文部科学省　次世代ＩＴ基盤構築のための研究開発「Web社会分析基盤ソフトウェアの研究開発」

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

国立情報学研究所 佐藤真一
東京大学　豊田正史、喜連川優
早稲田大学　山名早人

目次

多メディアWeb解析基盤の構築
及び社会分析ソフトウェアの開発

- 全体概要
- 達成状況
- 研究開発成果
- 独創性・優位性
- 中間評価指摘事項への対応
- 研究開発体制
- 成果の利活用
- 今後の展望

全体概要

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

目的

社会学、言語学、リスク管理、マーケティング等多様な社会分析ニーズに応じるために、膨大な多メディアWeb 情報を収集、蓄積し多様な解析を可能とする多メディアWeb情報解析基盤の構築と社会分析ソフトウェアの研究開発並びに実証を行う

背景

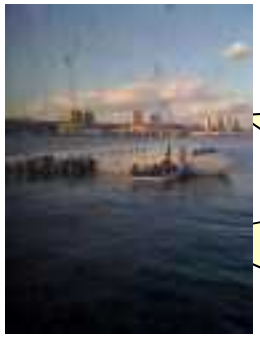
Web情報は人類社会の観測・調査・解析において新価値創出のために必要不可欠な情報源

- ■ 多メディア化が急速に進むと同時に、実世界情報と相互に及ぼし合う影響も拡大
- ■ 放送映像との密接な相互作用

課題

多メディアWeb情報の**収集・蓄積**、多メディア**内容解析**、高並列計算環境上での大容量・高スループット**解析基盤**、有効な**社会解析ソフトウェア**の実現が必須

イラン抗議デモ・チュニジア政変における
twitter/facebook/YouTubeの役割



「ハドソン川の奇跡」は
twitterの投稿写真で話題に

単一メディアではない
複数メディアの
有機的な統合による解析は
世界初

達成状況

- 21-22年度で、多メディアWeb基盤技術ならびに多メディアWeb要素技術において、必要となるツールの基本設計とプロトタイピングが終了
- 23-24年度に、実証アプリケーションのプロトタイプによる検証、具体的な社会分析ターゲットの選定に引き続き、大規模な実証実験を実施

研究開発成果

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

（１）多メディアWeb解析要素技術に関する研究

目標：先進的な技術、社会分析に耐えうる高精度、Webスケールに耐えるスケーラビリティ

●画像・映像キーワード抽出
画像・物体への自動キーワード付与，顔照合
TRECVID2010世界第1位，オープンソース公開

●画像・映像リンケージ情報検出
画像・映像コピー検出，物体検出　**TRECVID2011世界第1位**

●多次元解析高速化
Webデータ圧縮，高次元データ類似検索
世界最高の圧縮率・世界最高速Webクローラ

●多メディアWeb分析・可視化
トピック抽出，メディア間情報伝搬可視化
統計的日本語係り受け解析器オープンソース公開
世界初の画像・テキスト時系列頒価３次元可視化・分析

（２）多メディアWeb基盤技術に関する研究

目標：類を見ない巨大なアーカイブ、スケーラブルな処理を支える処理基盤実現

●多メディアWeb収集・蓄積
テキスト，画像，動画の大規模時系列収集
300億コンテンツ規模のアーカイブ構築（アジア圏最大級）

●データインテンシブスケジューリング
高可用性，レイテンシ最小化のためのスケジューリング手法
マイクロ秒レベルでのレイテンシ制御実現

（３）多メディアWeb統合処理に関する研究

目標：解析要素技術と基盤技術を統合利用するプラットフォームの実現

●多メディアWeb統合処理
共有プラットフォーム構築
150万エントリ大規模意味カテゴリ辞書構築
リアルタイム分散解析ミドルウェアQueueLinker公開
画像・映像キーワード抽出，リンケージ技術のWebAPI構築

（４）多メディアWeb解析の実証評価に関する研究

目標：様々な分野で利用できる社会分析ソフトウェアの実現

●多メディアWeb解析実証評価
- 時系列話題解析エンジンの構築
- **放送映像，ウェブ，ブログを対象とした３次元可視化解析**システムの構築
- **1年半のニュース映像（6000時間以上）と対応ブログ画像（46,000画像）との照合**の実現
- **多メディア間話題画像伝搬解析**の実現
- マイクロブログからのTV視聴者メッセージ自動抽出の実現
- **TVとマイクロブログの連動解析**の実現

研究開発成果

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

（１）多メディアWeb解析要素技術に関する研究

目標：先進的な技術、社会分析に耐えうる高精度、Webスケールに耐えるスケーラビリティ

●画像・映像キーワード抽出
画像・物体への自動キーワード付与，顔照合
TRECVID2010世界第1位，オープンソース公開

●画像・映像リンケージ情報検出
画像・映像コピー検出，物体検出　**TRECVID2011世界第1位**

●多次元解析高速化
Webデータ圧縮，高次元データ類似検索
世界最高の圧縮率・世界最高速Webクローラ

●多メディアWeb分析・可視化
トピック抽出，メディア間情報伝搬可視化
統計的日本語係り受け解析器オープンソース公開
世界初の画像・テキスト時系列頒価３次元可視化・分析

（２）多メディアWeb基盤技術に関する研究

目標：類を見ない巨大なアーカイブ、スケーラブルな処理を支える処理基盤実現

●多メディアWeb収集・蓄積
テキスト，画像，動画の大規模時系列収集
300億コンテンツ規模のアーカイブ構築（アジア圏最大級）

●データインテンシブスケジューリング
高可用性，レイテンシ最小化のためのスケジューリング手法
マイクロ秒レベルでのレイテンシ制御実現

（３）多メディアWeb統合処理に関する研究

目標：解析要素技術と基盤技術を統合利用するプラットフォームの実現

●多メディアWeb統合処理
共有プラットフォーム構築
150万エントリ大規模意味カテゴリ辞書構築
リアルタイム分散解析ミドルウェアQueueLinker公開
画像・映像キーワード抽出，リンケージ技術のWebAPI構築

（４）多メディアWeb解析の実証評価に関する研究

目標：様々な分野で利用できる社会分析ソフトウェアの実現

●多メディアWeb解析実証評価
- 時系列話題解析エンジンの構築
- **放送映像，ウェブ，ブログを対象とした３次元可視化解析**システムの構築
- **1年半のニュース映像（6000時間以上）と対応ブログ画像（46,000画像）との照合**の実現
- **多メディア間話題画像伝搬解析**の実現
- マイクロブログからのTV視聴者メッセージ自動抽出の実現
- **TVとマイクロブログの連動解析**の実現

画像・映像キーワード抽出技術（画像・映像意味分類技術）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 与えられたショットに対し、写っている物体種別、情景種別、画像・映像の種別などに基づいて、自動的に概念レベルの意味分類を行う技術（車、建物、スポーツなど）
- 正解データつきの学習用映像（数百時間規模）で意味分類器を学習
- インターネット映像でも高性能となることを確認済み


爆発・炎、車

歩く・走る、建物、車

爆発・炎、建物、山

歩く・走る、山、スポーツ

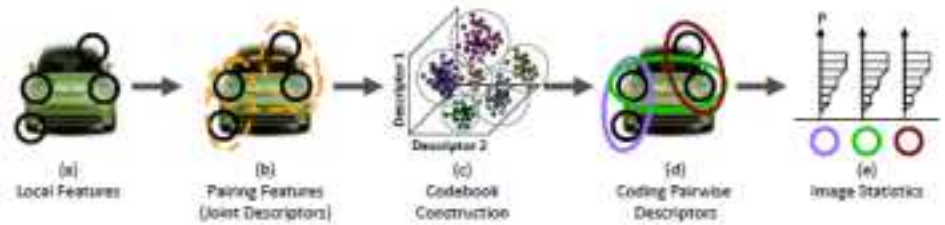

スポーツ、歩く・走る


水辺、山、車

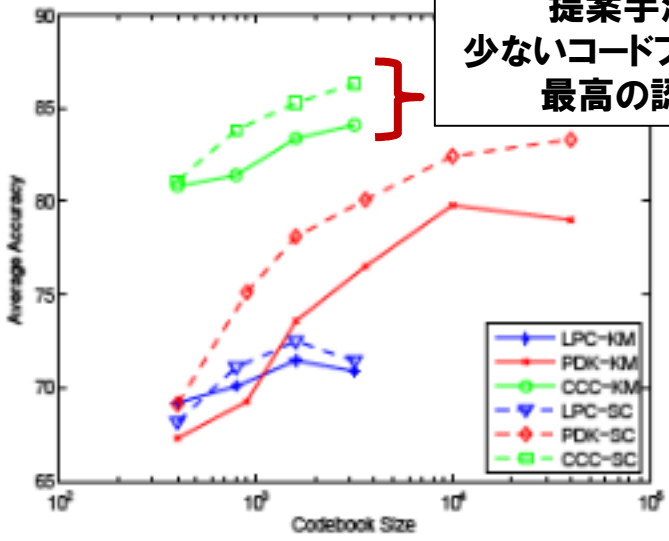

飛行機


顔

キーワード付与結果例



局所特徴量のペアを使うことにより飛躍的に認識精度を向上



提案手法により少ないコードブックサイズで最高の認識精度

コンピュータビジョントップ国際会議ECCV, ICCV
機械学習トップ国際会議NIPS

画像・映像キーワード抽出技術（顔検出・追跡・照合）

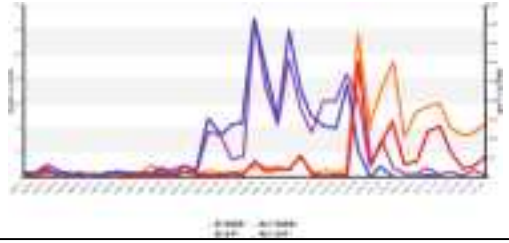
多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 顔特徴空間中の密度の解析とバギングを用いたリランキングにより、Webや放送映像から、特定人物認識器を自動構築可能
- 局所特徴に基づく高精度の顔照合を実現
- 未知の顔に対する名前推定を実現
- TRECVIDインスタンスサーチタスクにおいて顔照合性能の高さを実証

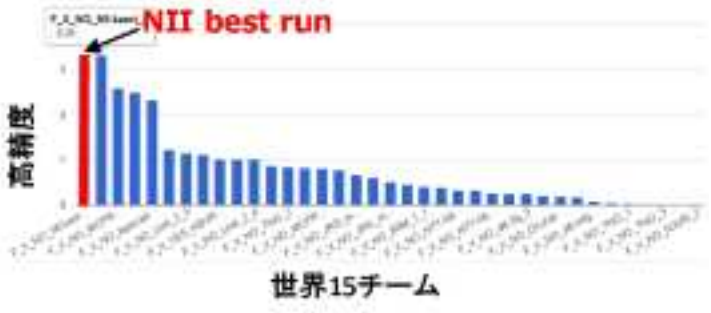
クエリ

検索結果

TRECVID2010 INSにおいて
左記クエリのみを手掛かりに
180時間映像から上記ショットを
正しく検索



Web上顔画像とNHKニュース10年分（6,000万顔画像）を照合し、顔の同定ならびに放送映像中の出現時間・言及数を解析

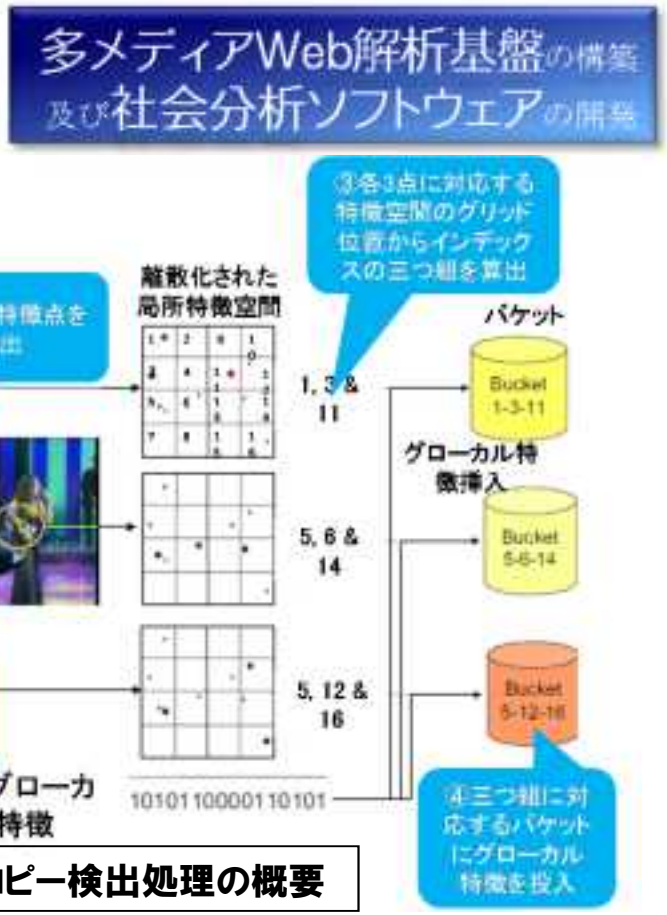


トップ国際会議CVPR, ICDM
TRECVID2010世界第一位

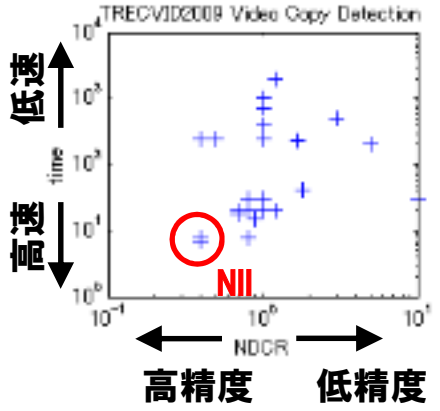
画像・映像リンケージ技術（映像コピー検出）

- 局所特徴量に基づく手法であり、部分的隠れや映像編集などに頑健
- 大局特徴と局所特徴の利点を併せ持ったグローカル特徴
- 空間的配置を考慮したハッシュによる高速照合
- きわめて高速なコピー検出が可能

Web画像と放送映像とのリンケージ例



画像・映像コピー検出処理の概要

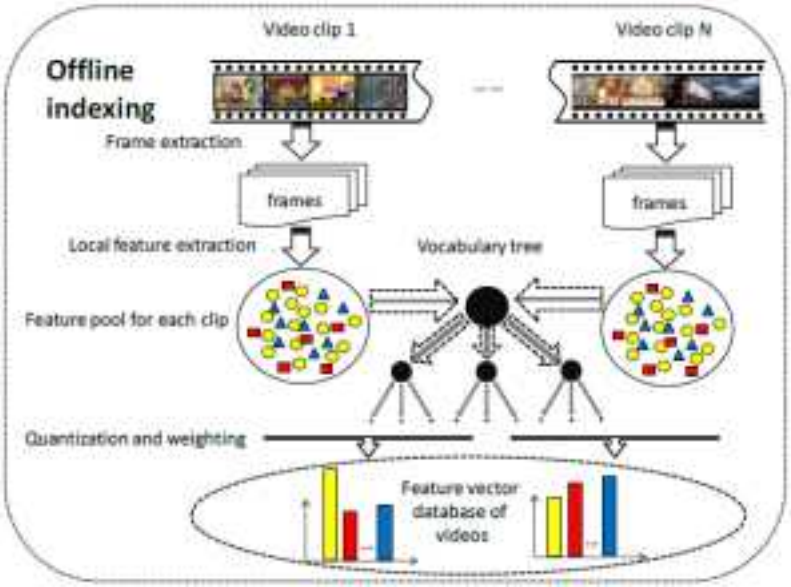


TRECVID2009映像コピー検出タスク
高精度・世界最高速

画像・映像リンケージ技術（同一物体検索）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 画像間で共起する物体を検出する技術
- 物体による画像・映像検索
- Webと放送映像とで共起する物体の検出



問い合わせ画像

任意の画像でコマーシャル映像数万本を検索

検索結果

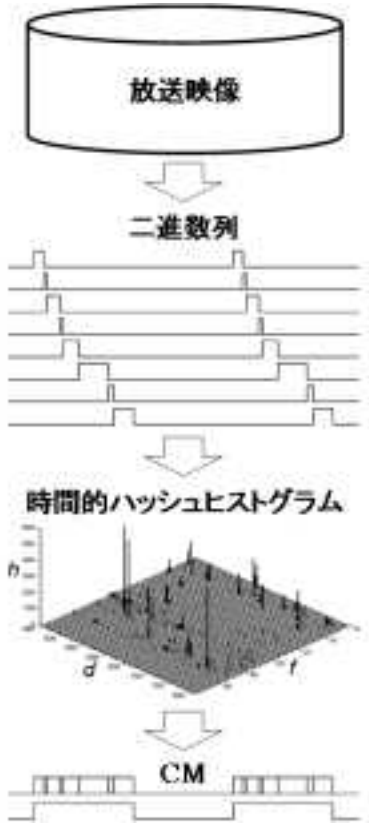
TRECVID2011インスタンスサーチタスク
検索精度**世界第一位**

画像・映像リンケージ技術（コマーシャル映像）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 映像アーカイブ中に繰り返し現れる15-30秒の映像はCMであるとして検出ならびに同定
- 高精度
  - 検出適合率96.6%
  - 検出再現率99.5%
  - 位置特定精度97.4%
- 超高速：1か月間分の映像の処理時間は60分以下
- 映像アーカイブ中のCMの出現に基づく各種統計量を算出可能

コマーシャルに基づくマーケティング戦略の解析



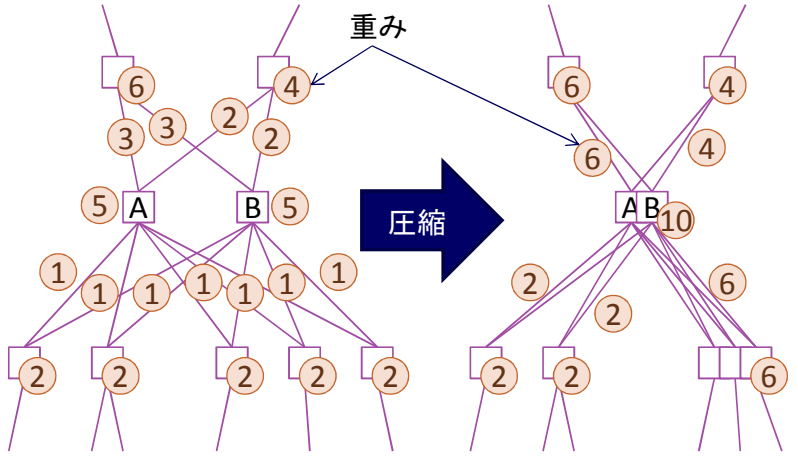
超高速コマーシャル検出・同定処理の概要

高精度・世界最高速
国際ジャーナル採択（IEEE Trans. CSVT）

多次元解析高速化技術
(Webリンク圧縮技術－LittleWeb)

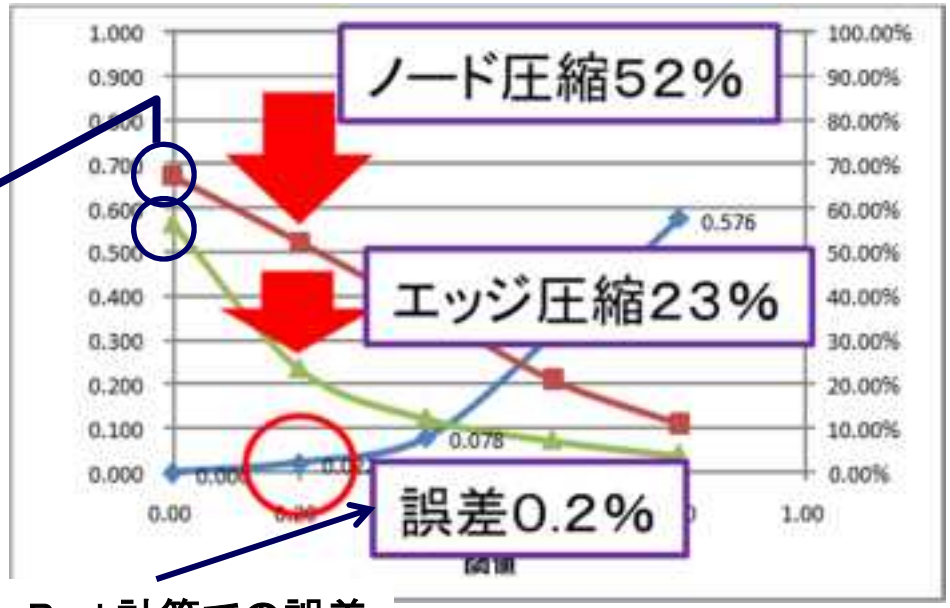
多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 多次元解析で用いるデータを圧縮する技術
  - エッジとノードから構成される有効グラフを対象とし，圧縮状態での構造解析を実現
- 高圧縮
  - ノード圧縮率52%，エッジ圧縮率23%を達成
  - 既存圧縮手法との併用可能
- 実証実験
  - 対象：3946万ノード，9.4億エッジのWebグラフ
  - PageRank算出において解析時間67%削減
  - 計算時間3.8時間（従来比3倍の高速化）



・ノードへの入力・出力が「同一」or「類似」のものをまとめ上げ圧縮
・ノード及びエッジの重みは再計算

- 圧縮率は利用者が設定可能
  - 可逆圧縮時もノード圧縮68%，エッジ圧縮57%の性能



PageRank計算での誤差

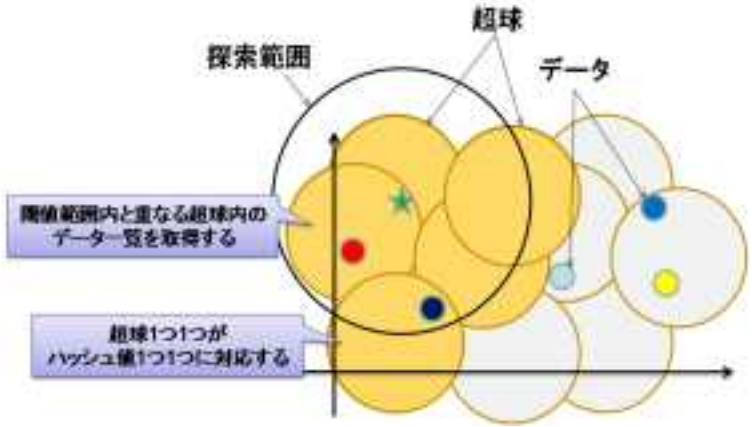
世界最高レベルの圧縮率
DEIM発表

多次元解析高速化技術
（高次元Data高速類似度検索技術）

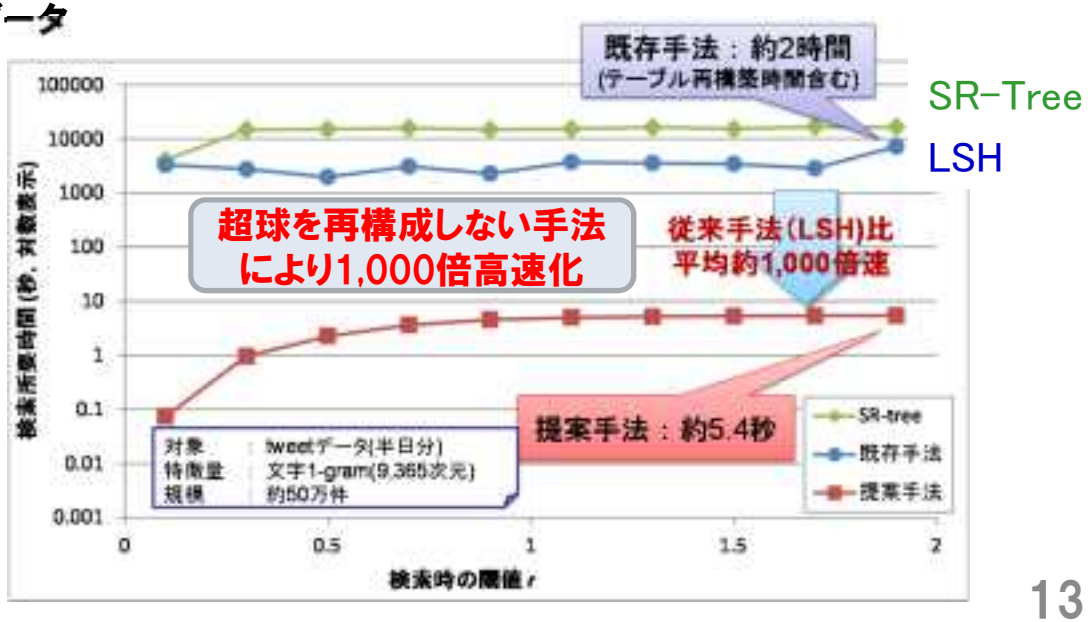
多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 多次元解析における類似度検索を高速化する技術－ResizableLSH

- アイデア
  - LSHの欠点克服（閾値変更不可，ハッシュ再計算要）
  - ハッシュ値間の類似度と多次元空間での類似度を同一視できる空間分割の実現
- 実証実験
  - Tweet 50万件，9365次元の特徴データ
  - 既存手法（SR-tree, LSH）比、1000倍の高速化
  - リアルタイムでの高次元データ解析を可能に



Resizable-LSHでのデータ、超球、探索範囲の関係



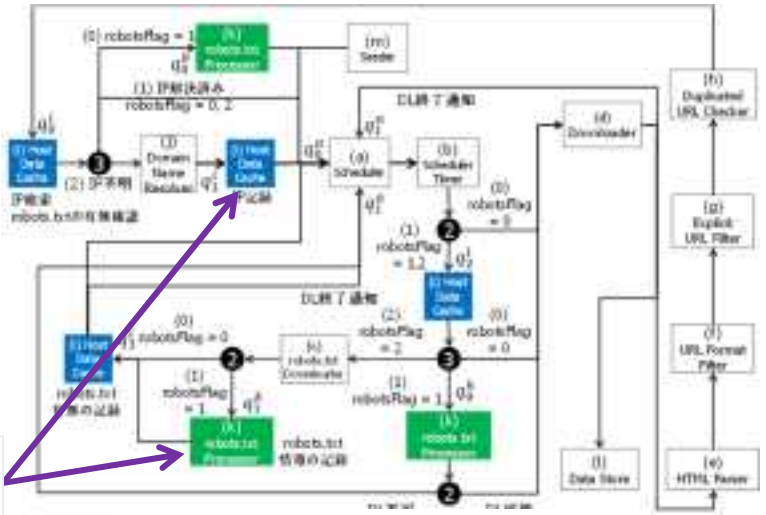
高次元データのリアルタイム類似度判定を可能に

IPSJ研究会（AL,DBS） 発表

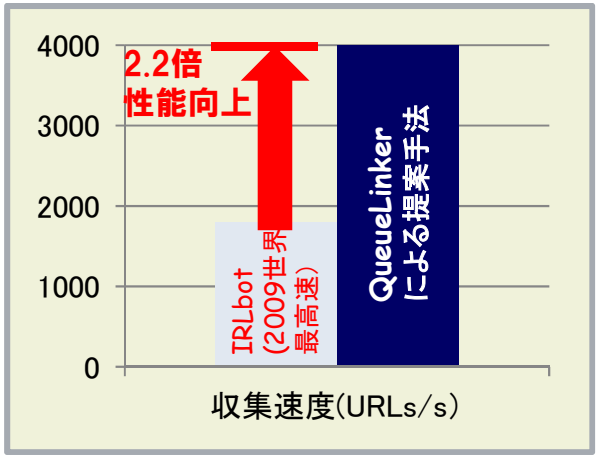
多次元解析高速化技術
（最高速・並列分散型Webクローラ）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

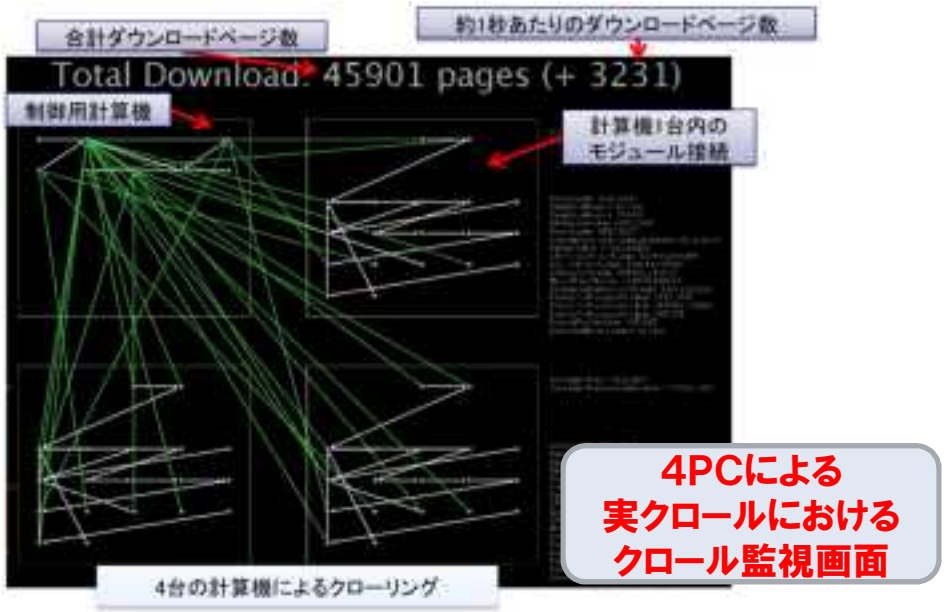
- 多メディアWebデータの高速収集を実現
  - 多メディア統合処理で構築した「**リアルタイム分散解析ミドルウェアQueueLinker**」を用いて構築
- 実証実験
  - 13モジュールを全てデータ並列化し実装
  - 実クロールにおいて、**毎秒4,000URL収集の世界最高速を達成**



並列分散型Webクローラ構成図



世界最高速（4,000URL/秒）
TOD論文誌掲載



4PCによる実クロールにおけるクロール監視画面

多メディアWeb分析・可視化技術（高速固有表現抽出アルゴリズム）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 大規模ウェブデータを社会分析に利用可能とする
  人物名や製品名などの固有表現の超高速抽出技術

人物名

健康上の理由により療養休暇中のアップルのCEO，
スティーブ・ジョブズ氏は次期iPadやiPhoneの開発に
関与していると明かしました

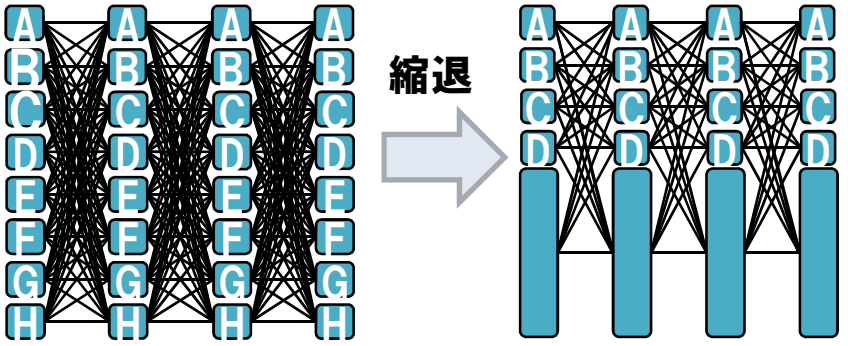
製品名

会社名

- メモリ効率を保持しながら従来より最大**300倍の高速化**を実現
  - 縮退ラティスによる探索空間の削減法を提案

解析速度（文数/秒）

アルゴリズム	結合タグ	Supertag
Viterbi	77	1.1
CarpeDiem	51	0.26
提案手法	1600	300



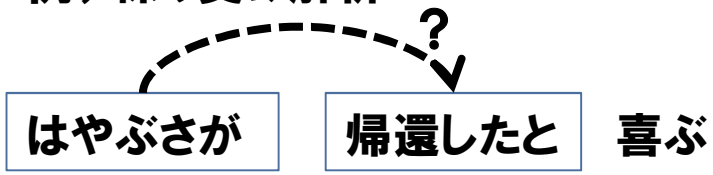
自然言語処理における最高峰の国際会議ACL2010に採択
Kaji et al. Efficient Staggered Decoding for Sequence Labeling

多メディアWeb分析・可視化技術
（組合せ素性に基づく分類器の学習）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

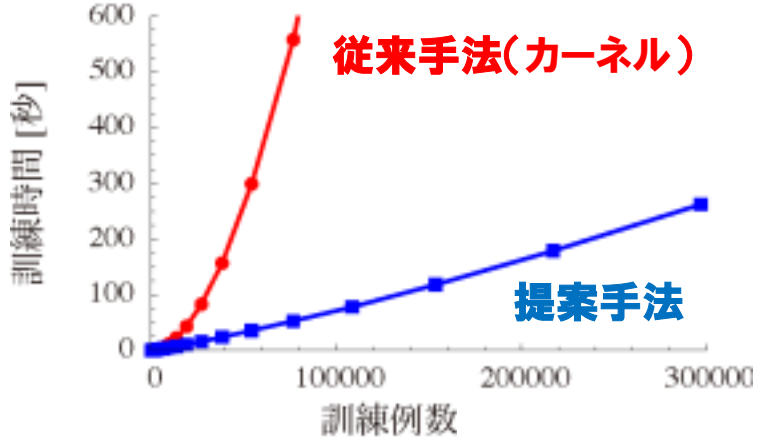
- 係り受け解析の頑健化には**多数の素性の組合せ**を考慮し，**大量の訓練例**を用いた学習が必要

例）係り受け解析



基本素性：品詞（細分類），活用，距離
組合せ素性：品詞×活用，品詞×品詞細分類，
品詞細分類×活用，品詞×距離・・・

- 組合せ素性を用いた大規模学習のための逐次学習法を提案
  - 低頻度素性に関する組合せを多項式カーネルで効率的に計算
    学習の時空間効率を制御
  - 素性の組合せを頻度を考慮して再分割し部分計算結果を再利用
    学習の規模耐性を向上



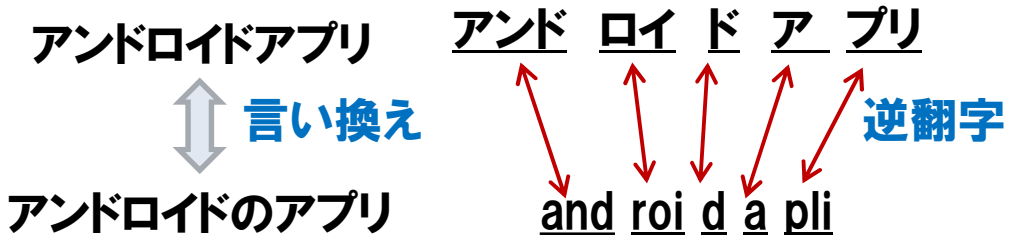
空間効率を保ちつつ学習を最大250倍高速化

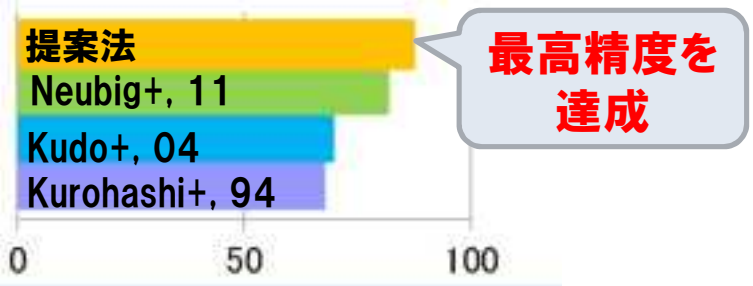
トップ会議 COLING 2010 採択　高速学習器 opal をオープンソースで公開

多メディアWeb分析・可視化技術（先進的な言語解析技術）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

■ 言い換えと逆翻字に基づく高精度複合名詞分割

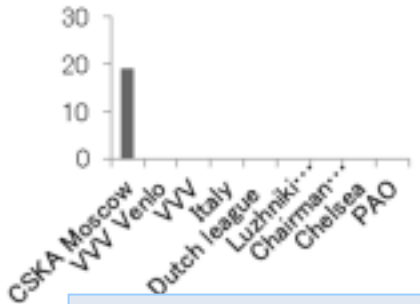


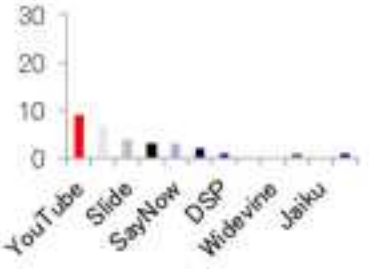


自然言語処理におけるトップ国際会議EMNLP2011に採択
言語処理学会2012年最優秀論文賞を受賞

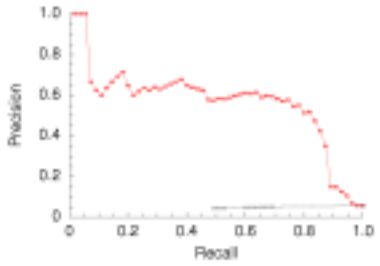
■ 恒久性に基づくイベント（関係）分類技術を世界で初めて提案

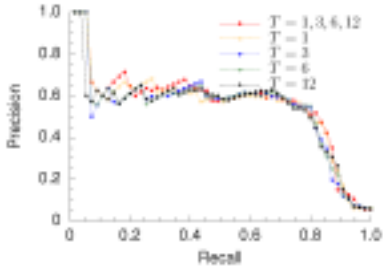
時系列テキストからの特徴量抽出





イベント分類実験の結果



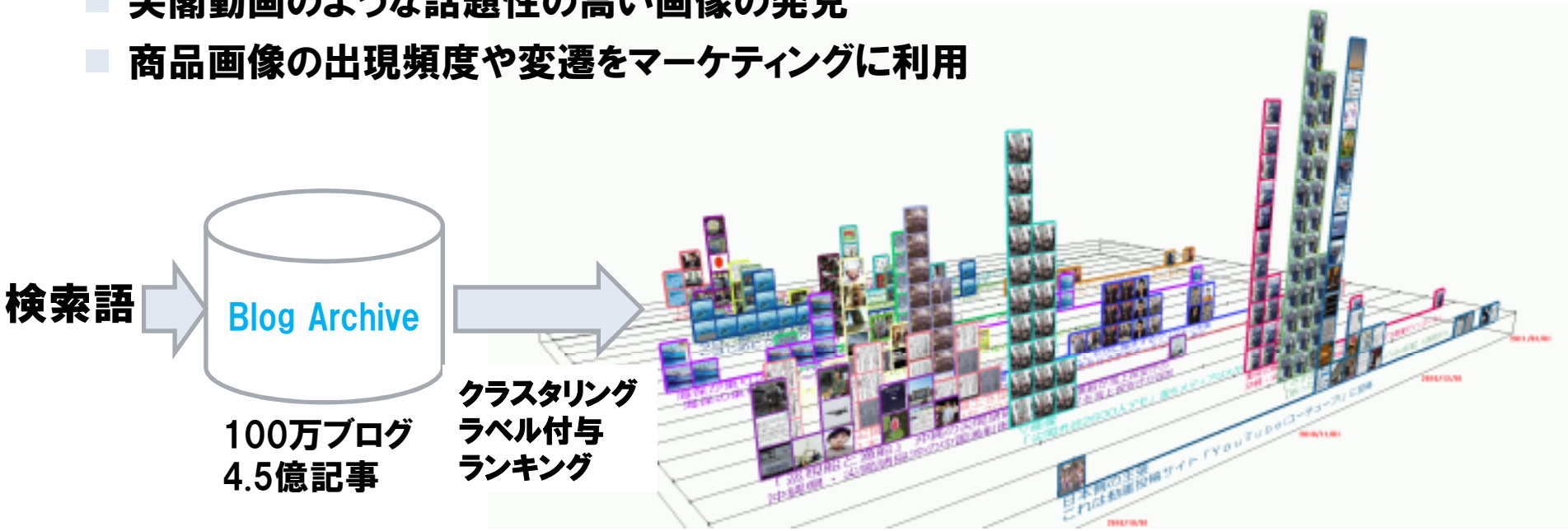


自然言語処理におけるトップ国際会議EMNLP2012に採択

多メディアWeb分析・可視化技術（CGM画像の組織化）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- CGM上の話題の推移を把握可能にするため
  画像を詳細な話題に分類し、ラベル付与・ランキングを行う
  - 報道に使用された画像の調査
  - イベントや集会の様子の視覚的な把握
  - 尖閣動画のような話題性の高い画像の発見
  - 商品画像の出現頻度や変遷をマーケティングに利用



分類画像と放送映像の照合により
放送・Web間の話題伝搬追跡を可能とする

研究開発成果

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

（１）多メディアWeb解析要素技術に関する研究

目標：先進的な技術、社会分析に耐えうる高精度、Webスケールに耐えるスケーラビリティ

●画像・映像キーワード抽出
画像・物体への自動キーワード付与，顔照合
TRECVID2010世界第1位，オープンソース公開

●画像・映像リンケージ情報検出
画像・映像コピー検出，物体検出　**TRECVID2011世界第1位**

●多次元解析高速化
Webデータ圧縮，高次元データ類似検索
世界最高の圧縮率・世界最高速Webクローラ

●多メディアWeb分析・可視化
トピック抽出，メディア間情報伝搬可視化
統計的日本語係り受け解析器オープンソース公開
世界初の画像・テキスト時系列頒価３次元可視化・分析

（２）多メディアWeb基盤技術に関する研究

目標：類を見ない巨大なアーカイブ、スケーラブルな処理を支える処理基盤実現

●多メディアWeb収集・蓄積
テキスト，画像，動画の大規模時系列収集
300億コンテンツ規模のアーカイブ構築（アジア圏最大級）

●データインテンシブスケジューリング
高可用性，レイテンシ最小化のためのスケジューリング手法
マイクロ秒レベルでのレイテンシ制御実現

（３）多メディアWeb統合処理に関する研究

目標：解析要素技術と基盤技術を統合利用するプラットフォームの実現

●多メディアWeb統合処理
共有プラットフォーム構築
150万エントリ大規模意味カテゴリ辞書構築
リアルタイム分散解析ミドルウェアQueueLinker公開
画像・映像キーワード抽出，リンケージ技術のWebAPI構築

（４）多メディアWeb解析の実証評価に関する研究

目標：様々な分野で利用できる社会分析ソフトウェアの実現

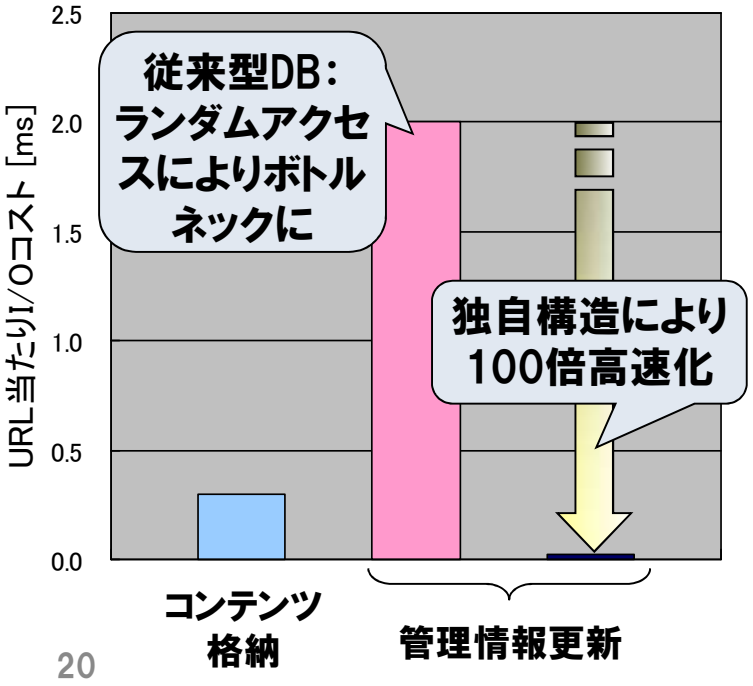
●多メディアWeb解析実証評価
- 時系列話題解析エンジンの構築
- **放送映像，ウェブ，ブログを対象とした３次元可視化解析**システムの構築
- **1年半のニュース映像（6000時間以上）と対応ブログ画像（46,000画像）との照合**の実現
- **多メディア間話題画像伝搬解析**の実現
- マイクロブログからのTV視聴者メッセージ自動抽出の実現
- **TVとマイクロブログの連動解析**の実現

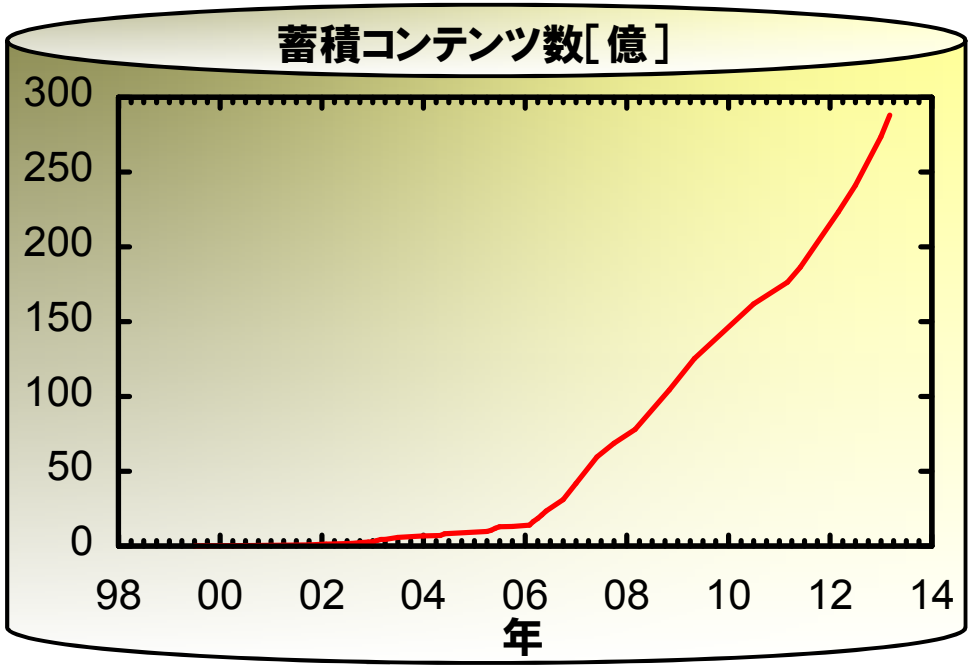
多メディアWeb収集・蓄積技術

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 20億URLの更新頻度に個別に適応する細粒度可変周期収集・蓄積技術（1分～1年周期）
- 14年間にわたり300億件規模（2013年3月）の日本語ウェブページ・画像を集積し、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

可変周期収集を実現するデータ管理技術





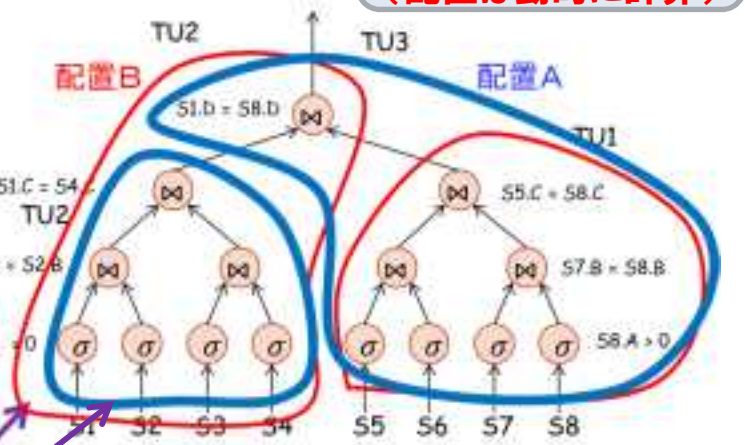
データインテンシブスケジューリング技術（オペレータ再配置）

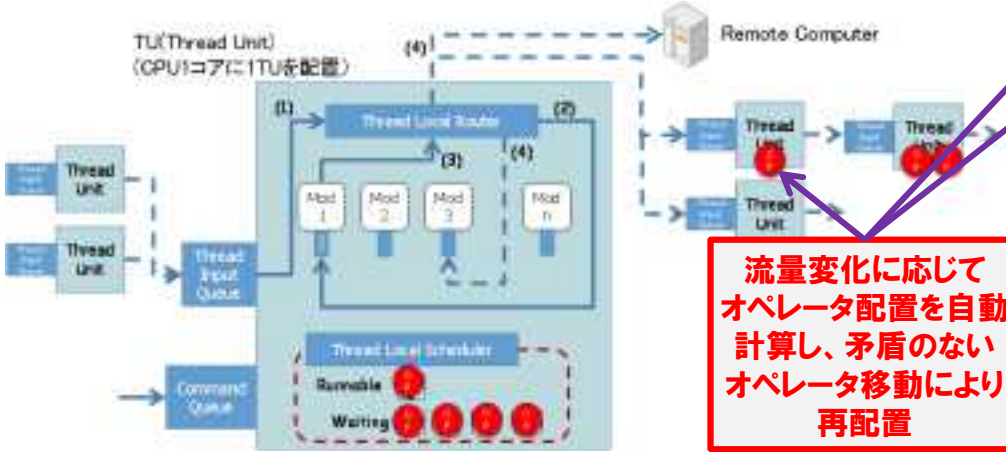
多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- レイテンシ最小化のためのオペレータ再配置法
  - ストリーム処理時の「データ流量変化」に対応したレイテンシ最小化のためのスケジューリング技術
  - 流量変化に応じた動的命令移動
- 実証実験
  - 右図に示す処理において，計算機資源（TU1，TU2，TU3）へのオペレータ配置を動的に制御
  - レイテンシ変動幅をμ秒レベルで制御することに成功し、アルゴリズム取引等への応用も可能に

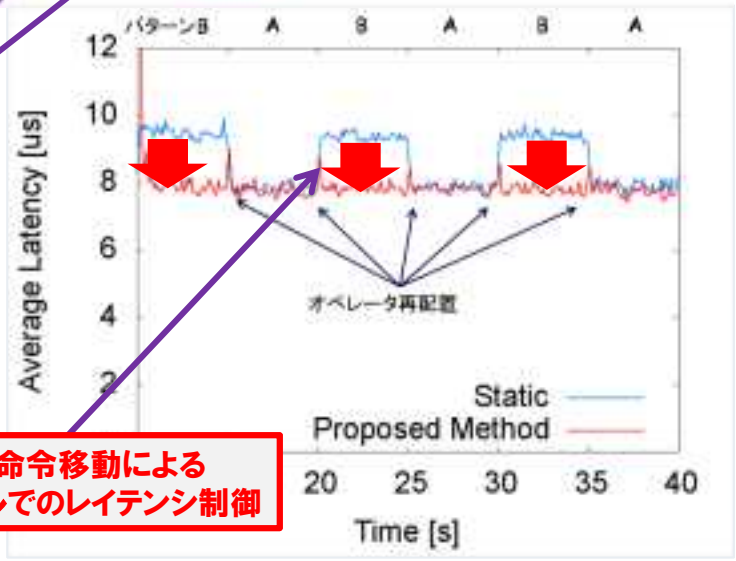
QueueLinkerで利用可

流量変化に応じて
配置A,配置Bを切替
（配置は動的に計算）





流量変化に応じて
オペレータ配置を自動
計算し、矛盾のない
オペレータ移動により
再配置

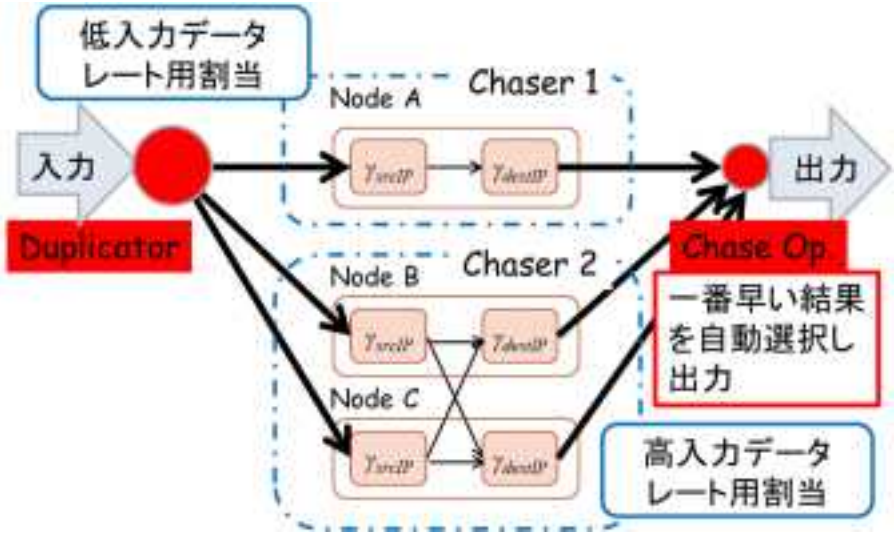


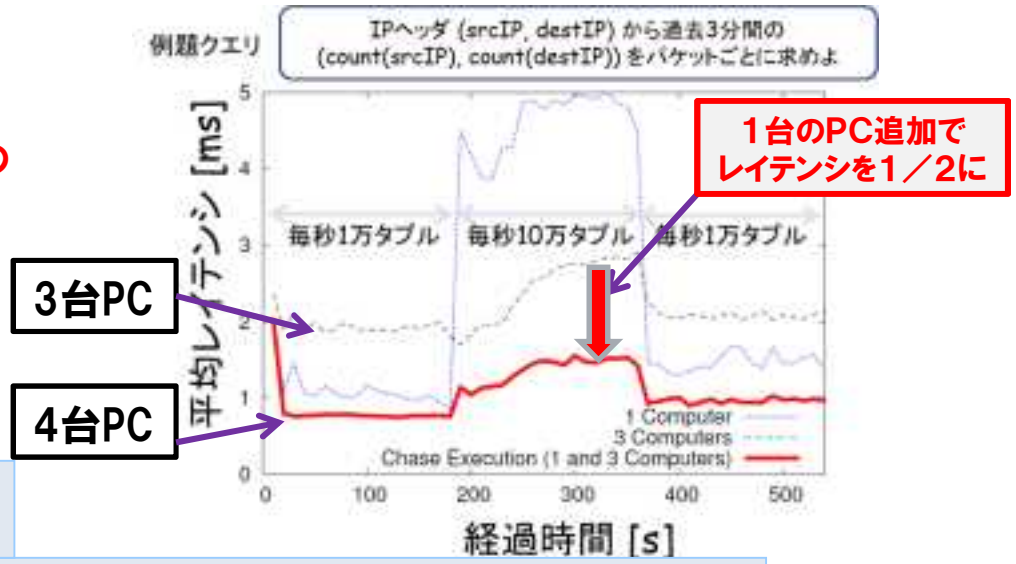
動的命令移動による
Msレベルでのレイテンシ制御

μ秒レベルでの再配置を実現
IEICE和文D論文誌掲載

データインテンシブスケジューリング技術（ChaseExecution）

- 高可用性とレイテンシ最小化を実現する実行手法
  QueueLinkerで利用可
- アイデア
  - 高可用性のために用意した計算機資源を常時利用する（※通常利用時と異なるオペレータ配置）ことにより，流量変化への瞬時対応
- 実証実験
  - IPヘッダ解析を事例として利用
  - 流量変化10倍時もレイテンシ変動を従来の1/2以下に低減できることを確認





高可用性と低レイテンシを実現

WebDB発表，DBSJ論文誌掲載，IPSJ-CS領域奨励賞受賞

研究開発成果

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

（１）多メディアWeb解析要素技術に関する研究

目標：先進的な技術、社会分析に耐えうる高精度、Webスケールに耐えるスケーラビリティ

●画像・映像キーワード抽出
画像･物体への自動キーワード付与，顔照合
TRECVID2010世界第1位，オープンソース公開

●画像・映像リンケージ情報検出
画像･映像コピー検出，物体検出　**TRECVID2011世界第1位**

●多次元解析高速化
Webデータ圧縮，高次元データ類似検索
世界最高の圧縮率・世界最高速Webクローラ

●多メディアWeb分析・可視化
トピック抽出，メディア間情報伝搬可視化
統計的日本語係り受け解析器オープンソース公開
世界初の画像・テキスト時系列頒価３次元可視化・分析

（２）多メディアWeb基盤技術に関する研究

目標：類を見ない巨大なアーカイブ、スケーラブルな処理を支える処理基盤実現

●多メディアWeb収集・蓄積
テキスト，画像，動画の大規模時系列収集
300億コンテンツ規模のアーカイブ構築（アジア圏最大級）

●データインテンシブスケジューリング
高可用性，レイテンシ最小化のためのスケジューリング手法
マイクロ秒レベルでのレイテンシ制御実現

（３）多メディアWeb統合処理に関する研究

目標：解析要素技術と基盤技術を統合利用するプラットフォームの実現

●多メディアWeb統合処理
共有プラットフォーム構築
150万エントリ大規模意味カテゴリ辞書構築
リアルタイム分散解析ミドルウェアQueueLinker公開
画像・映像キーワード抽出，リンケージ技術のWebAPI構築

（４）多メディアWeb解析の実証評価に関する研究

目標：様々な分野で利用できる社会分析ソフトウェアの実現

●多メディアWeb解析実証評価
- 時系列話題解析エンジンの構築
- **放送映像，ウェブ，ブログを対象とした３次元可視化解析**システムの構築
- **1年半のニュース映像（6000時間以上）と対応ブログ画像（46,000画像）との照合**の実現
- **多メディア間話題画像伝搬解析**の実現
- マイクロブログからのTV視聴者メッセージ自動抽出の実現
- **TVとマイクロブログの連動解析**の実現

多メディアWeb統合処理（QueueLinker）

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

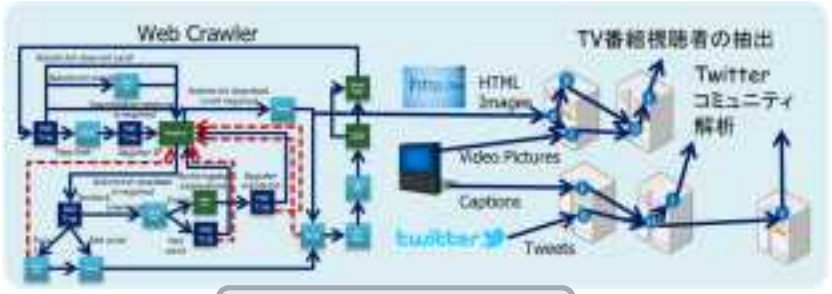
- リアルタイム分散解析の容易な実現を可能とする「リアルタイム分散解析ミドルウエア」

応用例：世界最高速・並列分散Webクローラ

- 特徴
  - モジュール実装と接続関係をProducer-Comsumerモデルにより記述するのみ（**データストリーム処理の容易な実現**）
  - 並列数・分散数を自由にユーザ制御可
  - モジュール内の並列制御はQueueLinker側で制御

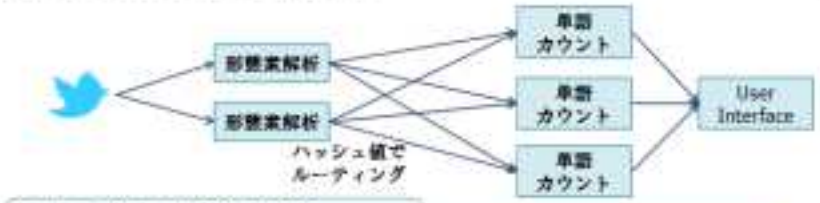
QueueLinker提供の基底クラス

名称	用途	ブロック処理の可否
PushModule	フィルタリング・算術演算を含む、ノンブロッキング処理一般	×
PullModule	ファイル I/O を含む、ブロッキング処理一般	○
SourceModule	外部データの取得	○
SinkModule	処理データの保存・表示	×



モジュール接続の記述例と実行方法

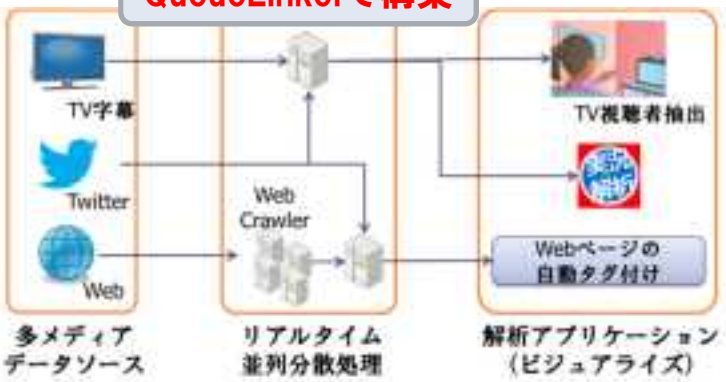
```
LogicalGraph graph          = new LogicalGraph();
LogicalVertex twitter       = graph.addLogicalVertex(TwitterDataSource.class);
LogicalVertex morphAnalyzer = graph.addLogicalVertex(MorphAnalyzer.class, 2);
LogicalVertex wordCount     = graph.addLogicalVertex(WordCount.class, 3, HMode.Hash);
LogicalVertex ui            = graph.addLogicalVertex(UI.class);

graph.addLogicalEdge(twitter,       morphAnalyzer);
graph.addLogicalEdge(morphAnalyzer, wordCount);
graph.addLogicalEdge(wordCount,     ui);

QueueLinkerClient client = QueueLinkerClientFactory.getClient();
QueueLinkerJob job = new QueueLinkerJob(graph);
JobHandle handle = client.startJob(job);
```



モジュールのデータ並列方式を指定
- Full：　指定なし（ラウンドロビン）
- Hash：　ハッシュ分割
- Customize：　ユーザ制御

「キュー」を「接続」⇒ QueueLinker

QueueLinkerで構築



オープンソース公開

研究開発成果

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

（１）多メディアWeb解析要素技術に関する研究

目標：先進的な技術、社会分析に耐えうる高精度、Webスケールに耐えるスケーラビリティ

●画像・映像キーワード抽出
画像･物体への自動キーワード付与，顔照合
TRECVID2010世界第1位，オープンソース公開

●画像・映像リンケージ情報検出
画像･映像コピー検出，物体検出　**TRECVID2011世界第1位**

●多次元解析高速化
Webデータ圧縮，高次元データ類似検索
世界最高の圧縮率・世界最高速Webクローラ

●多メディアWeb分析・可視化
トピック抽出，メディア間情報伝搬可視化
統計的日本語係り受け解析器オープンソース公開
世界初の画像・テキスト時系列頒価３次元可視化・分析

（２）多メディアWeb基盤技術に関する研究

目標：類を見ない巨大なアーカイブ、スケーラブルな処理を支える処理基盤実現

●多メディアWeb収集・蓄積
テキスト，画像，動画の大規模時系列収集
300億コンテンツ規模のアーカイブ構築（アジア圏最大級）

●データインテンシブスケジューリング
高可用性，レイテンシ最小化のためのスケジューリング手法
マイクロ秒レベルでのレイテンシ制御実現

（３）多メディアWeb統合処理に関する研究

目標：解析要素技術と基盤技術を統合利用するプラットフォームの実現

●多メディアWeb統合処理
共有プラットフォーム構築
150万エントリ大規模意味カテゴリ辞書構築
リアルタイム分散解析ミドルウェアQueueLinker公開
画像・映像キーワード抽出，リンケージ技術のWebAPI構築

（４）多メディアWeb解析の実証評価に関する研究

目標：様々な分野で利用できる社会分析ソフトウェアの実現

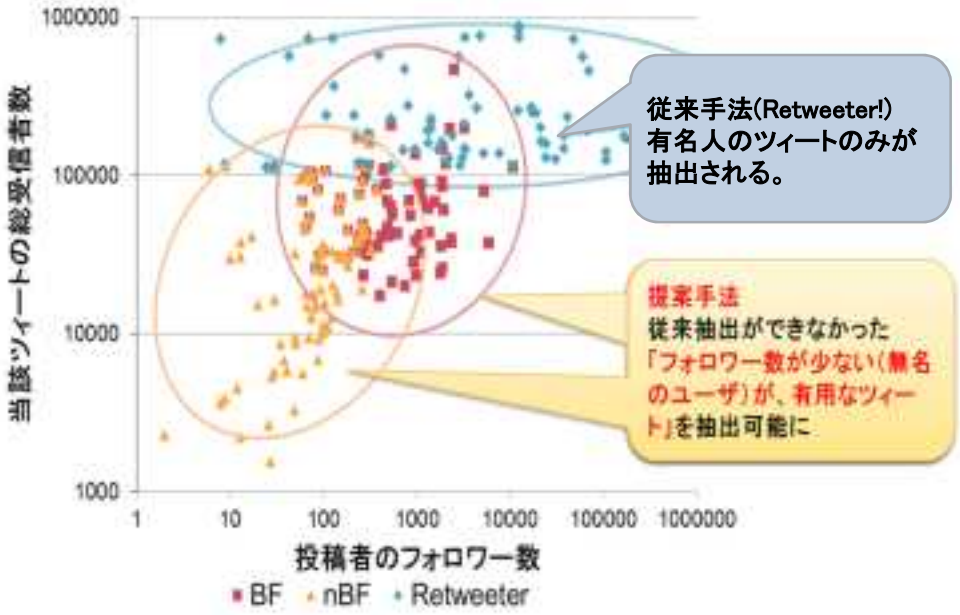
●多メディアWeb解析実証評価
- 時系列話題解析エンジンの構築
- **放送映像，ウェブ，ブログを対象とした３次元可視化解析**システムの構築
- **1年半のニュース映像（6000時間以上）と対応ブログ画像（46,000画像）との照合**の実現
- **多メディア間話題画像伝搬解析**の実現
- マイクロブログからのTV視聴者メッセージ自動抽出の実現
- **TVとマイクロブログの連動解析**の実現

解析事例①

有用tweet抽出－twichaRT

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 多メディアWeb解析において，重要な情報の一つであるtweetから**重要tweet(※)抽出**を実
  - (※)ユーザの有名・無名に関係なく，その内容に多くの人が興味を持つtweet
- アイデア
  - リツィート数だけでなく，フォロワー数で正規化してランキング



ICWSM2011併設WS発表・サービス公開(2012.1)

解析事例②
TV・Twitter解析

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- TVに連動するtweetの自動抽出，及びTV番組の盛り上がり部分の自動検出を実現

TV・Twitter解析

DEIM発表・サービス公開（2013.2）

TV連動tweet自動抽出

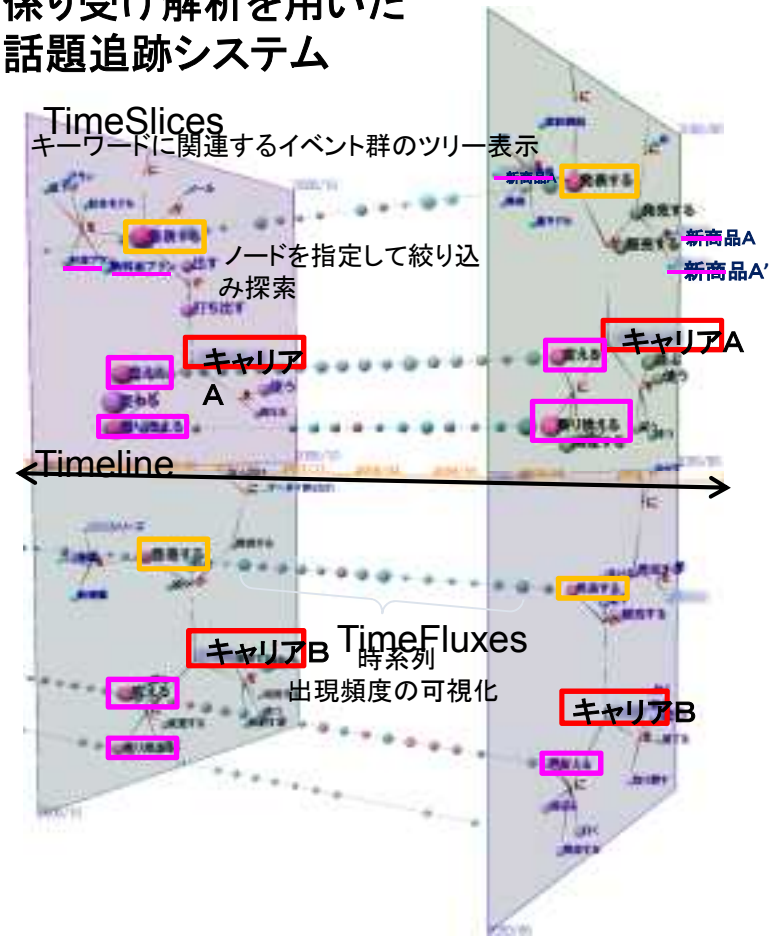
- TV番組視聴ユーザのtweetの特徴語＋字幕＋電子番組表利用

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

解析事例③
多メディア話題追跡システム

- Webグラフ及び係り受け関係の時系列変化を可視化し、インフルエンサー、人々の行動・興味の推移を追跡探索
- メディア間、話題間の比較分析が可能
  - メディアによるインフルエンサー、書き込み内容の差
  - 商品間、人物間の差

係り受け解析を用いた
話題追跡システム



国際会議IV2010採択、第72回情報処理学会全国大会 **大会優秀賞**
国際会議PVis2012採択、DEIM2011**優秀論文賞**

解析事例④

多メディアWeb統合解析

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

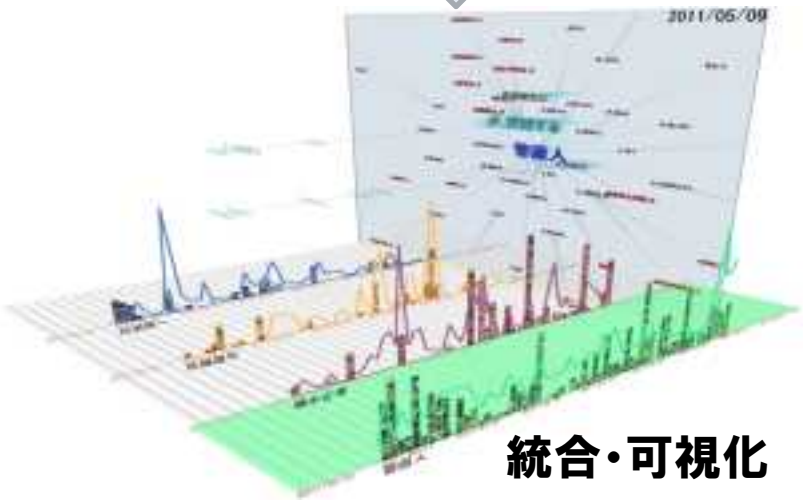
■ Web・放送映像上の話題比較・相互作用分析を実現

Web画像

Web・放送映像リンケージ

映像照合・テキスト比較
時系列の時差・頻度比較

放送映像アーカイブ



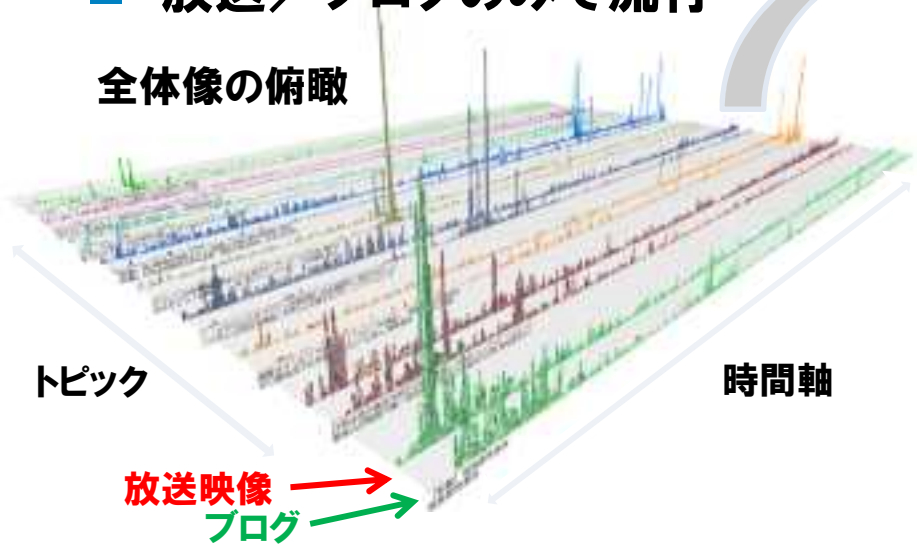
統合・可視化

解析事例④

放送・ブログを用いた相補的なイベント抽出・反応差分析

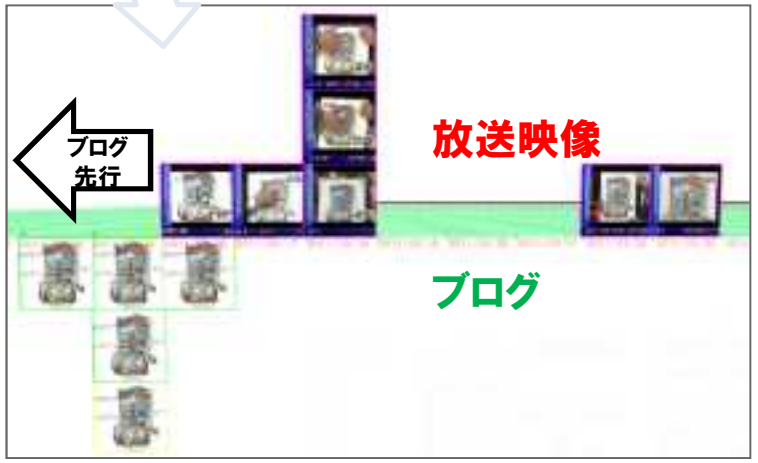
多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- 放送・ブログから類似画像クラスタ群を抽出、3次元空間に時系列可視化
- メディア間の相違、話題間の相違を探索
  - 放送先行で流行
  - ブログ先行で流行
  - 同時に流行
  - 相互に影響
  - 放送／ブログのみで流行



交差相関係数、コサイン類似度等から様々な反応差の画像群を絞り込み

詳細探索

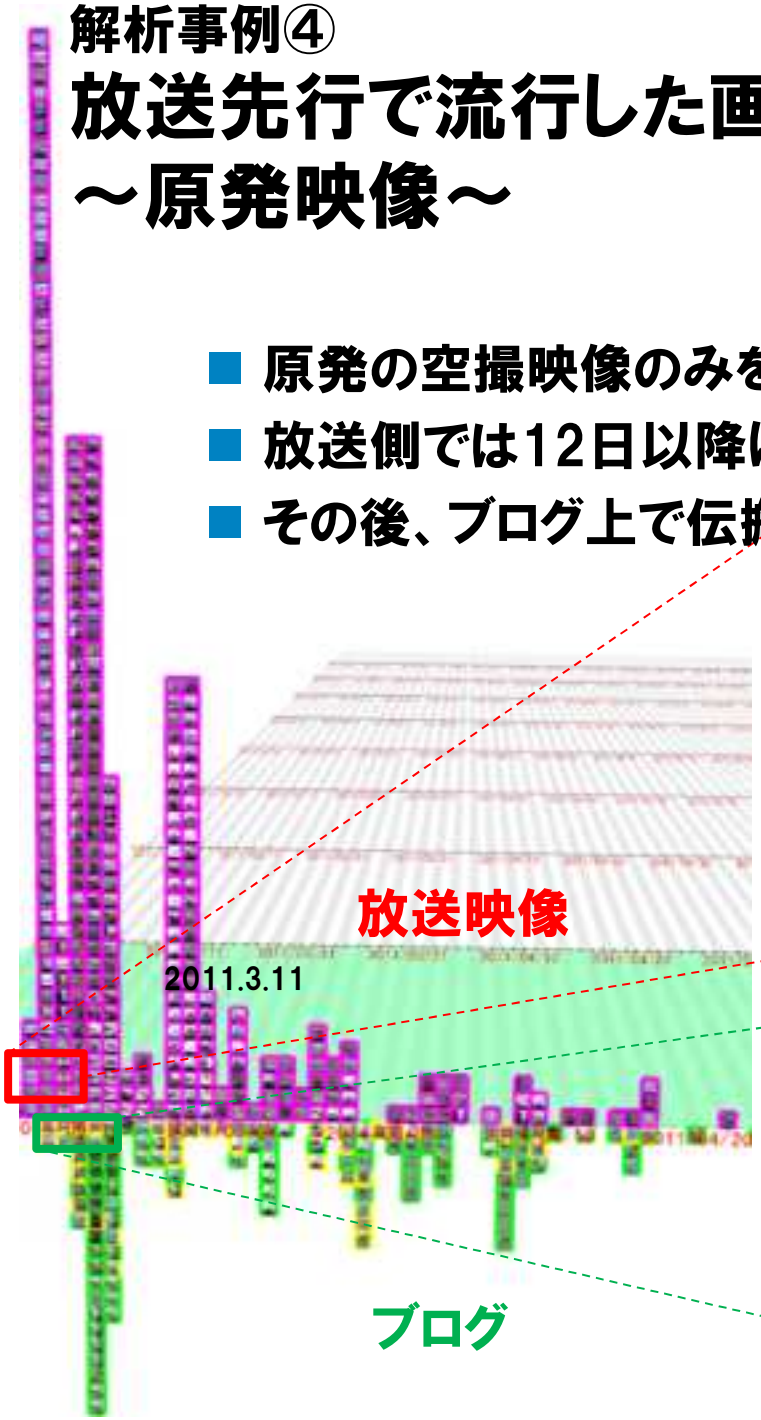


解析事例④

放送先行で流行した画像
～原発映像～

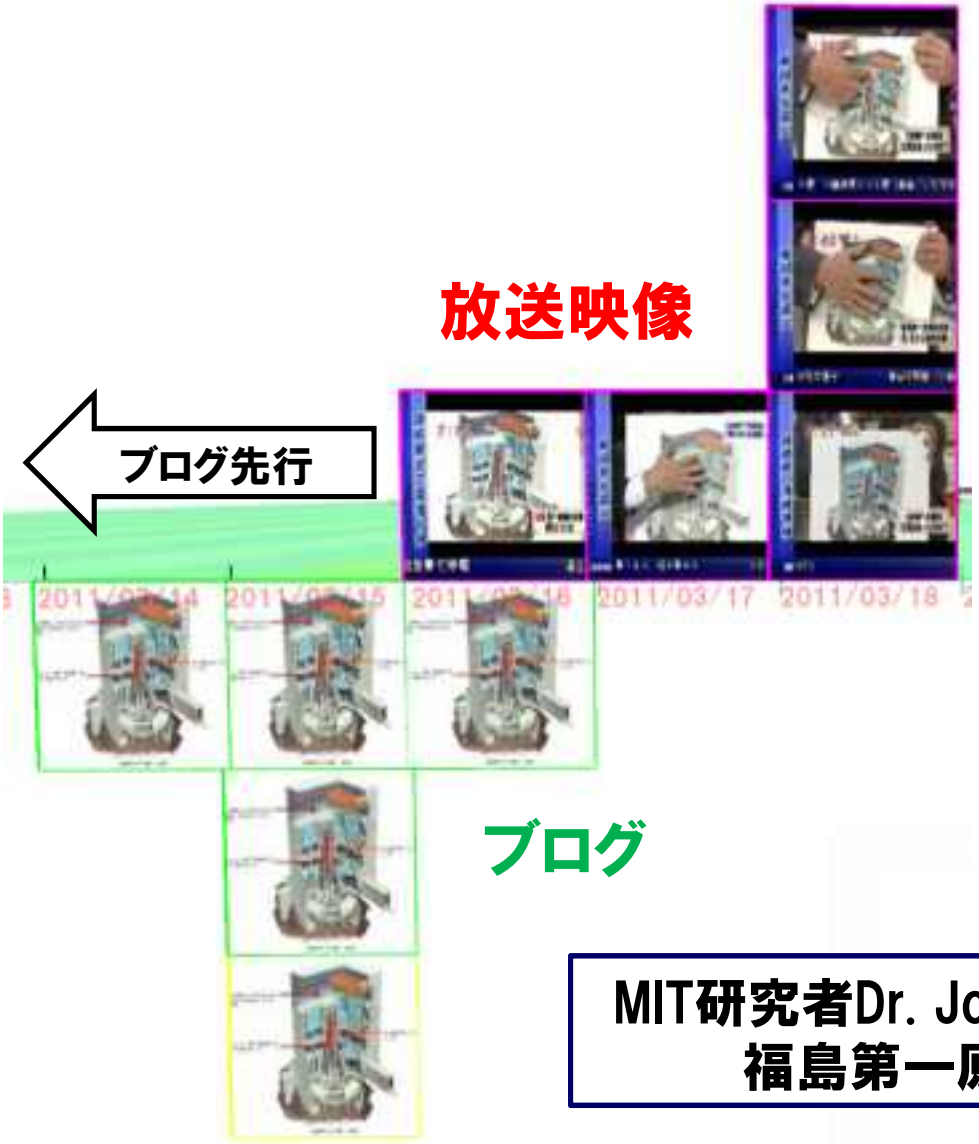
多メディアWeb解析基盤の構築
及び社会分析ソフトウェアの開発

- 原発の空撮映像のみを選択
- 放送側では12日以降に繰り返し出現
- その後、ブログ上で伝搬



解析事例④

ブログ先行で流行した画像
～原発事故解説図～



情報ソースへの
アクセスによる
詳細観測

ニュース動画

ブログエントリ

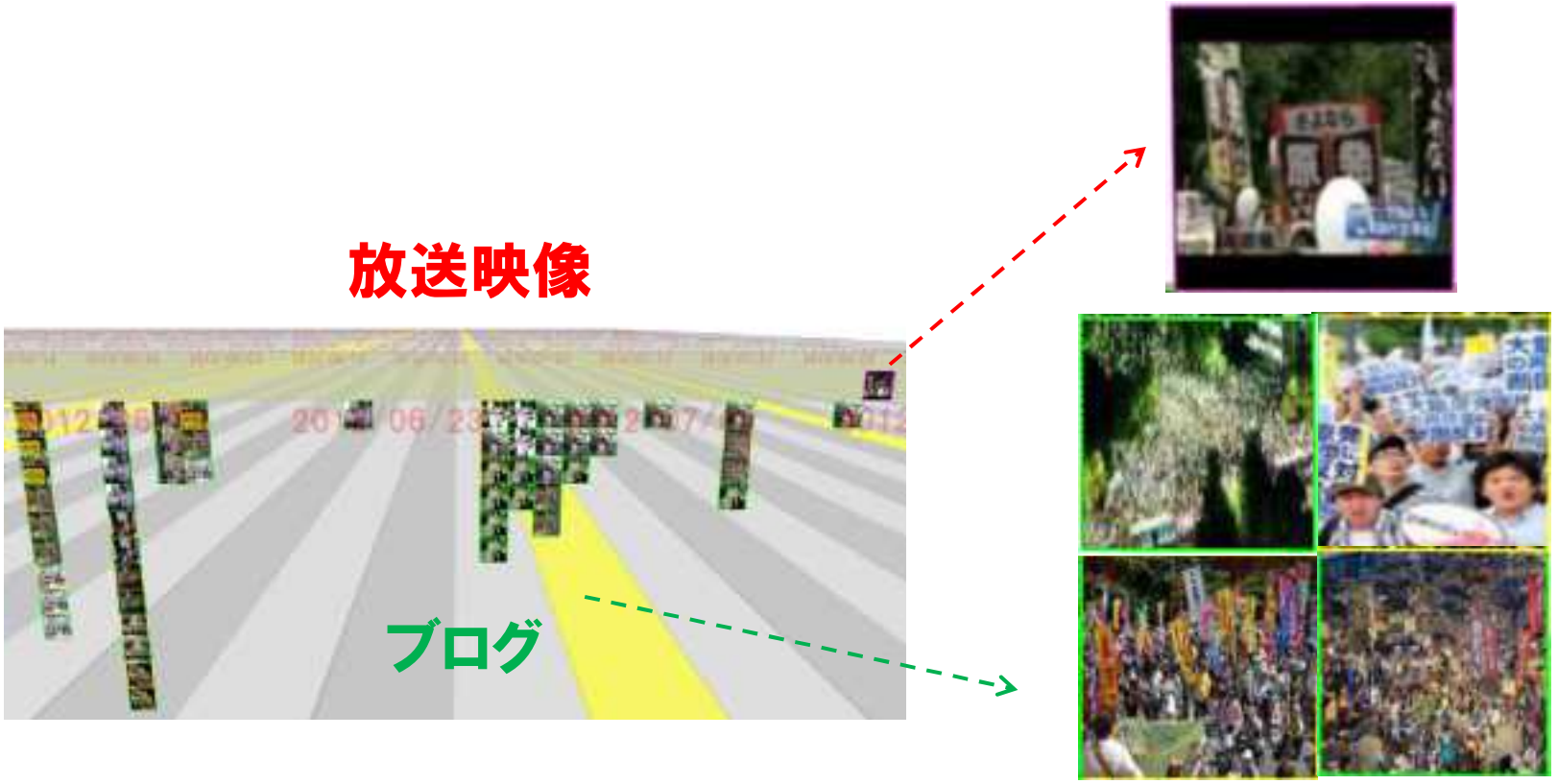
MIT研究者Dr. Josef Oehmenによる
福島第一原発事故解説

解析事例④

ブログのみで流行した画像
～デモ行進の様子～

多メディアWeb解析基盤の構築
及び社会分析ソフトウェアの開発

- デモ活動に対するウェブとテレビの反応の違い
  - テレビ報道では大きく扱われにくい
  - ブログでは賛否を含めて大きく取り上げられる傾向



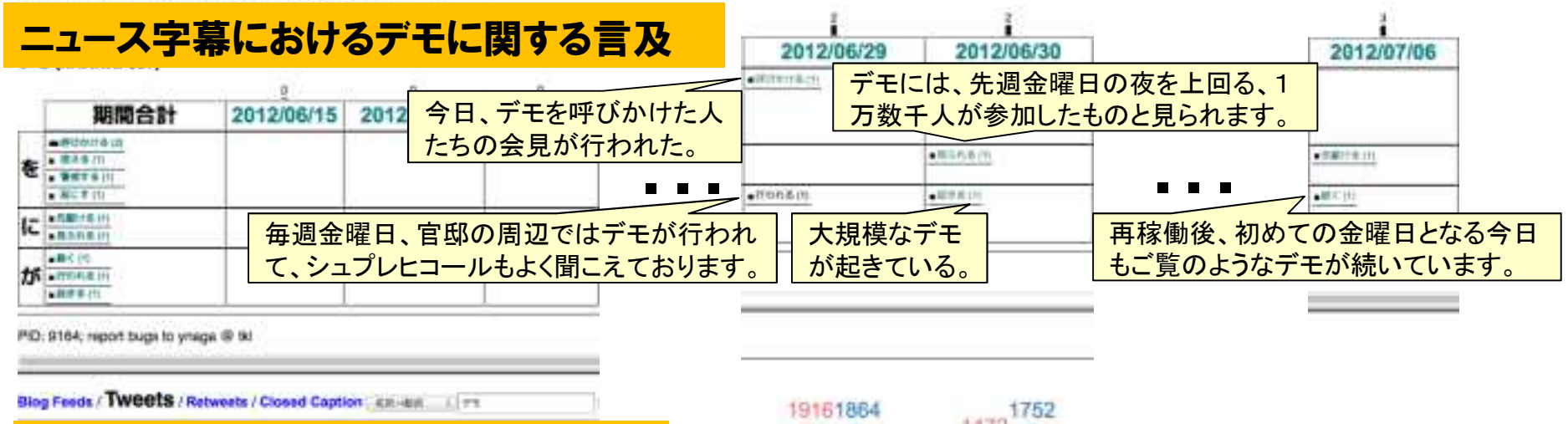
解析事例⑤

イベント・感情追跡システム

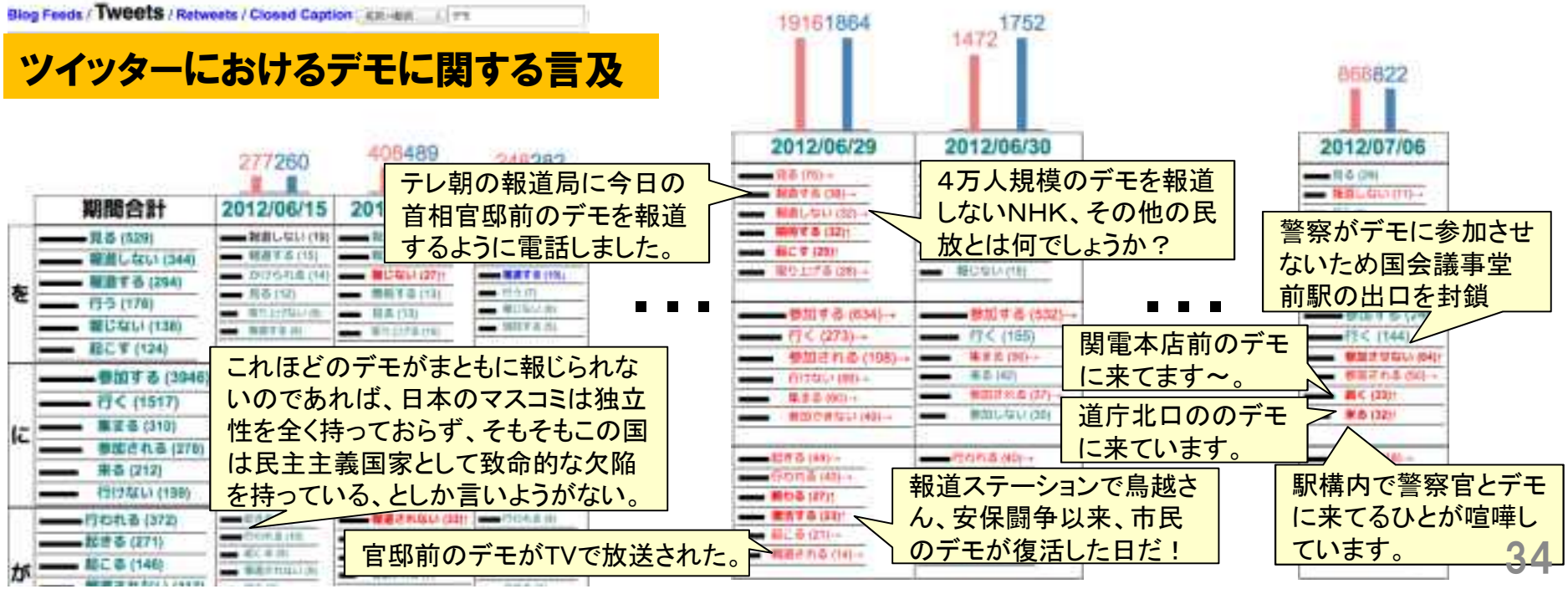
～頑健・高速な依存構造解析器J.DepPによる（ソーシャル）メディア解析～

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

ニュース字幕におけるデモに関する言及



ツイッターにおけるデモに関する言及



解析事例⑤

イベント・感情追跡システム

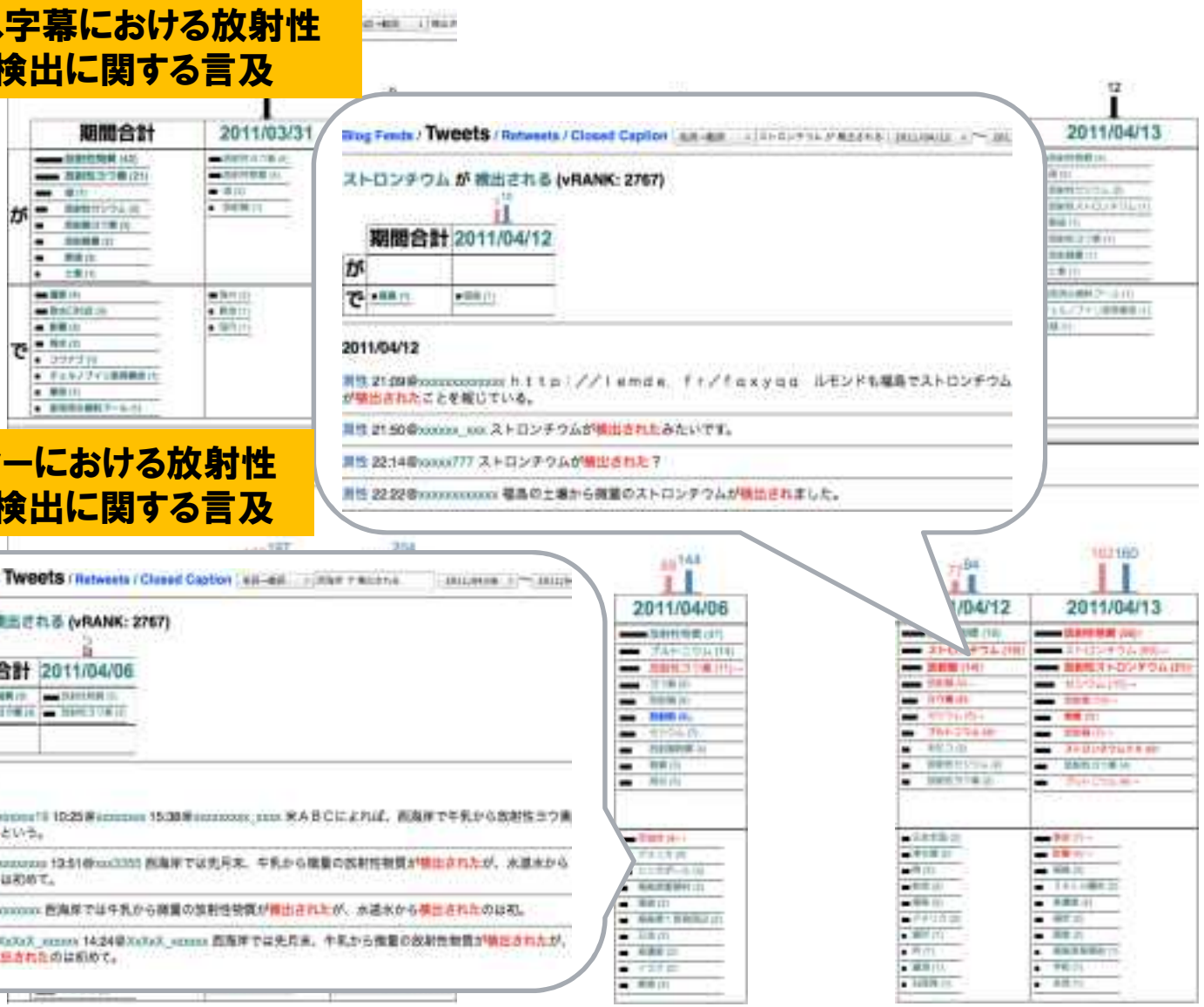
～頑健・高速な依存構造解析器J.DepPによる（ソーシャル）メディア解析～

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp

ニュース字幕における放射性物質の検出に関する言及

ツイッターにおける放射性物質の検出に関する言及



独創性・優位性

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

- **Webアーカイブ構築**
  300億URL、時系列収集、**世界最大級**
- **放送映像アーカイブ構築**
  多チャンネル、3年以上、30万時間以上、**国内外に例なし**
- **大容量・高スループット解析基盤**
  クラウド環境にて1日当たり1億ツイートを**リアルタイム解析可能**

- **高速／高精度画像・映像解析技術**
  TRECVID（2009コピー検出**世界最高速**、2010より二年連続物体検索**精度世界1位**）
  最高峰国際会議採択（ICCV, CVPR, ECCV, NIPS, ICDM）
  CM検出・同定**高精度・世界最高速**
- **自然言語処理に基づくWeb分析技術**
  最高峰国際会議採択（ACL, COLING, EMNLP）
  情報処理学会大会優秀賞
- **可視化技術**
  トップ国際会議採択（PVis）
  情報処理学会大会優秀賞、日本データベース学会年次大会優秀論文賞
- **高速Web構造解析技術**
  Webグラフ圧縮、**圧縮率で世界最高**
  高速最近傍探索、**高次元類似検索で世界最高速**

中間評価指摘事項への対応

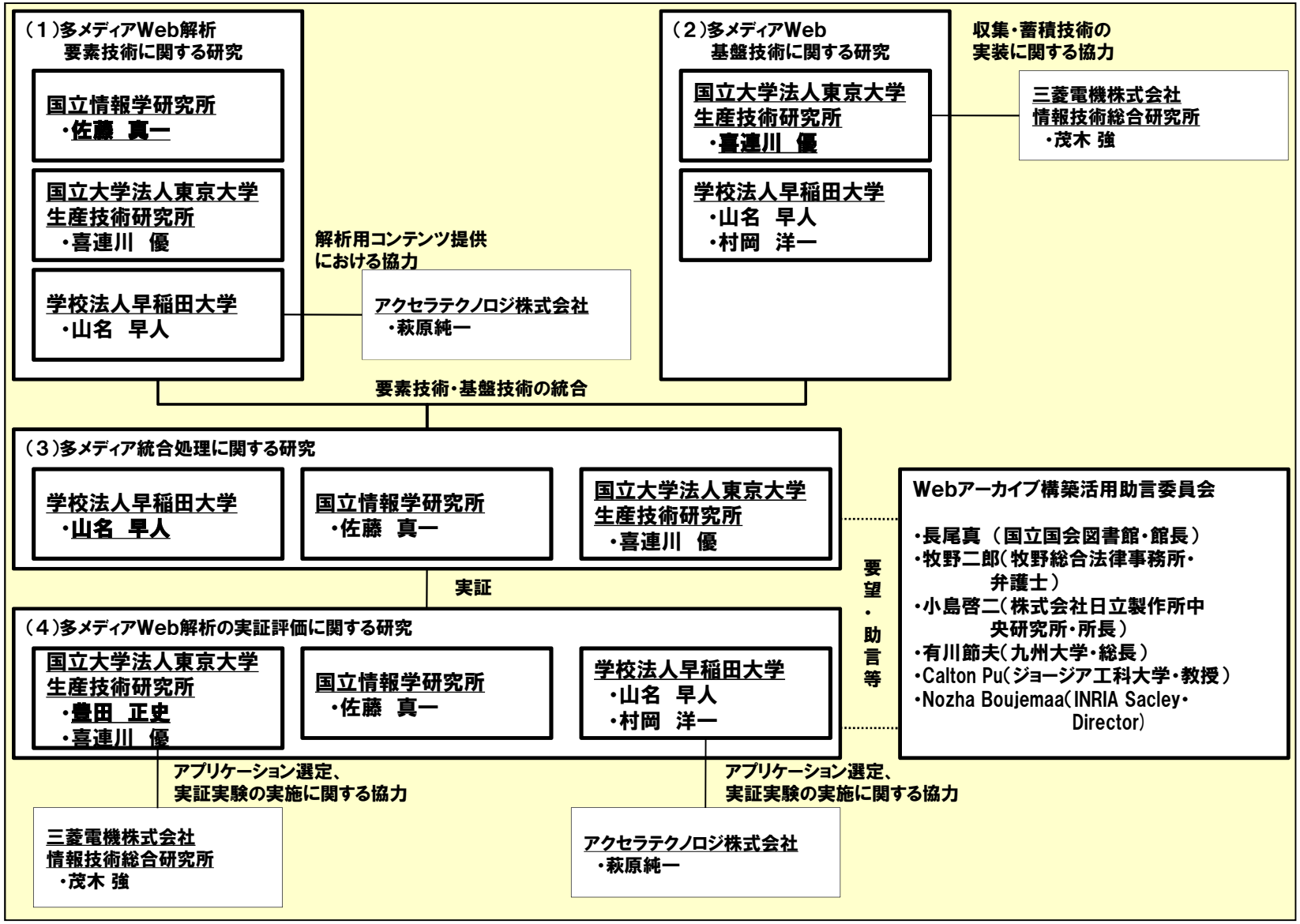
多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発

指摘点1：
今後、引き続き社会にとって価値のある成果となるよう本研究開発の実証段階において十分留意するとともに、社会にとってどのような価値のある成果が得られたのか、具体的なテーマをとりあげてアピールしていく必要がある。
対応：
具体的なテーマの一つとして、イベントの時系列的な流れに対するWeb（blog, twitter）と放送映像の反応と相互作用を解析することによる社会分析の実証実験を行い、東日本大震災、北朝鮮ミサイル問題、ロンドンオリンピック等のイベントに対する効果的な社会分析となっていることを示した。

指摘点2：
今後、実利用可能なソフトウェアの開発に向け、実証実験を行っていく上で社会学者、メディア研究者、広告代理店などの応用開発やサービスについて実務経験をもった人材の意見を取り入れていくことが重要である。
対応：
社会学者、メディア研究者、広告代理店などの応用開発やサービスについて実務経験をもった人材の意見の取り入れを図った。具体的には、NII共同研究の枠組みにより、東大・慶大・名大・筑波大・国文研等の社会学者・メディア研究者・言語学者らとの連携を実施した。また、言語学者、経営学者、広告会社、シンガポールHP、証券会社、Twitter社らとの既存の連携についても強化し、意見のくみ上げなどを通して実証実験につなげた。

研究開発体制

多メディアWeb解析基盤の構築及び社会分析ソフトウェアの開発



成果の利活用

- 本研究開発で開発したソフトウェアモジュールは**オープンソース**として提供
  - 画像・映像意味分類システム
  - 組み合わせ素性を用いた分類器
  - 高速系列ラベリングアルゴリズム
  - データインテンシブスケジューリング
- 実証実験等を協力企業等と連携して実施することを通し、新しい社会分析が可能であること示し、**企業との連携による研究開発成果の実用化**を促す。
- 本研究開発で構築したWebアーカイブについては、様々な学術利用を可能とするために適切な検索・データ取得インタフェースを提供し、**研究者向けに公開**することを検討している。ただし、著作権法等法的に問題ない範囲での公開を検討していく。

人材育成

	ポスドク	博士後期課程	修士課程
国立情報学研究所	4名	2名	
	2012.10 Google 2013.4 NTT	在学中	
東京大学	2名		
	2012.4 特任准教授 2012.7 特任准教授		
早稲田大学	1名	3名	7名
	2013.4 明治大学准教授	2012.4 本学助手2名 2013.4 日本IBM東京基礎研究所	ヤフー、楽天、Mixi、証券系システム他

今後の展望

多メディアWeb解析基盤の構築
及び社会分析ソフトウェアの開発

- **本事業に関する今後の計画**
  - アーカイブについては、当面は引き続き管理・運用を進める。公開についても引き続き検討を進める。
  - 技術的成果については、実利用・実用化の促進に努める。

- **今後の研究開発への展望**
  - 大規模アーカイブの整備、研究用の公開、解析のための計算資源などのインフラ整備等が重要
  - アーカイブのコンテンツのグローバル化は極めて有望
  - データの大規模化に伴う、よりシビアな要望に応えうる技術の研究開発の推進
  - 多様なデータを扱うさまざまな技術の適切な融合
  - アーカイブに基づくベンチマークの整備による研究開発の特段の推進

国際シンポジウム開催

多メディアWeb解析基盤の構築
及び社会分析ソフトウェアの開発

- 2013年3月13日に国際シンポジウムを開催
- 2件の基調講演（海外著名研究者）
- 3件の講演並びに10件のポスター・デモにより成果公開
- 参加者138名
- 電機メーカ（23%）、ITサービス（18%）、放送・通信（12%）等、企業からの参加が大半