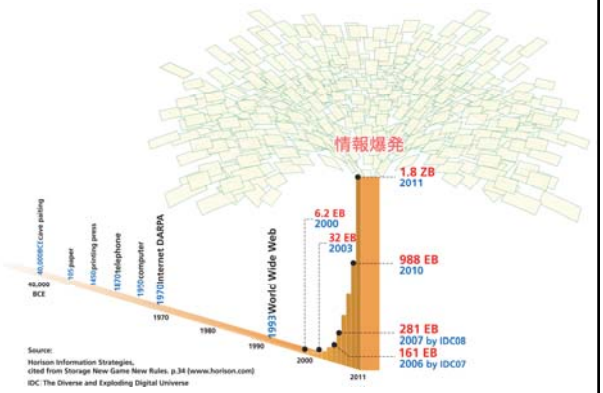
情報爆発時代に向けた新しいIT基盤技術の研究

平成17年発足　文部科学省特定領域研究

喜連川優
東京大学　生産技術研究所

The Information Explosion (Info-plosion)

info-plosion

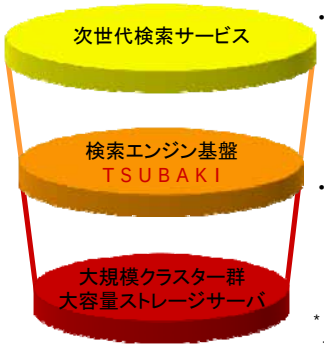


ディープNLPオープンサーチエンジン基盤

info-plosion 情報爆発

ＴＳＵＢＡＫＩ　（黒橋、新里）

（皆でいじれるコンポーザブル型爆発サーチエンジン）

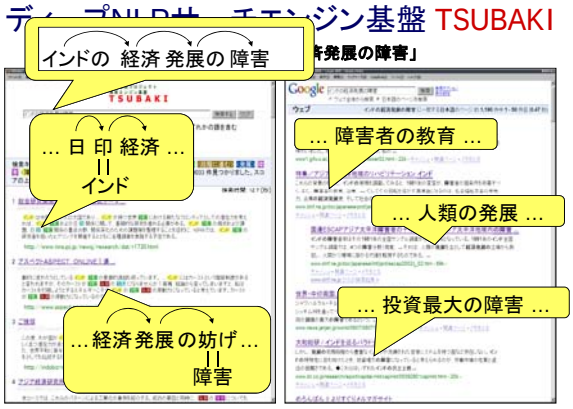


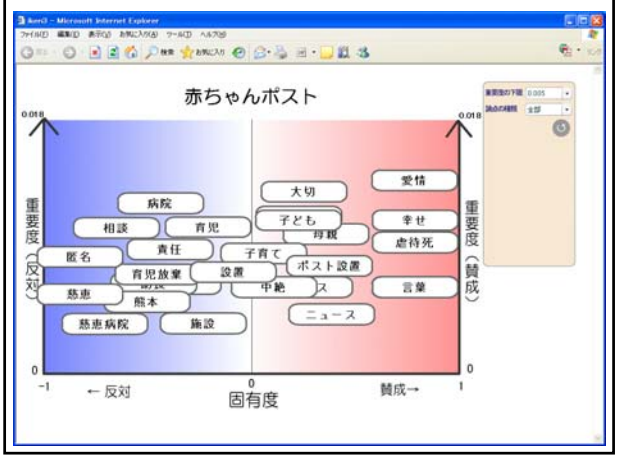
- 日本語ウェブ文書5,000万件を検索対象とした開放型検索エンジン基盤
  - 高度ウェブ処理用標準フォーマットによりウェブ文書を管理
  - 構造的言語処理によるインデックス
  - 無制限に利用可能なAPI
  - 透明性・再現性のある検索結果
- 計算機環境*
  - 計算ノード32台 (計64CPU)
    - CPU: 3.60GHz × 2
    - メモリ: 2GB
    - 内蔵ディスク: 600GB (計20TB)
  - ファイルサーバー2台 (計12.5TB)

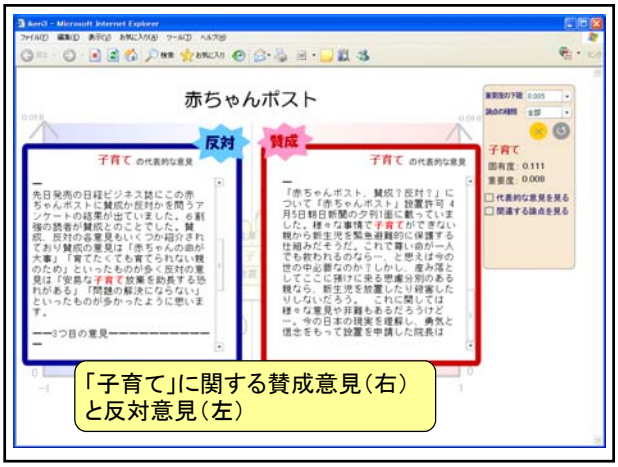
* 今年度末に、64CPU、100TBストレージ、メモリ4GBを追加・増強予定

ディープNLPサーチエンジン基盤 TSUBAKI

info-plosion 情報爆発







Webアーカイブからの辞書、知識ベース自動作成

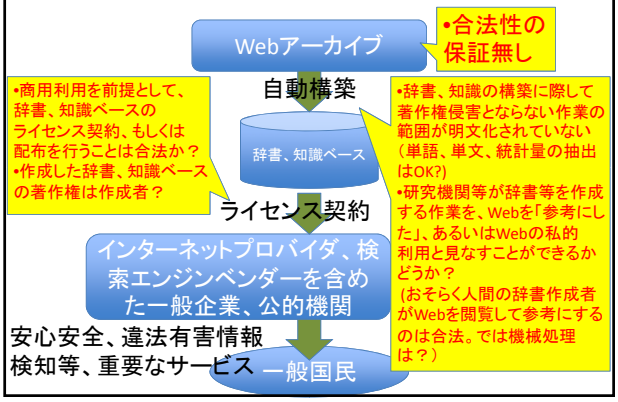
社会における辞書、知識ベース

- 機械学習等技術の進歩により、Webアーカイブをもとに大規模で有用な辞書や知識ベースを自動構築することは現在十分可能
- 辞書/知識ベースとその社会的意義の例
  - 違法有害情報検知で有用な辞書
    - アダルトビデオ女優の名前のリスト
    - 猥褻な日本語表現、犯罪に関係の深い日本語表現
  - 安心安全の担保で有用な知識ベース
    - 食品と有害物質の関係(ぎょうざと残留農薬、アジサイとシアン化合物による中毒)
    - 製品とその欠陥(DHAサプリメントと副作用の出血、ガスコンロとその欠陥)等々
- これらの辞書、知識ベースを、Web上で広く利用可能とすることで、安心安全の確保、違法有害情報フリーなWebなどを国民が享受することができる
- これら、Webから自動的に作られた辞書、知識ベースは現在合法的か？

辞書作成技術の例

- 「中毒」がトラブルであることを認識するには。。
  - Webアーカイブ上で「中毒」の周辺に現れる単語を収集
    - … 中毒で死亡する…
    - …中毒で入院する…
    - …中毒で体調を崩す…
  - そのパターンを統計的手法により自動的に学習
    - 学習された規則:「Xで死亡する」、「Xで入院する」等のパターンに現れる単語Xはトラブルの可能性が高い
    - この規則により、
      - 癌、肝臓がん、毒、ガス中毒、交通事故、人身事故、。。。等の単語もトラブルであることが認識できる
  - おそらく、この種の表現は数万語単位で存在し、人手での収集は現実的ではない
  - また、新規に問題になるトラブルも継続的に収集する必要がある⇒自動化は必須

現状の法的問題点



今後のニーズ、技術の進化を踏まえた法整備の必要性

- 少なくとも、現状、以下のようなニーズ、技術の可能性がある
  - 辞書、知識ベース作成によるイノベーション支援
    - 例:iSP細胞開発において有効利用可能な物質、遺伝子をWeb上に公開されたデータから自動的に列挙して知識ベースとし、体系的に実験
      - 現在はおそらく研究者がWebページを人手で閲覧して決定⇒膨大な手間
      - あらかじめ知識ベースをWebから作成しておくことで研究は加速
  - 機械翻訳の品質向上のためのWebデータ利用
    - Webをもとに膨大な対訳文を抽出、あるいは合成し、機械翻訳システムの統計的パラメータの計算で利用
    - 対訳データ抽出や合成が著作権侵害とならないか？
    - こうした利用が可能であるならば、日本文化、知的財産の発信において有用
  - 重要、かつ大量の情報を素人に分かりやすく説明するため、Web情報を一部自動改変、あるいは要約
    - おそらく現行法では非常にグレー、もしくは非合法
    - 問題は一種のデジタルデバイドであり、社会の効率、安全性を高める上で重要
- 他にも様々な技術のニーズとシーズが存在