

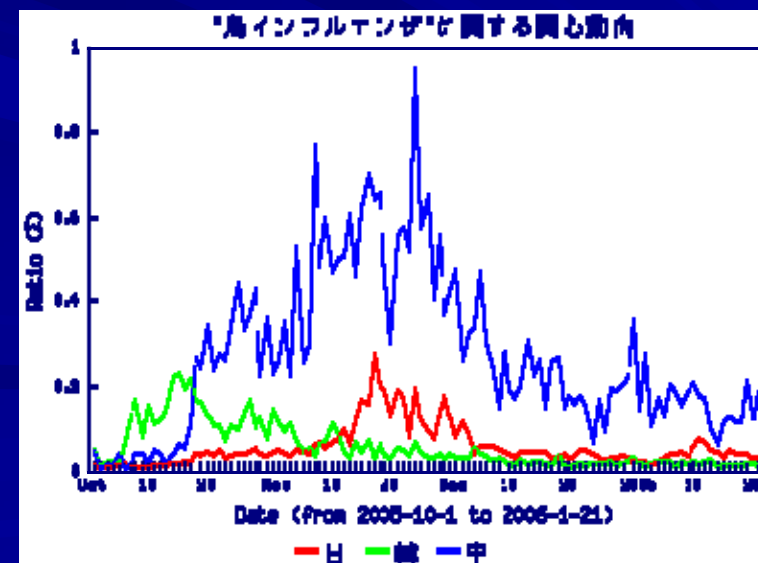
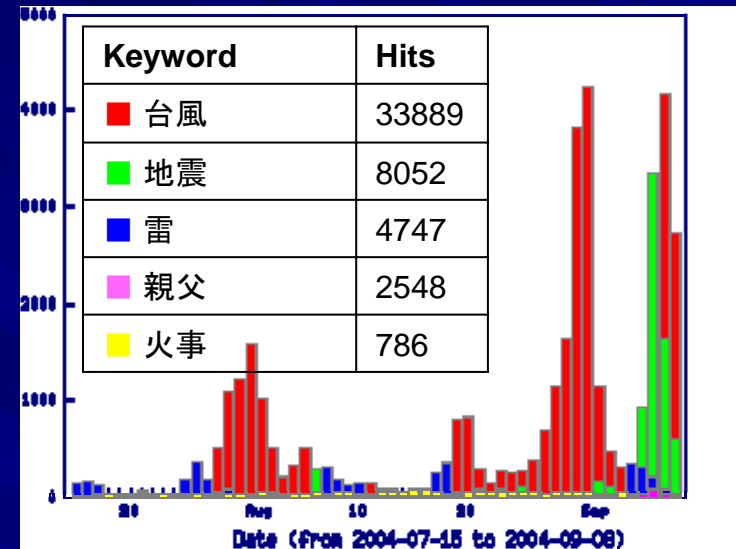
多言語マイニング： Weblogから世界の関心を探る

東京大学人工物工学研究センター
価値創成イニシアティブ
(住友商事) 寄附研究部門
福原知宏

2006年12月9日(土)
科研「情報爆発IT基盤」: NLPミーティングにて発表

はじめに

- Weblog記事を用いた社会的関心の分析について、多言語での関心分析を織り交ぜながら述べる
- 分析の視点
 1. 社会的関心のパターン
 2. 共起語を用いた分析
 3. 多言語関心分析
 4. 感情表現を用いた分析
 5. 実世界データとの比較



社会問題と関心

■ 社会問題と社会の関心

－ 様々な社会問題

- 地震, 台風, 津波, テロ, 地球温暖化, 異常気象, 狂牛病, SARS, 鳥インフルエンザ, 情報漏洩, いじめ問題

...

－ 問題の解決に向けて

- 問題に対する社会の関心を把握することが重要!
 - － 人々がどのような問題に関心を寄せているかを知る

地震
(阪神淡路大震災 (1995))

SARS

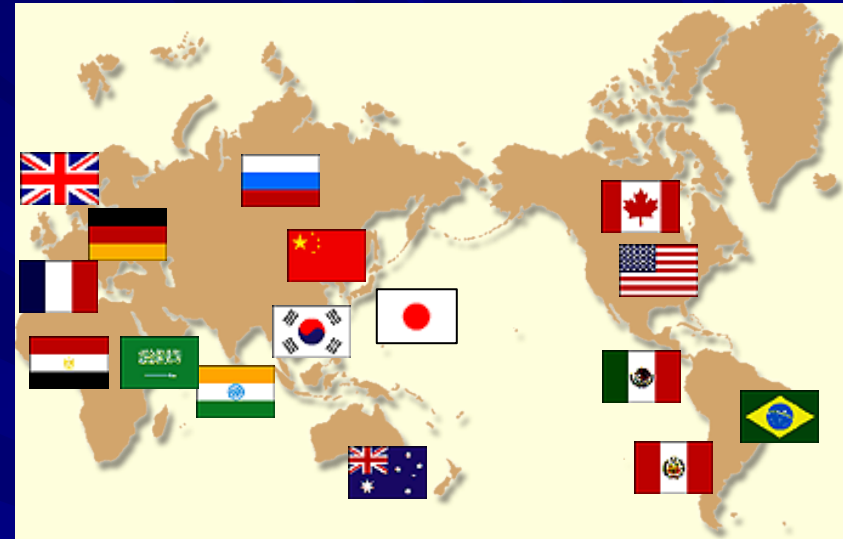
狂牛病

関心の言語横断比較

- 日本だけでなく世界の人はどんな問題に関心を抱いているか？また、それらの問題は日本とどう関係しているのか？

－ 世界の関心を知ることで日本の役割も見えてくるのではないか？

- 各国のブログ記事を集めて解析する



鳥インフルエンザ

SARS

BSE

自然災害

イラク問題

北朝鮮

核実験

ミサイル発射

Weblog記事を用いた 関心解析システム: KANSHIN

■ Weblog記事を言語横断的に収集し、各言語コミュニティの関心を解析するシステム

— 現在、日中韓英の4言語を対象

■ 機能

— 記事検索機能

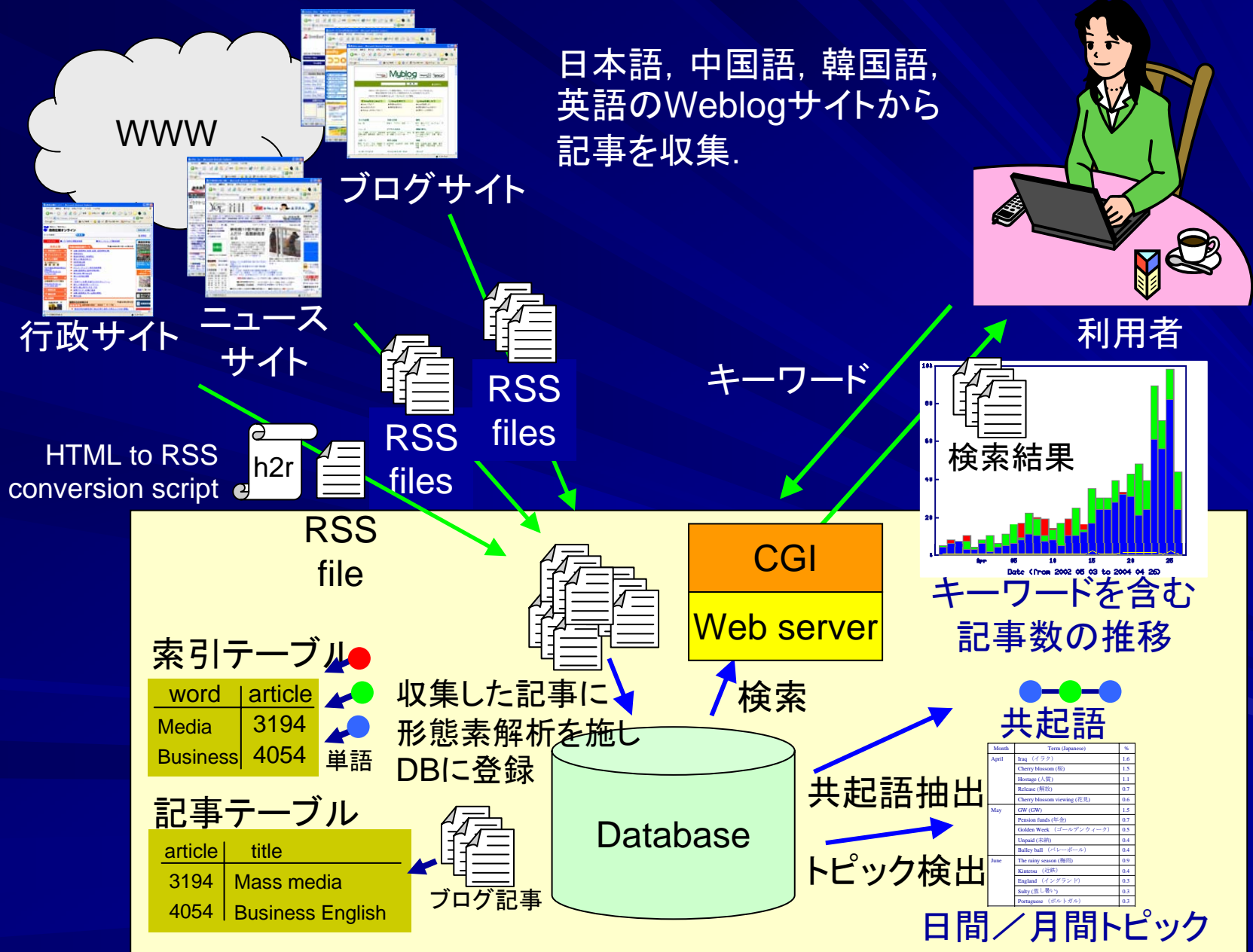
— 話題抽出機能

■ 月間トピック, 日間トピック

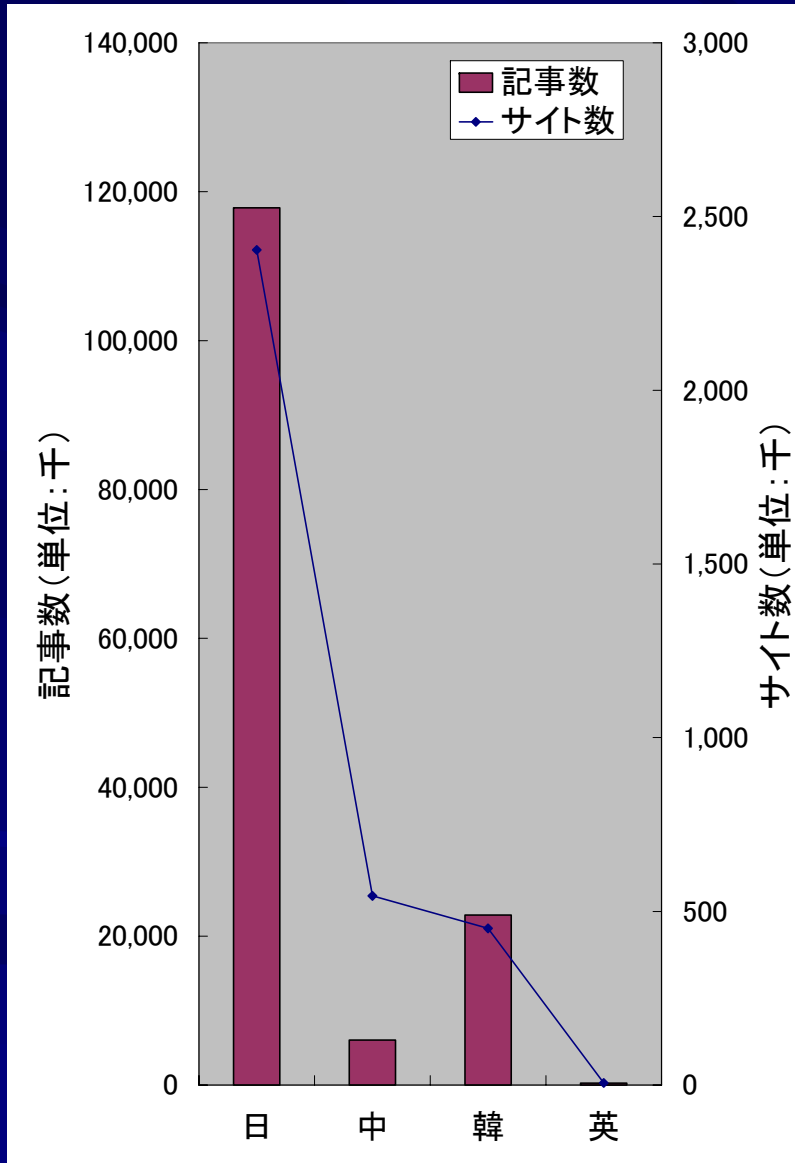
— 共起語検索機能



システム全体像



登録件数



■ 収集開始時期, 経過日数

- 2004/3/18~; 996日(日本語)
- 2004/12/1~; 738日(中国語)
- 2005/8/1~; 634日(韓国語)
- 2006/11/10~; 29日(英語)

■ 記事数

- 日: 117,815,931記事
- 中: 6,052,115記事
- 韓: 22,850,875記事
- 英: 255,747記事

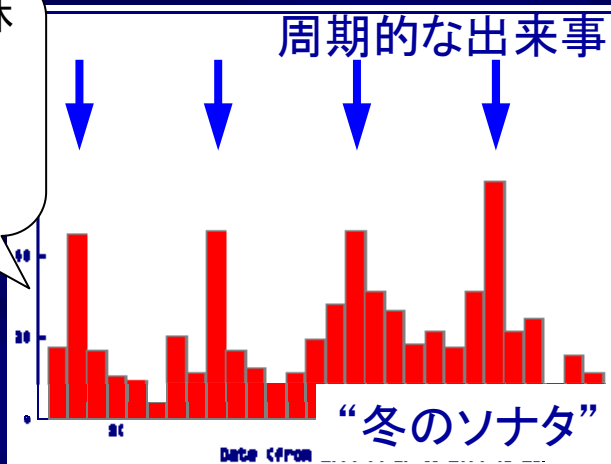
■ ブログサイト数

- 日: 2,403,799サイト
- 中: 544,349サイト
- 韓: 451,301サイト
- 英: 6,261サイト

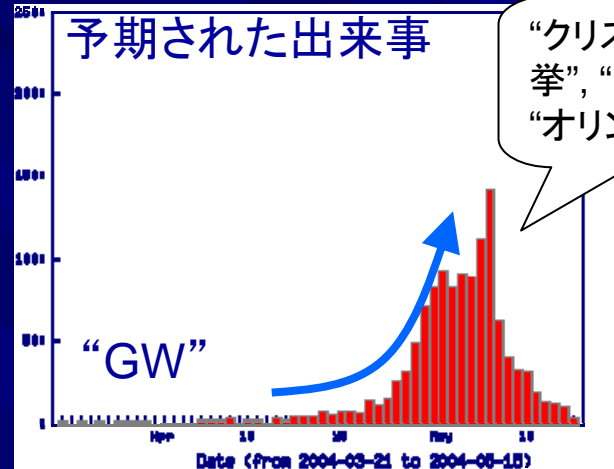
社会的関心の分類

Weblogに見られる関心パターン

“給料日”, “休日”, “BBQ”, “家族連れ”, “結婚式”, “同窓会”



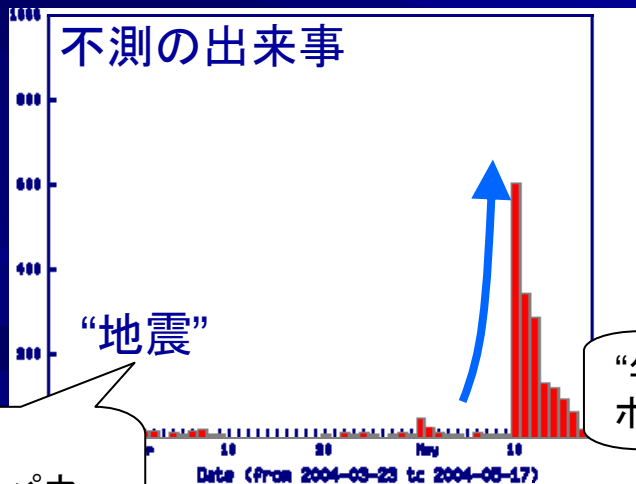
I. 周期型



“クリスマス”, “選挙”, “台風”, “夏”, “オリンピック”

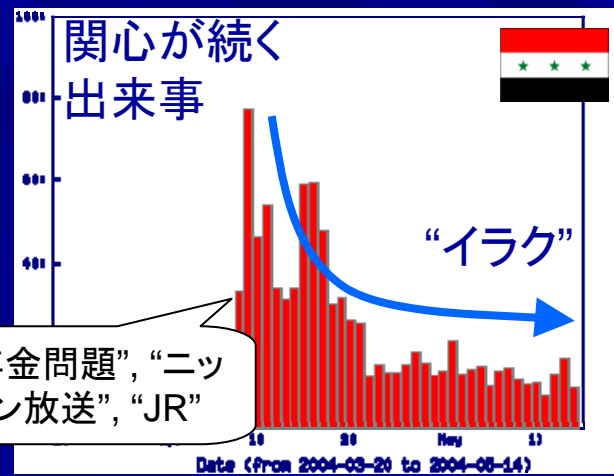
II. 漸次増加型

“酸性雨”, “大気汚染”, “水不足”



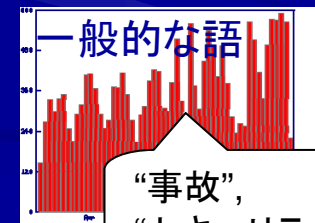
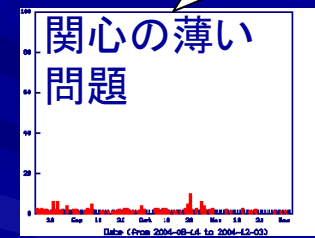
“津波”, “ヨハネ・パウロ”, “脱線”

III. 突発型



“年金問題”, “ニッポン放送”, “JR”

IV. 関心持続型



“事故”, “セキュリティ”

V. その他

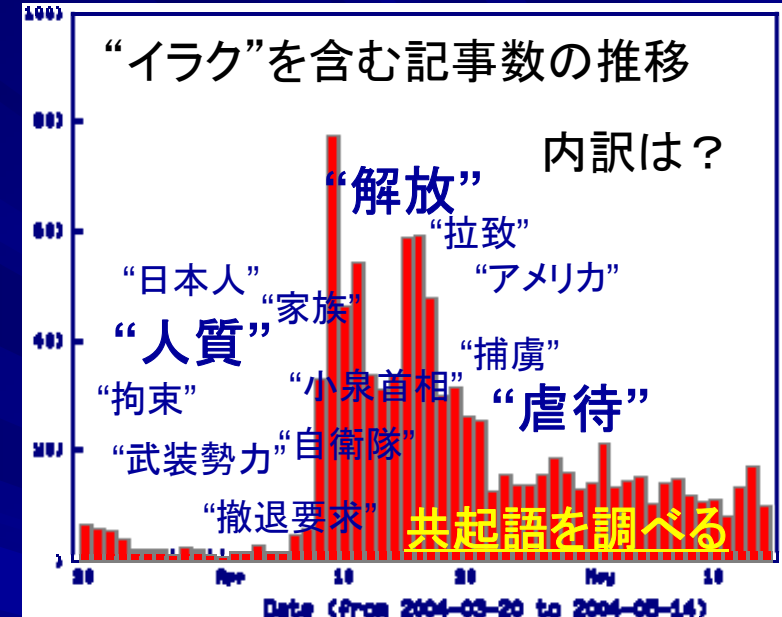
共起語を用いた分析

どんな言葉と共起していたか？

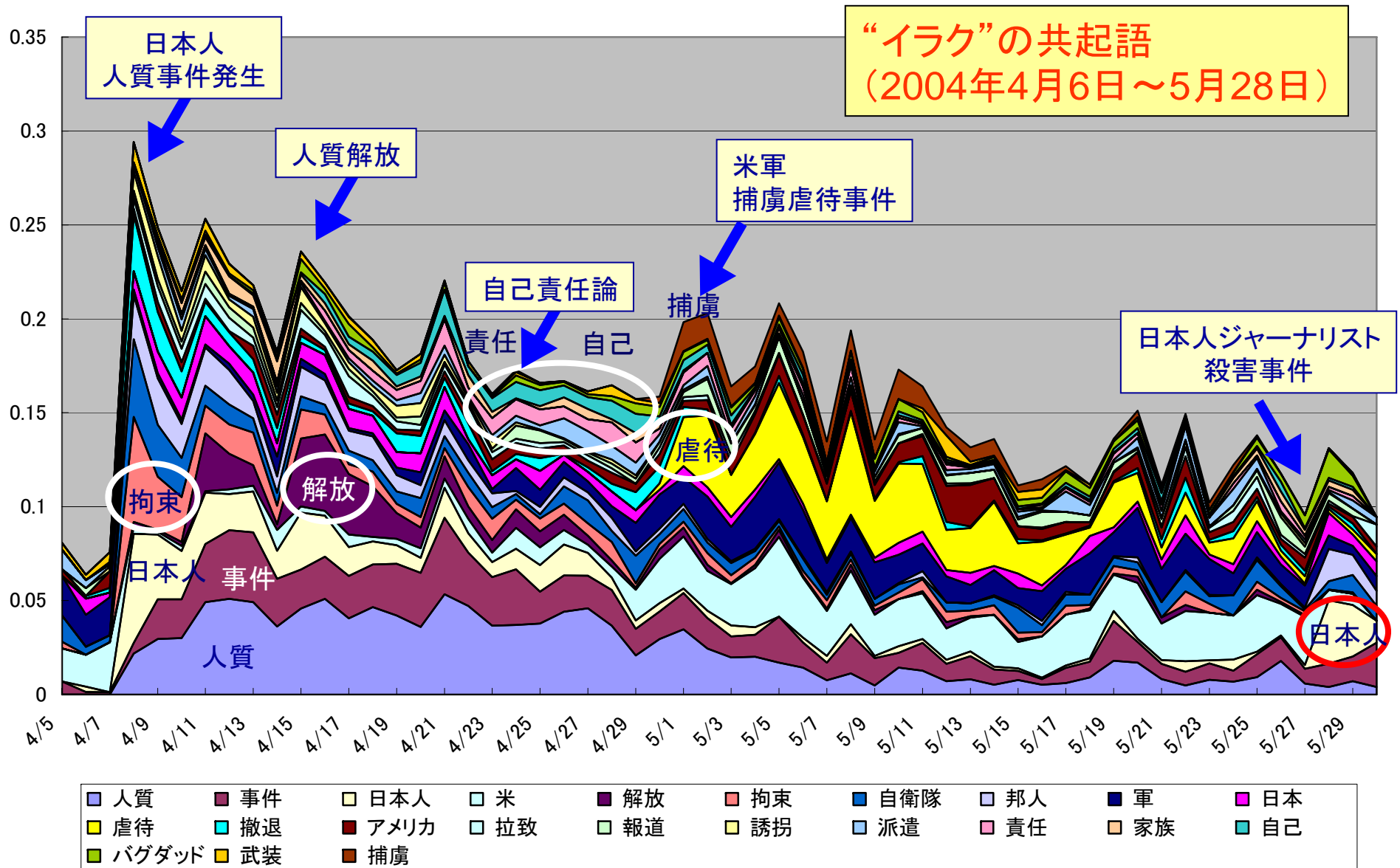
問題の焦点は何か？

共起語を用いた分析

- 出来事のどの部分に焦点が当てられていたか？
 - － 記事数だけでは分からない
 - その出来事がどの言葉と共起していたか、共起語を知る必要がある。
- 時間軸上の共起語の推移を調べることで焦点の推移が分かる。

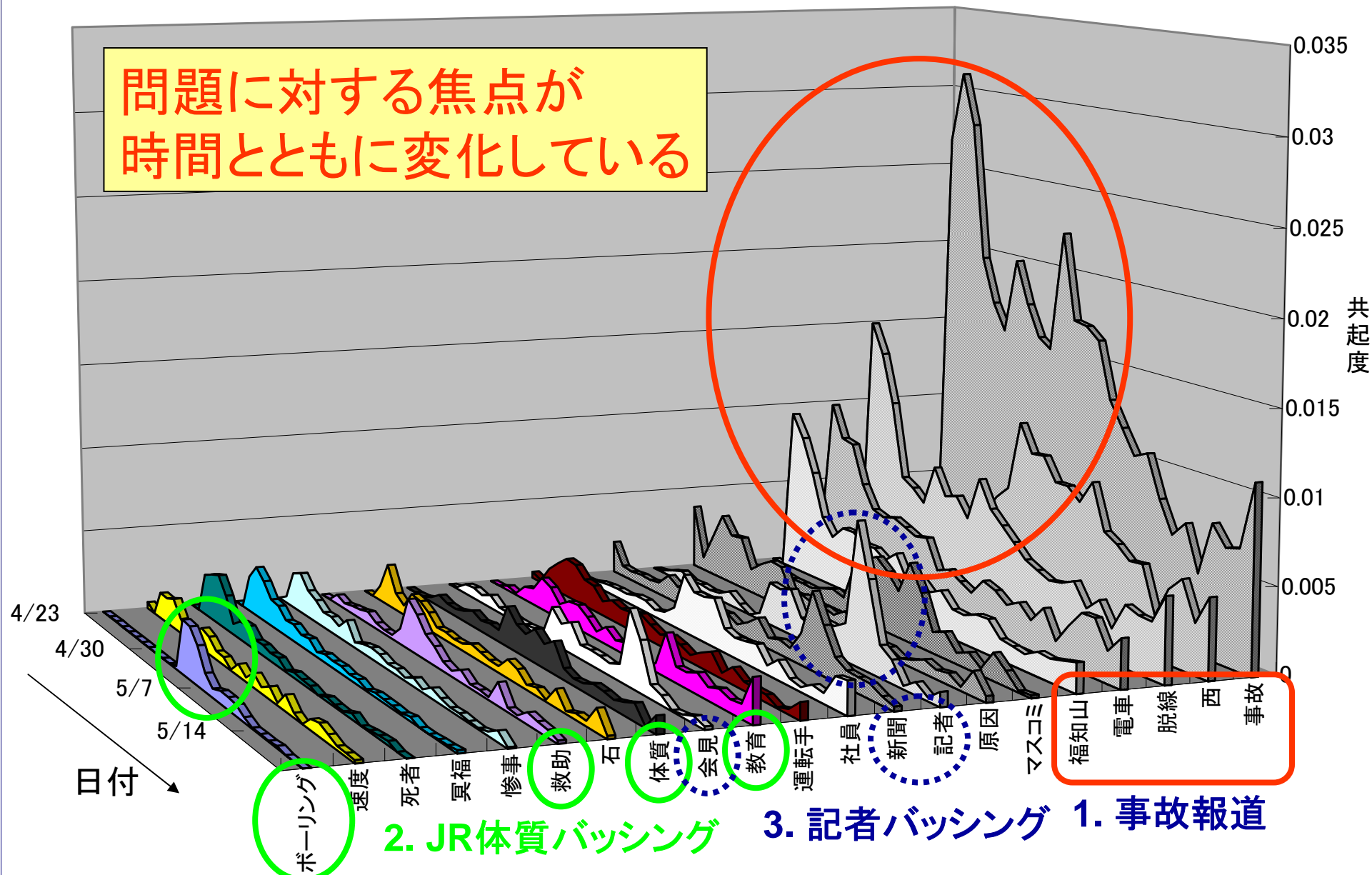


共起に基づく分析(日本語ブログ記事) 2004年4月イラク人質事件



共起に基づく分析(日本語)

2005年4月JR西日本福知山線脱線事故

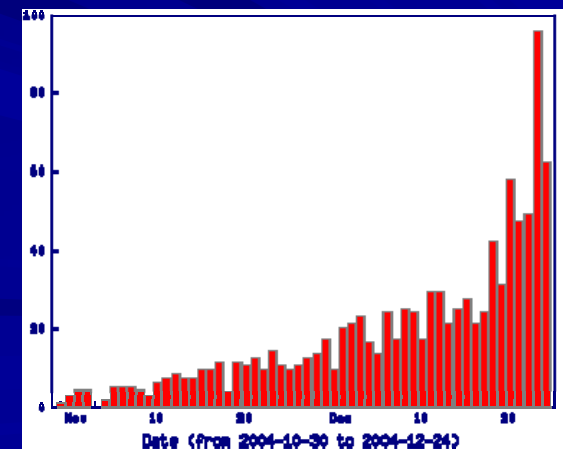
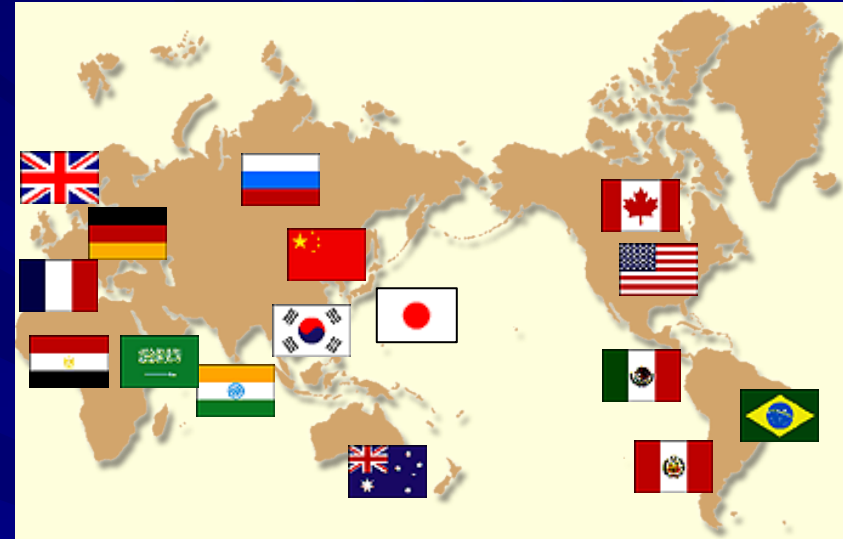


言語横断型関心分析

国内と海外の関心を比較する

言語横断型関心分析

- 各国の関心を知りたい
 - ある出来事に関する各国の関心を知りたい
- あるキーワードについて他の言語コミュニティの関心と比較する
 - 入力
 - 日本語キーワード
 - 出力
 - それぞれの言語での検索結果



Wikipediaを使った対訳表現検索

入力



ワールドカップ



Wikipedia:
日本語記事

他の言語

- Alemannisch
- العربية
- Български
- Bosanski
- Català
- Český
- Cymraeg
- Dansk
- Deutsch
- Ελληνικά
- English

他の言語の
同一エントリーへのリンク

出力



축구 월드컵

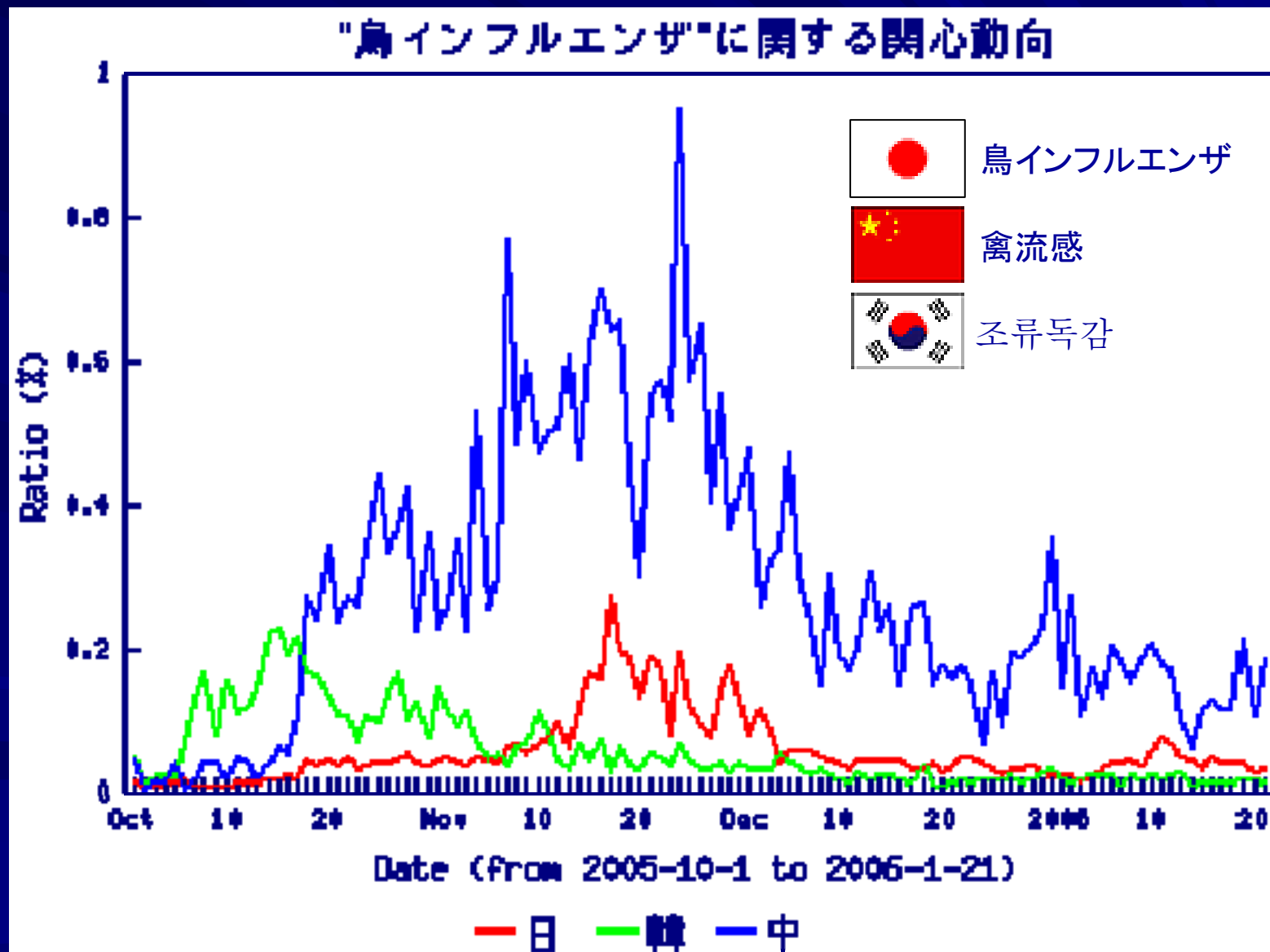


世界盃足球賽

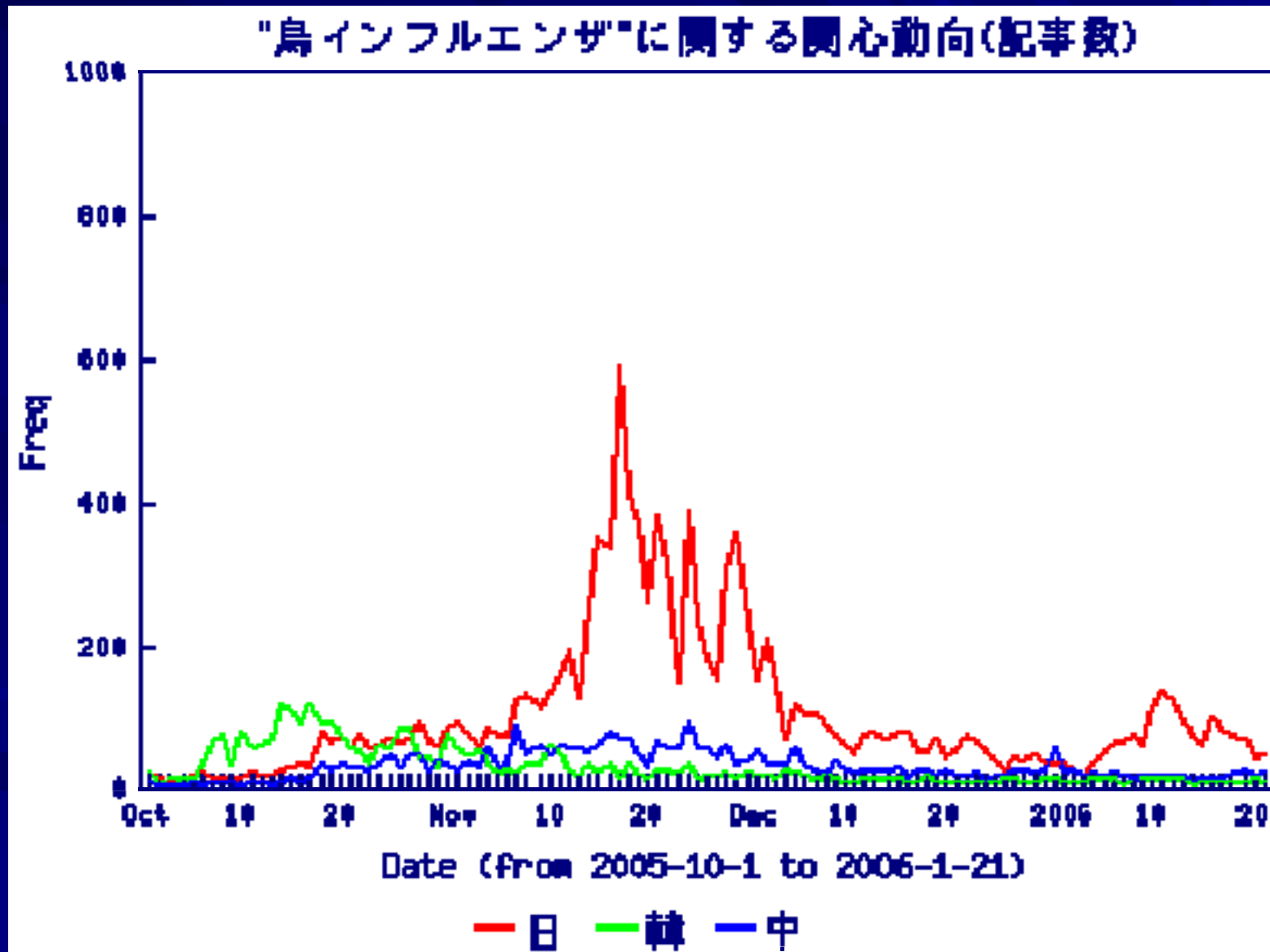
■ 処理概要

1. 日本語キーワードをWikipedia で検索
2. (もしエントリがあれば)それぞれの言語へのリンクをたどり、対訳表現を得る
3. 得られた対訳表現を用いて記事検索を行う

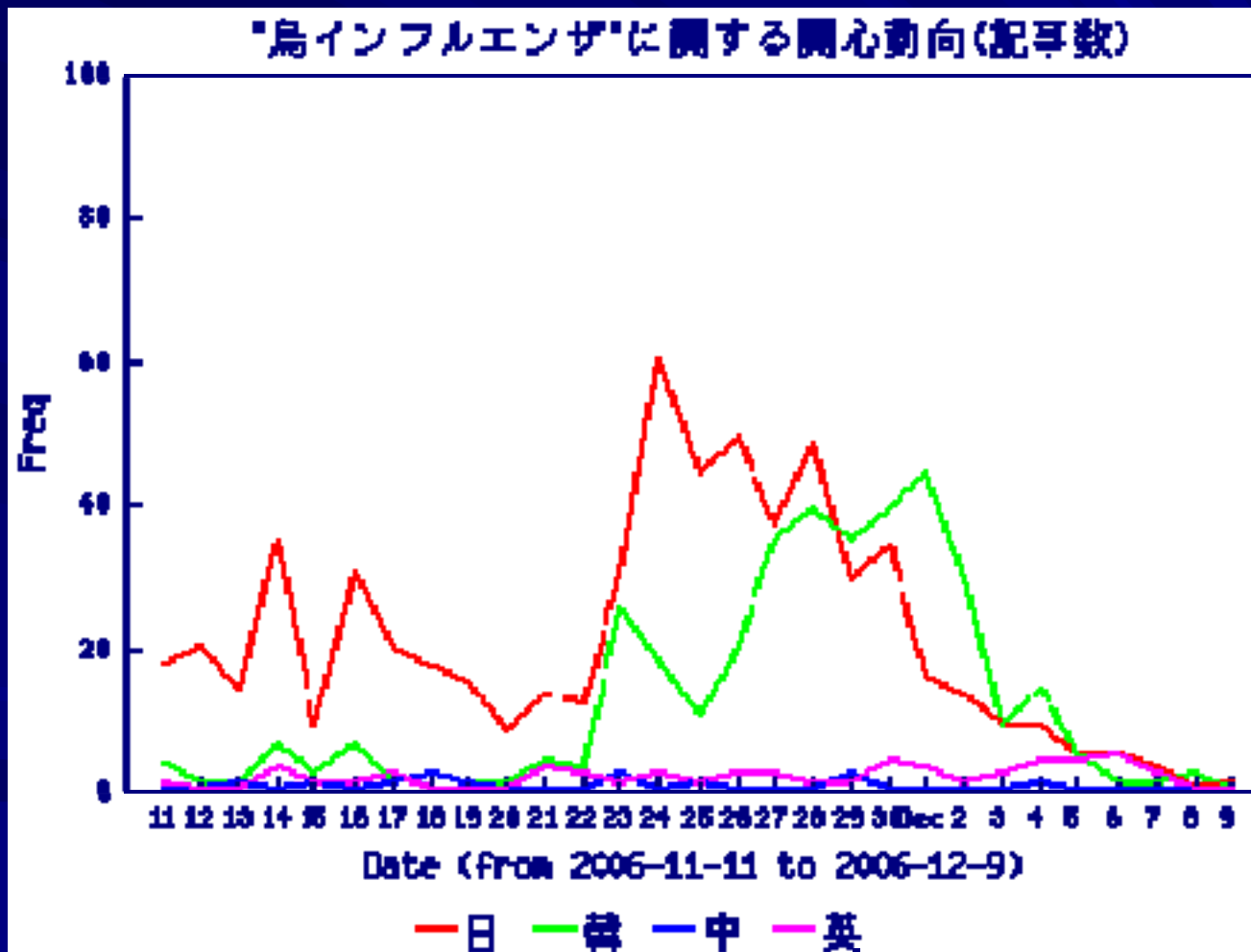
鳥インフルエンザに関する関心比較



記事数の比較

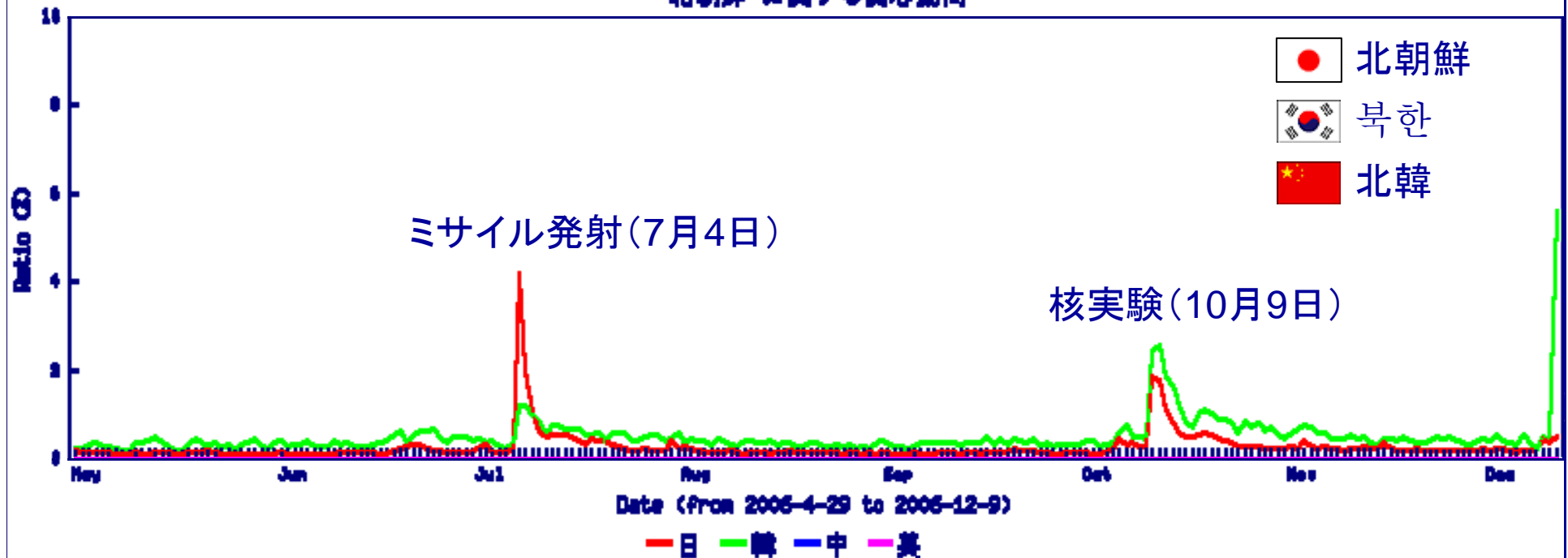


鳥インフルエンザに関する 今年の関心



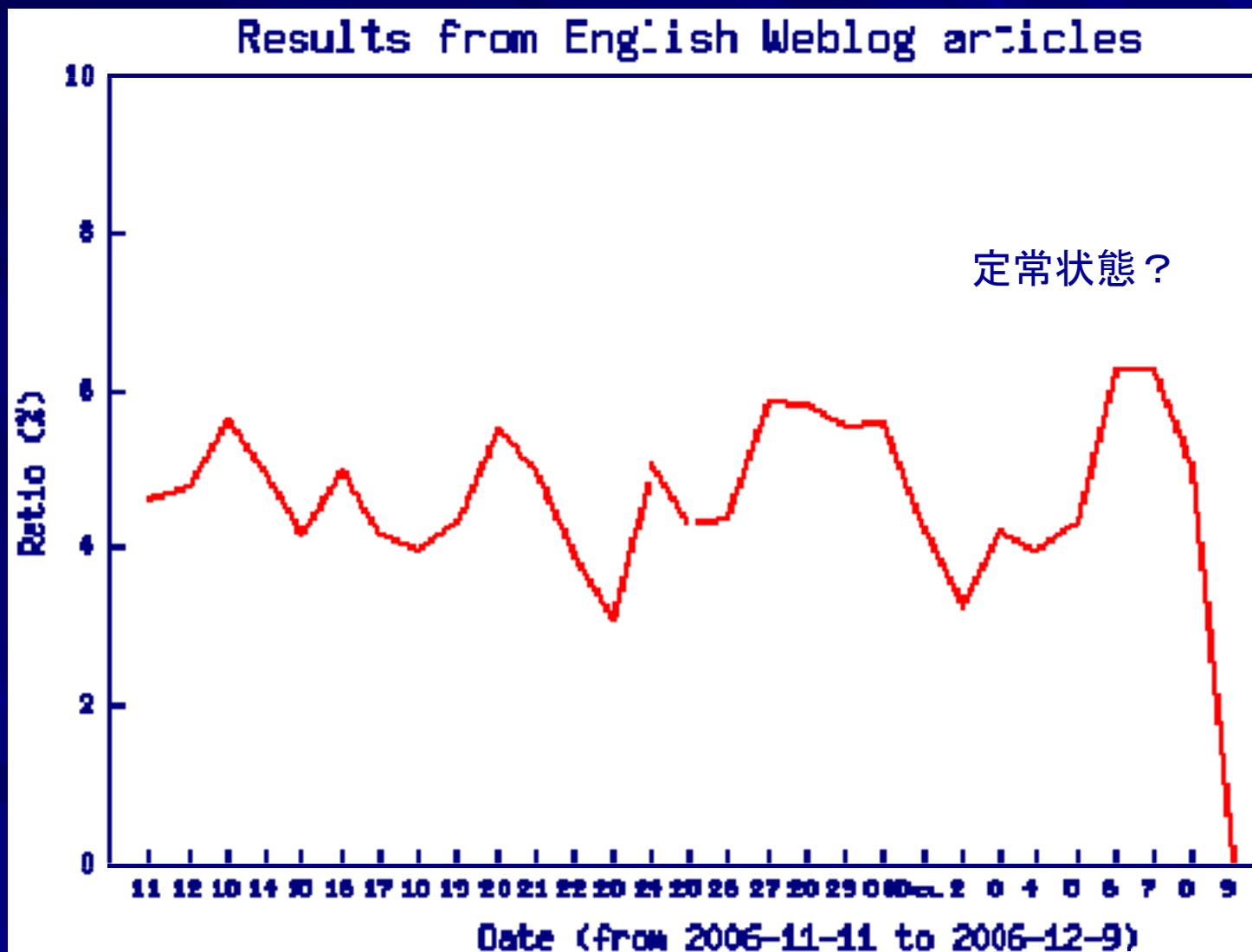
北朝鮮に関する日中韓関心比較 (2006/4/29-12/9)

"北朝鮮"に関する関心動向



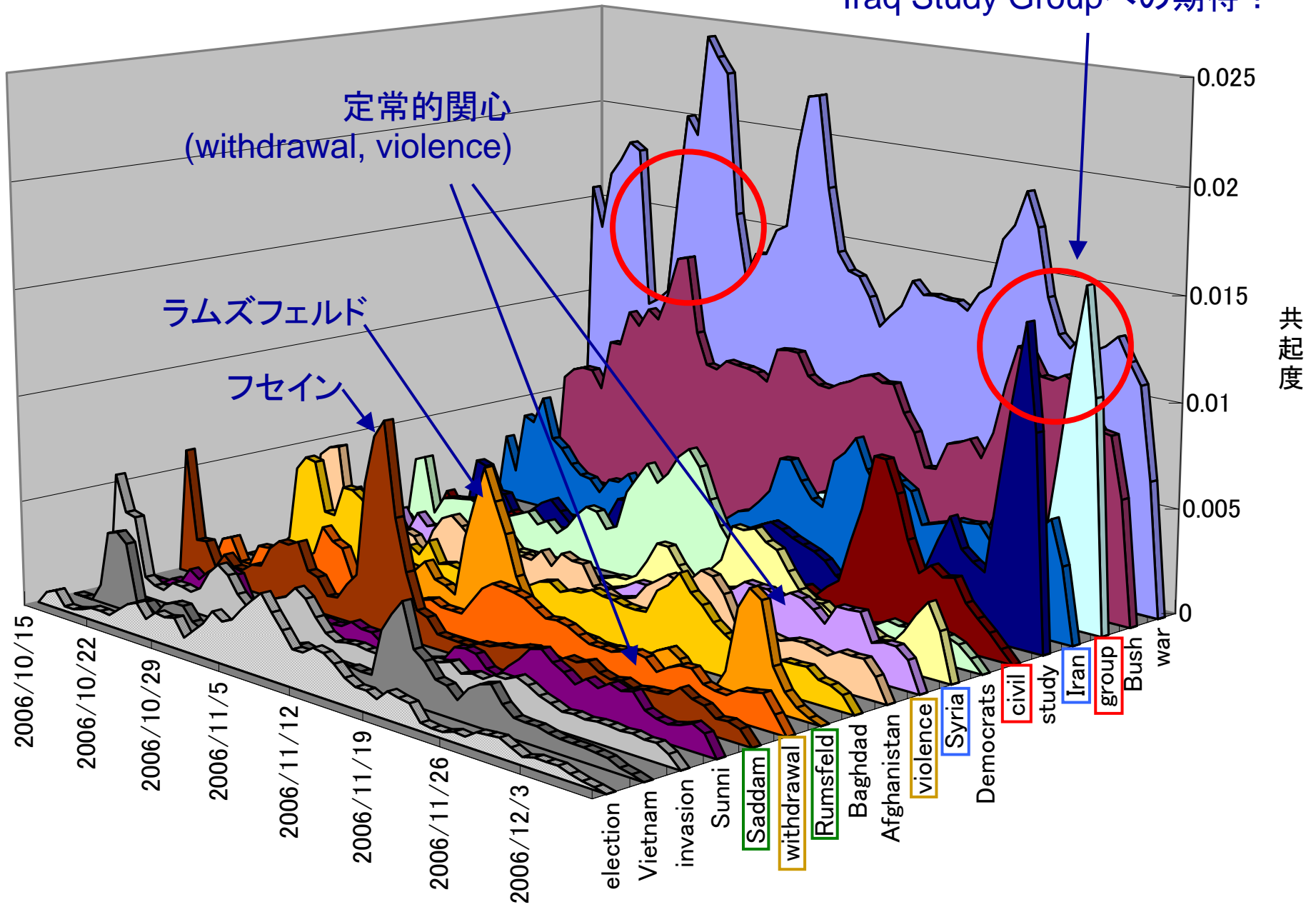
日本と韓国の関心は一致している
中国語ブログは反応していない？
(キーワードが良くない??)

英語圏の“Iraq”に対する関心

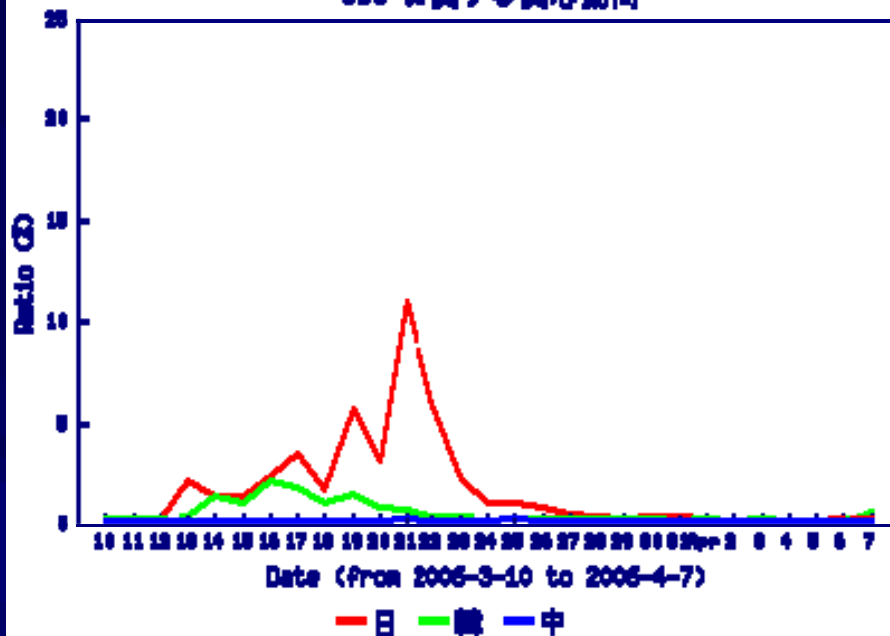


“Iraq”の共起語 (移動平均3を適用)

Iraq Study Groupへの期待?



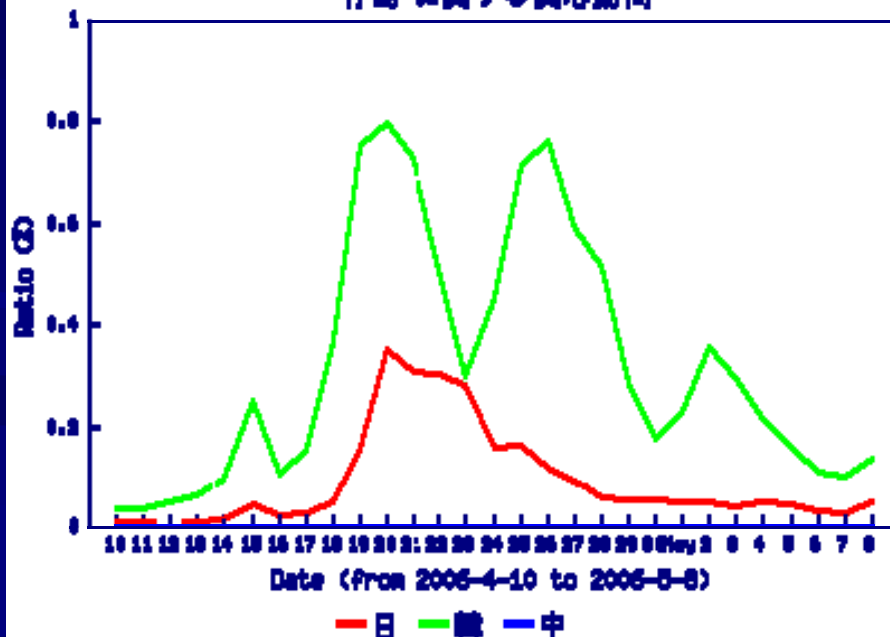
'abc'に関する関心動向



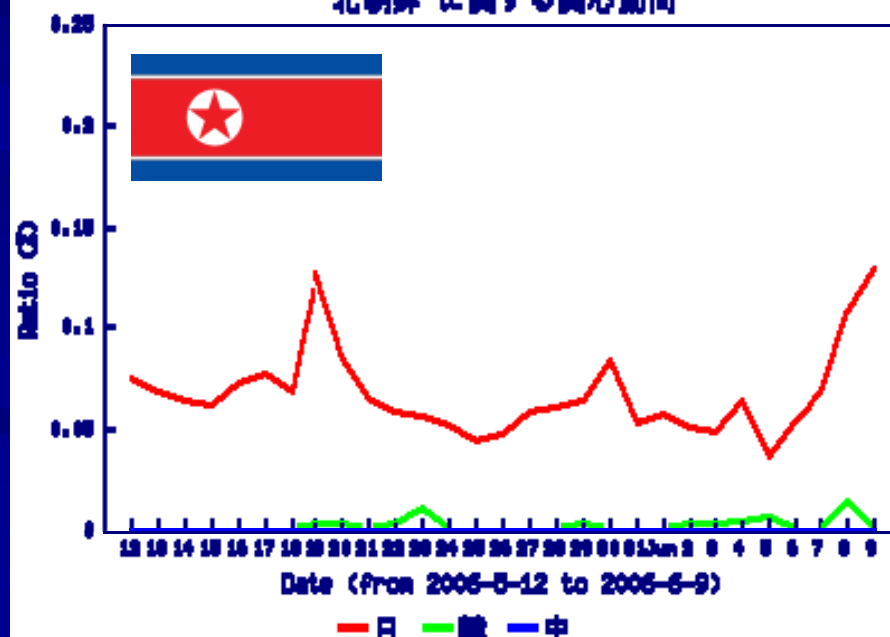
'ワールドカップ'に関する関心動向



'竹島'に関する関心動向



'北朝鮮'に関する関心動向



言語横断検索の課題

■ 一般名称の取得

– 正式名ではなく一般に使われている言葉を知りたい.

■ 例：北朝鮮

 日：朝鮮民主主義人民共和国

 中：朝鮮民主主義人民共和国

 韓：조선민주주의인민공화국

 英：North Korea

■ 韓国では「北」とも「북한」とも. 中国では「北韓」, 英語では“N.Korea”, “DPRK”という呼び方もある.

感情表現を用いた分析

人は事件や事故に対して
どのような感情を抱いていた
か？

社会の出来事と感情

■ 企業や行政に対する社会の感情的反応

－ 例)

- 雪印集団食中毒事件
- 東京電力トラブル隠し問題
- 年金問題
- 靖国神社参拝

－ 中国, 韓国での受け止め方

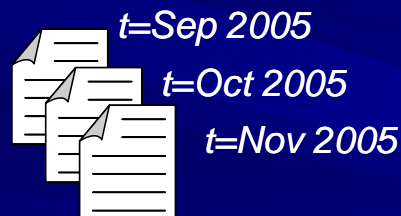
■ 社会の出来事に対して人々がどのような感情を抱いたか？

フランス暴動事件

NHK記者放火事件

感情と話題推移の把握： 感情表現を用いた話題検出手法

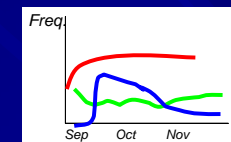
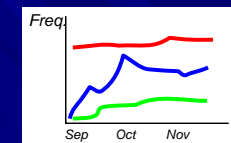
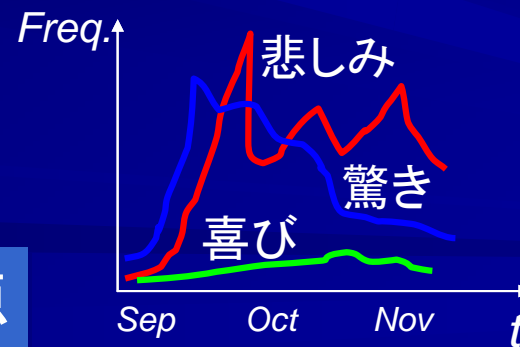
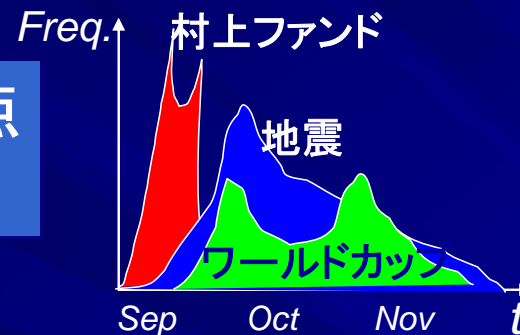
1. 話題推移の観点 (トピックグラフ)



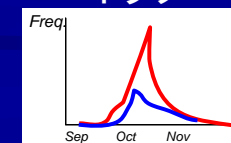
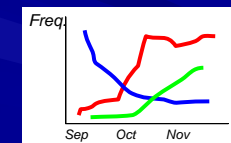
時系列

テキスト集合：
(タイムスタンプ付き
テキスト集合)

2. 感情推移の観点 (感情グラフ)



感情を指定



話題を指定

感情表現の例

■ 不安

- 不安, 心配, 怯えた, 気掛かり, 胸が騒ぐ, 安心できない, 青ざめた

■ 悲しみ

- 悲しい, 涙が, 目頭を押さえた, すすり泣く, 嗚咽を

■ 怒り

- 怒り, 苛立ち, 語気を強めた, 許せない, 憤った, 声を荒げた, 厳しく非難

■ 安心・喜び

- 安心, 喜んだ, ホッとした, 躍り上がった, 笑顔で, 胸をなでおろした

■ 苦悩

- 苦悩, 苦渋, 頭を抱えて, やりきれない, 憂うつ,

■ 疲労

- 疲労, 疲れ, ぐったり, うんざり, 途方に暮れた, 疲れ果て, 疲労困ぱい, 残念,

■ 不満

- 不平, 不満, 納得いかない, ぶ然と, 難色を示した, 不公平, 不服,

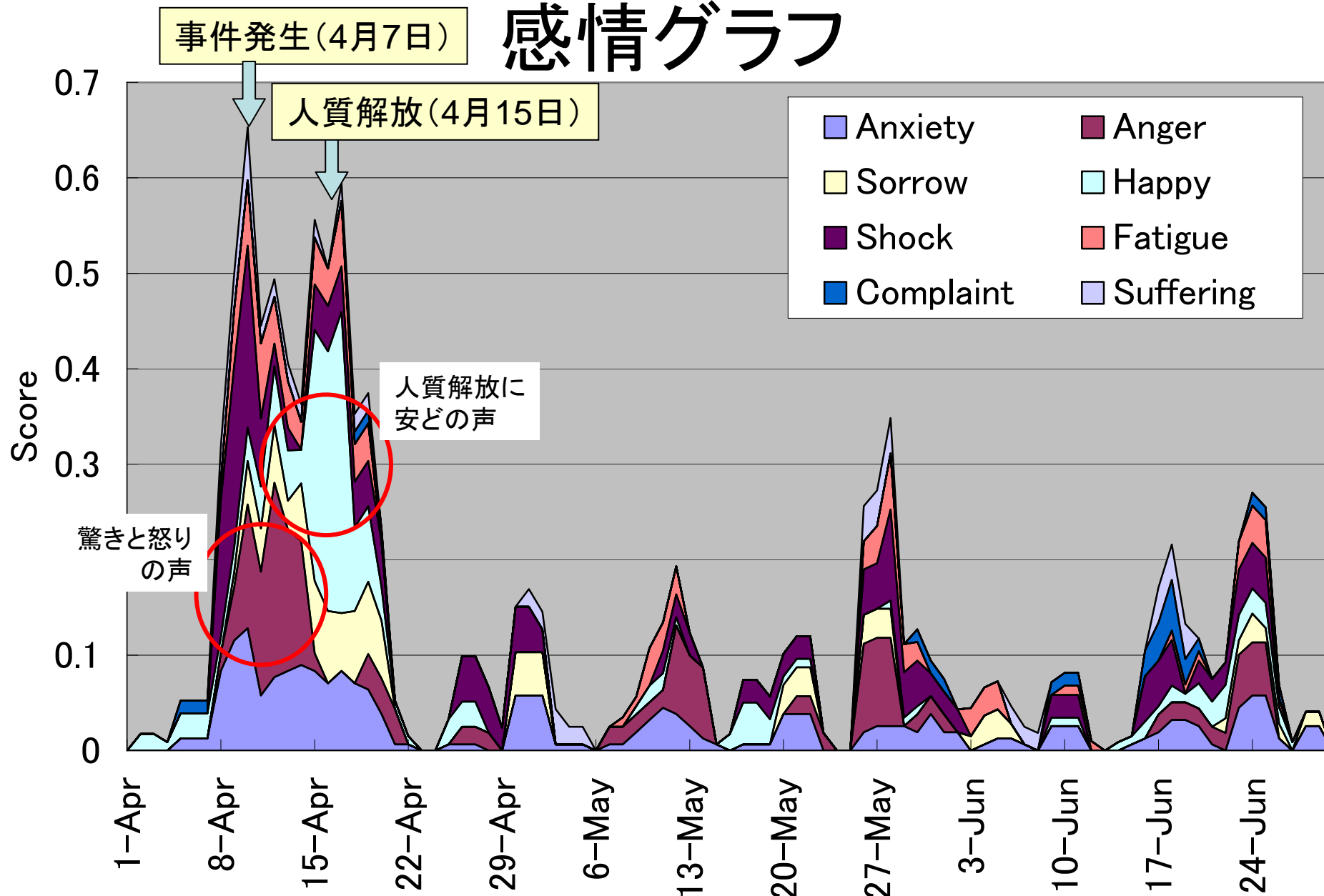
■ 衝撃

- ショック, 驚き, 衝撃, 啞然と, 慌てて, 興奮して, 動転して

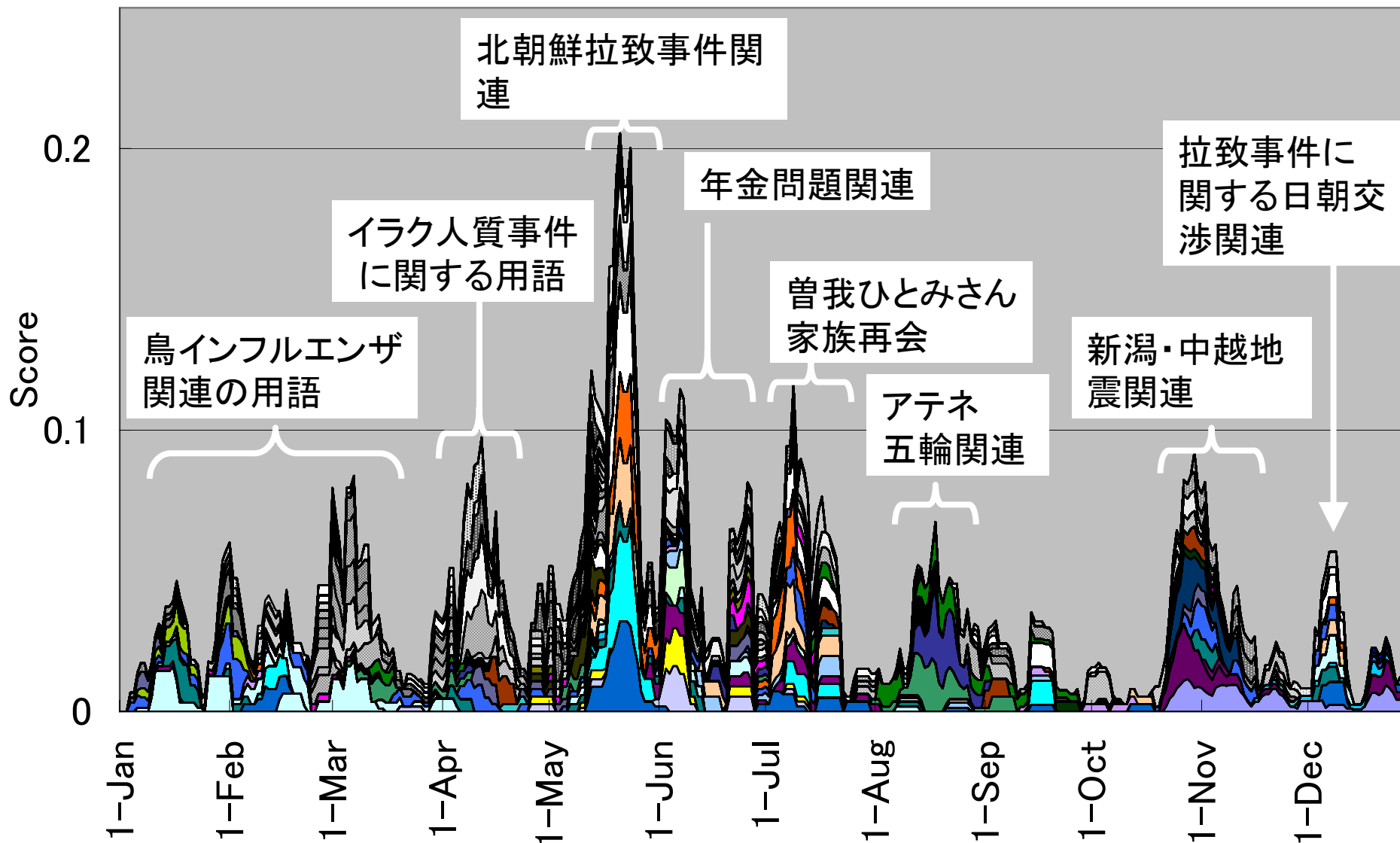
※2003年に発行された全国紙4紙, 地方紙3紙のうち化学工場事故に関する記事から383個の感情表現を手動で抽出

2004年4月イラク人質事件に関する

感情グラフ



“不安”に対するトピックグラフ (朝日新聞2004年版から作成)



実世界データとの関連

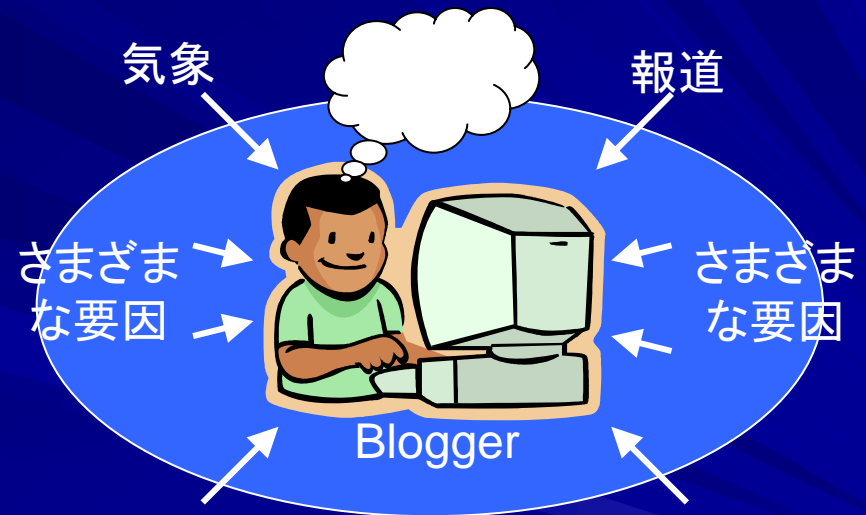
実世界データとの関係

- ブログは実世界の様々な影響を受けている.

- e.g., 気温, 天候, 報道 (テレビ, ラジオ, 新聞), 社会情勢, 文化・伝統, 人間関係など

- 社会的関心と実世界データとの関係を探る

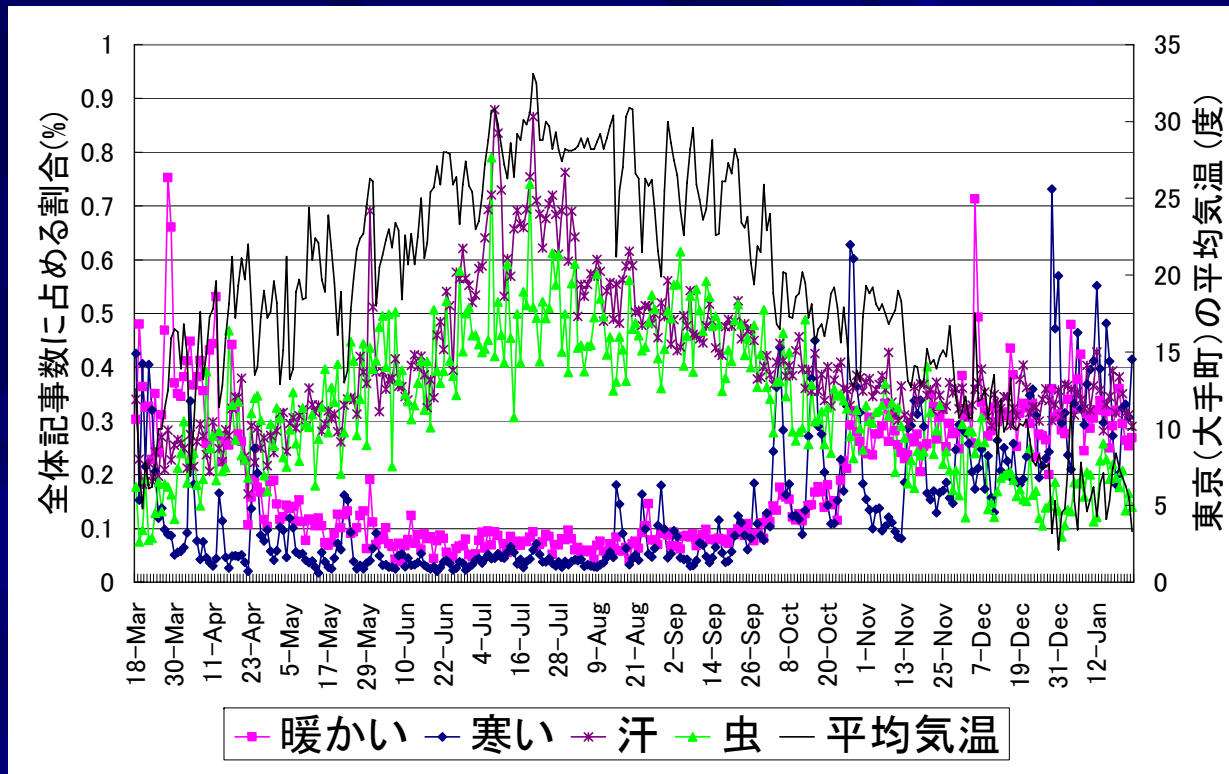
- 個人や社会の関心は実世界の影響をどの程度受けているのか?
- ここでは気温との関係について調査した



社会情勢

文化・伝統

気温と関心の関係

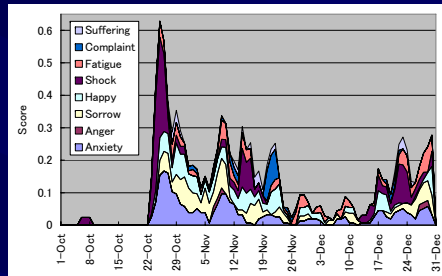


- 2004年の平均気温(東京・大手町)とWeblog中に出現する用語の関係を見る
- 正の相関を持つ言葉
 - e.g, “汗”, “虫”, “蚊”, “ゴキブリ”, “日焼け”, “エアコン”, “暑苦しい”, “ビール”, “スイカ”, “枝豆”, “汗だく”, “ゴーヤ”, “マンゴー”, “きゅうり”, “朝顔”, “涼しい”...
- 負の相関を持つ言葉
 - e.g, “暖かい”, “寒い”, “セーター”, “ヒーター”, “暖房”, “手袋”, “ぬくぬく”, “風邪”, “スノボ”, “ゲレンデ”, “雪山”, “真冬”, “寒空”....

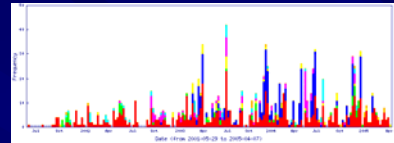
まとめと今後の課題

- Weblog記事を用いた言語横断的関心分析研究の現状について述べた.
- 今後の課題
 - 対象言語の拡張
 - 露, 仏, 伊, 独, 西, 葡...
 - 各言語コミュニティにおける感情の把握
 - “核実験”について各国の人々はどう感じていたか？
 - 社会問題に対する関心の言語横断的な比較

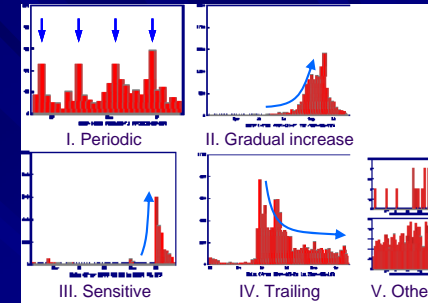
まとめ：関心に対する視点



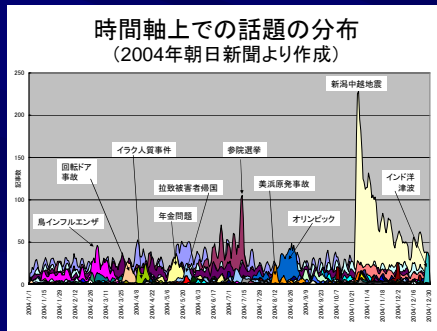
感情に注目した分析



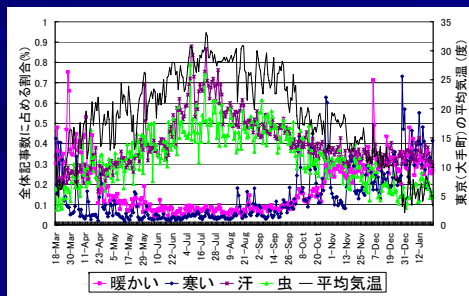
個人(キーパーソン)の関心



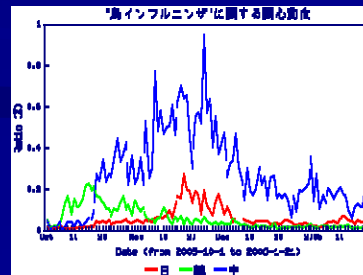
社会の関心



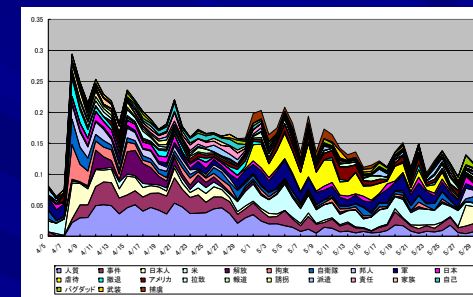
話題の自動検出



実世界データとの関連



海外の関心との比較



共起語を用いた
話題の焦点に関する分析

TODO

- 新聞, テレビ, 雑誌との比較
- 個人属性との関係
- (年代, 性別, 住所...)
- コミュニティ単位での分析