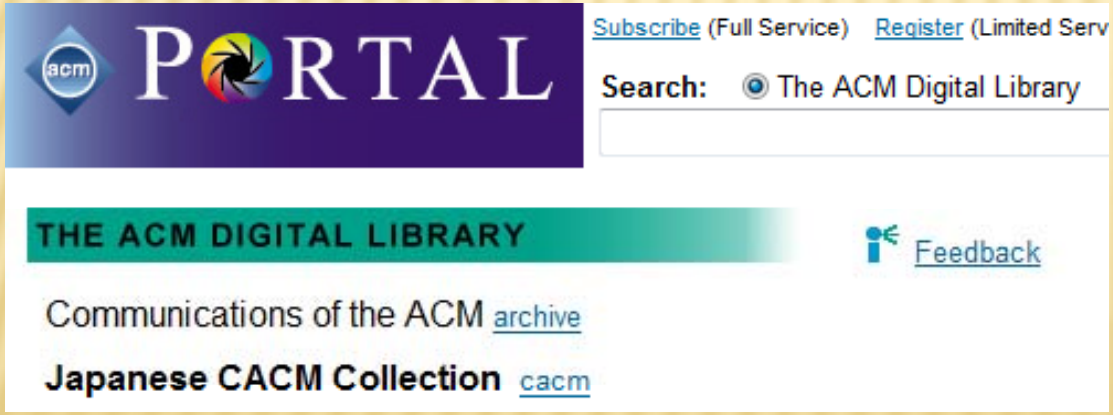
検索エンジンの「今」とその信頼性
ー未来へ向けてー

早稲田大学　理工学術院
情報理工学科　山名早人
http://www.yama.info.waseda.ac.jp/~yamana/
2011/01/17

WHO AM I?

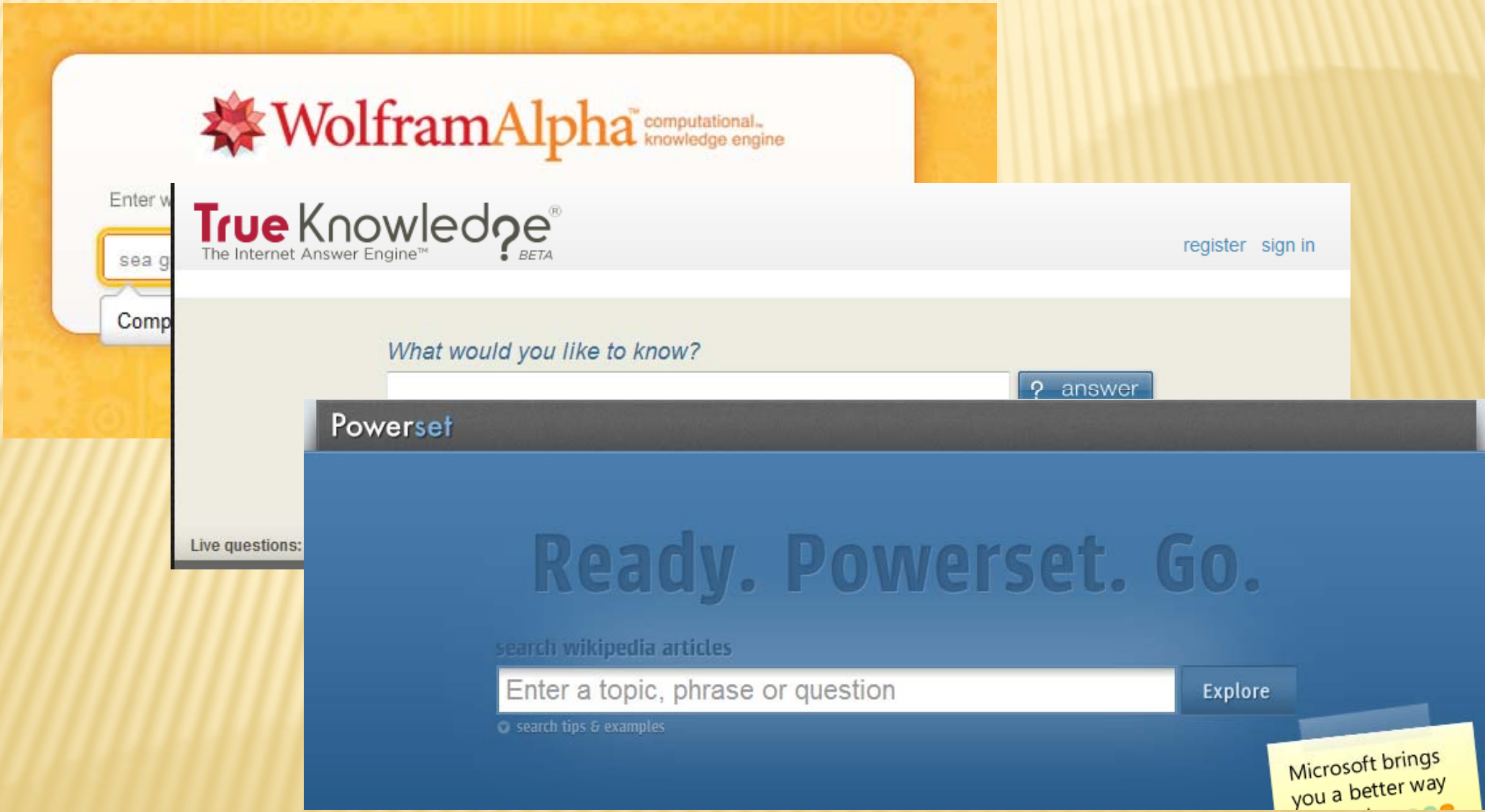
- 山名早人
  - 監訳：Google Hacks, Optimizing Web site等
  - 情報処理学会データベースシステム研究会主査
  - 電子情報通信学会和文Ｄ論文誌編集副委員長
  - CACM日本語版編集委員長
  - 元IEEE Computer Society Japan Chapterチェア



こんな経験はありませんか？

そして、こんな経験も？

新しい「検索エンジン」？



AGENDA

1. 日本における検索エンジンの歴史
2. WEBの規模と現状
3. 検索エンジンの信頼性ーランキングー
4. 検索エンジンの信頼性ー検索結果数ー
5. 新しい検索エンジンと未来

1. 日本における検索エンジンの歴史

日本における検索エンジンの歴史（１）

1994.12 千里眼 *日本初の検索エンジン* ～1999.3 by 早稲田大学(田村氏)

1995.8 NETPLAZA ～2001.1 by 日本電気

1995.9 Info Navigator ～1999.11 by 富士通

1995.10 HOLE-IN-ONE ～2000.5 by 日立国際ビジネス

1995.11 TITAN Internet Information Navigation System by NTT

1996.1 ODiN A Happy New Year Open Documentary Information Navigator by 東京大学 (原田氏)

1996.4 RCAAU Mo-n-do-u(問答) by 京都大学(河野先生)

1996.10

1997.3

1997.7

1998.6

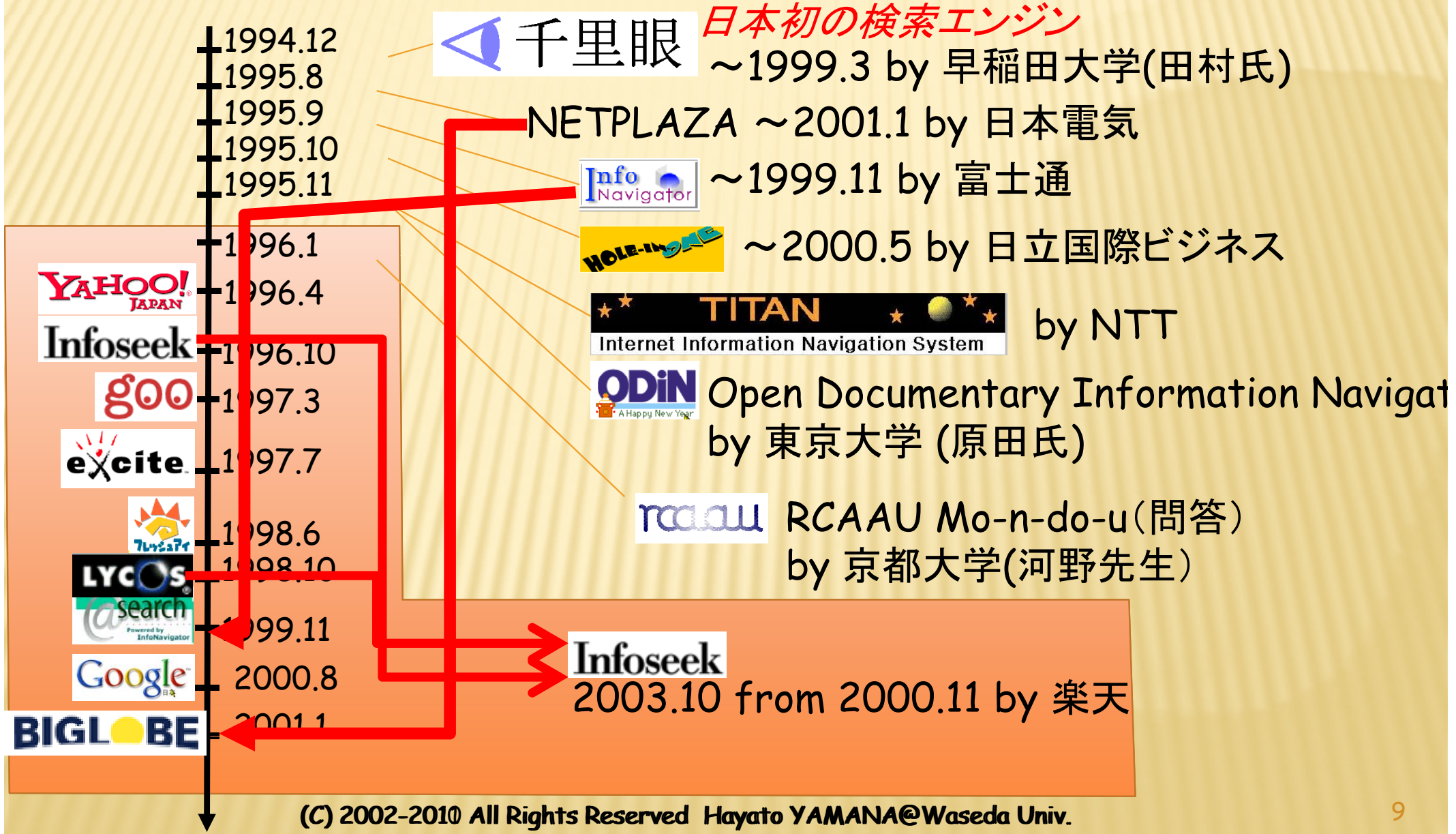
1998.10

1999.11

2000.8

2001.1

日本における検索エンジンの歴史（２）



主要な検索エンジンのバックエンド

- YAHOO! JAPAN
  - 1996.4 powered by Original
  - 1998.5 powered by goo
  - 2001.4 powered by Google
  - 2004.5 powered by YST
- NETPLAZA, BIGLOBE (NEC)
  - 1995.9 powered by Original
  - 2000.11 powered by Google
- InfoNavi/@nifty (Fujitsu)
  - 1995.10 powered by Original
  - 2001.4 powered by Google
  - 2009.夏 powered by YST
- goo (NTT)
  - 1997.3 powered by Original
  - 2003.12 Google
- msn
  - Original
  - 2009.5 bing Beta へ

- Infoseek Japan
  - 1996.10 powered by Original
  - 2003.9 powered Google
- Excite Japan
  - 1997.7 powered by Original
  - 2002.1 powered Google
- Baidu 百度
  - 2008.1 powered by Original
- NAVER BETA
  - 2009.6 powered by Original

日本オリジナルな検索
エンジンは消滅
そして、
Google/Yahoo!+Bing/
Baidu/NAVERの4強へ

検索エンジン創世記の大学における主な研究者たち

千里眼

apan)

Google

ODiN
A Happy New Year

oratorie

n)

Google

UEDA lab. @ Waseda Univ.
Verno

Google

南山大学

ほとんどの研究者はGoogleへ

2. WEBの規模と現状

世界に[illegible]？

- 1997年[illegible]
  - 大規[illegible]なりから統計[illegible]
- 1999年[illegible]
  - 360[illegible]2500のWeb[illegible]的に算出。
- 2011年 1月：1448億と推定
  - 平均530ページ/サーバ（2004-2005年収集の85億ページの平均）
  - 530page×27,330万Webサイト[3]≒1448億

January 2011 Web Server Survey

In the **January 2011** survey we received responses from **273,301,445** sites.

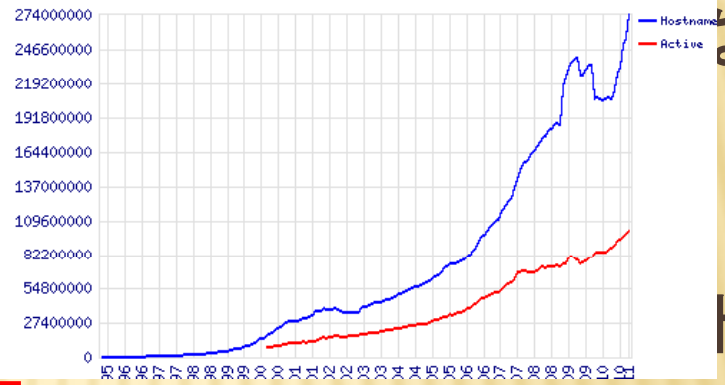
nginx once again saw the largest increase in market share amongst the top web servers, gaining 0.88 percentage points or 3.59M hostnames. nginx no top 1 million sites. 2.4M of the additional nginx hostnames are hosted by Ecatel, which appears to have migrated from Apache to nginx this month. ngi nginx with a 8.23% share.

Lighttpd was the only other web server to increase its share of hostnames this month, increasing by 558k hostnames, 531k of which are hosted by Secu

Apache remains the dominant force, commanding a 59.13% share and an increase of 10.1M hostnames. The majority of the Apache growth was in the U the United States is concentrated in AmeriNOC and Softlayer, together accounting for 7.34M hostnames. Germany's growth, on the other hand, is co holding pages for fhe3rz.net.

Microsoft made modest gains this month, adding 669k hostnames despite losing 1.2 percentage points in market share.

Total Sites Across All Domains
August 1995 - January 2011



[1]S.Lawrence, C.L.Giles:"Searching the World Wide Web", Science, Vol.280, No.5360, pp.98-100 (1998)
[2] S.Lawrence, C.L.Giles:"Accessibility of Information on the Web", Nature, Vol.400, pp.107-109 (1999)
[3] -: Netcraft Home Page, http://www.netcraft.co.uk/

CUILのインデックス数

GOOGLEの2009年8時点のGOOGLE

- Googleはどのぐらいの情報を持つか？
  - 「site:トップレベルドメイン名」で検索結果数を調査
  - インターネットに接続する２６９トップレベルドメインを調査→約400億ページ
  - この数字は2006年時とほとんど変化せず

TechCrunch

What's Hot: Android | Apple | Facebook | Google | Groupon | Microsoft | Zynga | Classics | CES 20

Highlights and Videos from CrunchGear's CES coverage »

Cuil Fails to Be Acquired

Michael Arrington

いいね！ Buzz 176 Tweet 348 16 Digg

Sep 19, 2010

63 Comments

As we reported last week, search engine **Cuil** was **unceremoniously shut down** on Thursday, and there were reports that employees were told to go home and forget about getting paid.

New sources tells us that Cuil was in the final stages of an acquisition as of last Wednesday, and everything was in place except the final signatures. Then the deal fell apart for some reason.

Or put another way, Cuil found one last way to fail.

There are certain assets, particularly algorithms and patents, that may have some value to certain companies, we've heard from one of our sources.

A complication may have been over employees, which were supposed to go with the deal and be taken care of by the buyer.

Regardless, our understanding is that Cuil is trying to regroup and get the site back live, and another deal, or the old deal, may be closed soon.

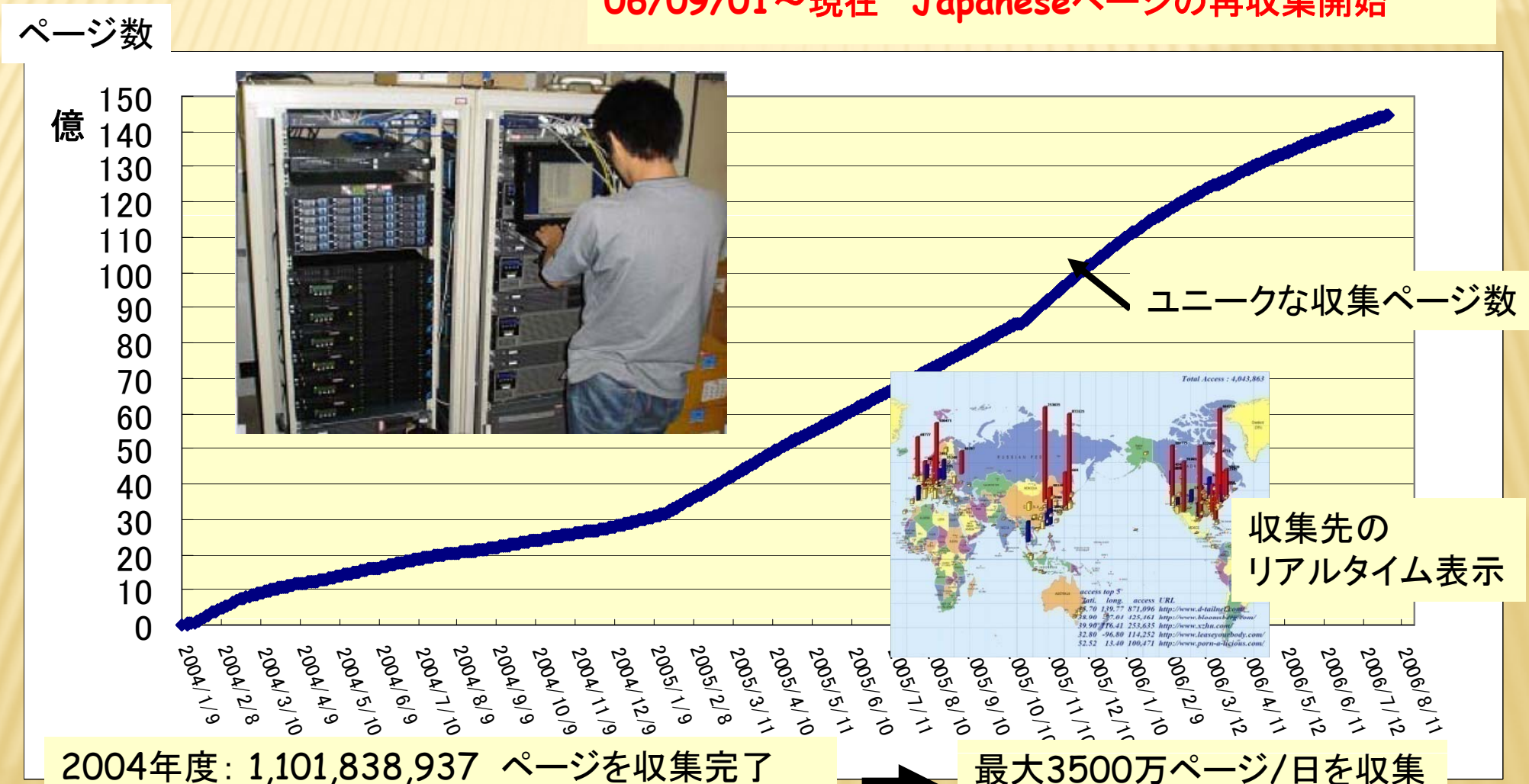
Either way, at best it's a soft landing. More details as we gather them. There are only a very few buyers who'd have much interest in Cuils assets – particularly Google and Microsoft.

http://techcrunch.com/2010/09/19/cuils-fails-acquired/

実績- WEBデータ収集

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

04/01/19　3拠点(早大,NTT,IDC)で収集開始[30CPU]
05/01/17　2拠点追加(早大,NII) [合計50CPU]
05/10/21　3拠点にマシン追加[合計80CPU]
06/09/01～現在　Japaneseページの再収集開始



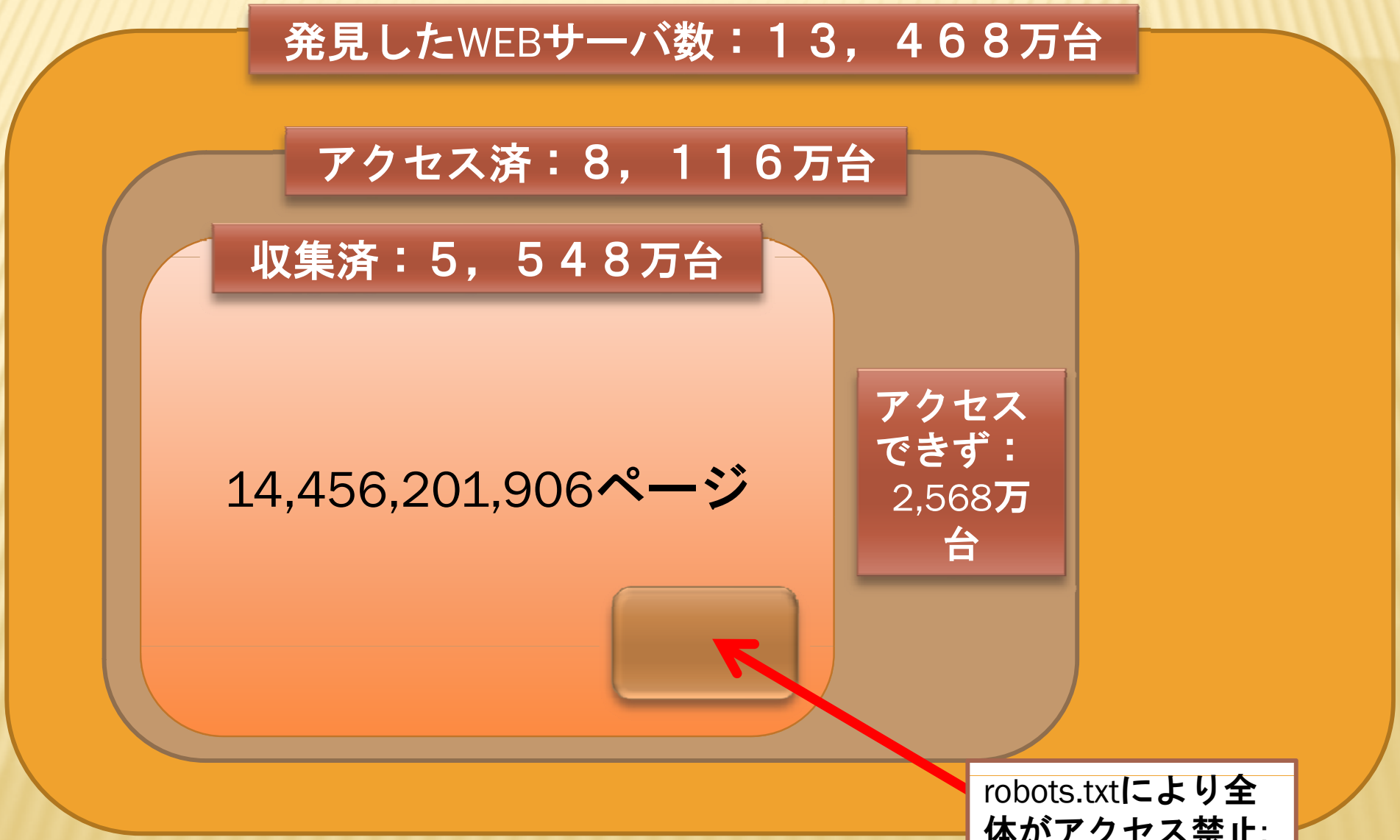
2004年度：1,101,838,937　ページを収集完了
2005年度:12,669,681,455 ページを収集完了
2006/7末:14,456,201,906　ページを収集完了

→ 最大3500万ページ/日を収集
(平均約1000万ページ/日)

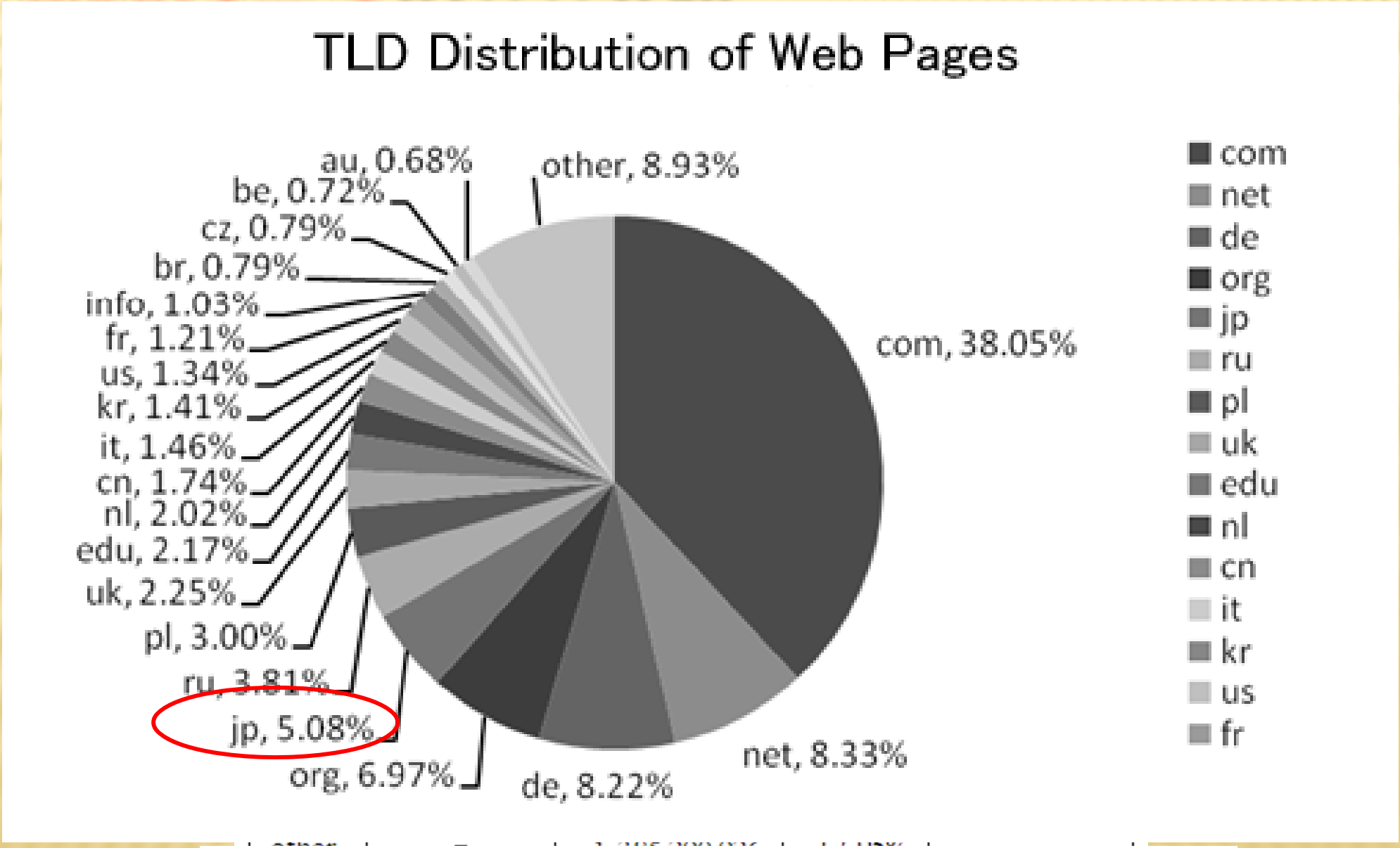
起点と収集方針

- 起点
  - 約600万のWebサーバリスト
    - 2004.1以前に我々が持つ起点リストを利用
      - 1998-2000　分散収集実験他
- 収集方針
  - 起点から最大15ホップ先までを収集
  - 収集間隔は1秒（順次5秒,15秒と変更）
  - テキストのみを収集（バイナリは拡張子で排除）
  - 2005.1以降　6時間収集５時間サスペンド
  - 2005.7以降,　Webサーバへの負荷軽減対策

収集済WEBサーバ数

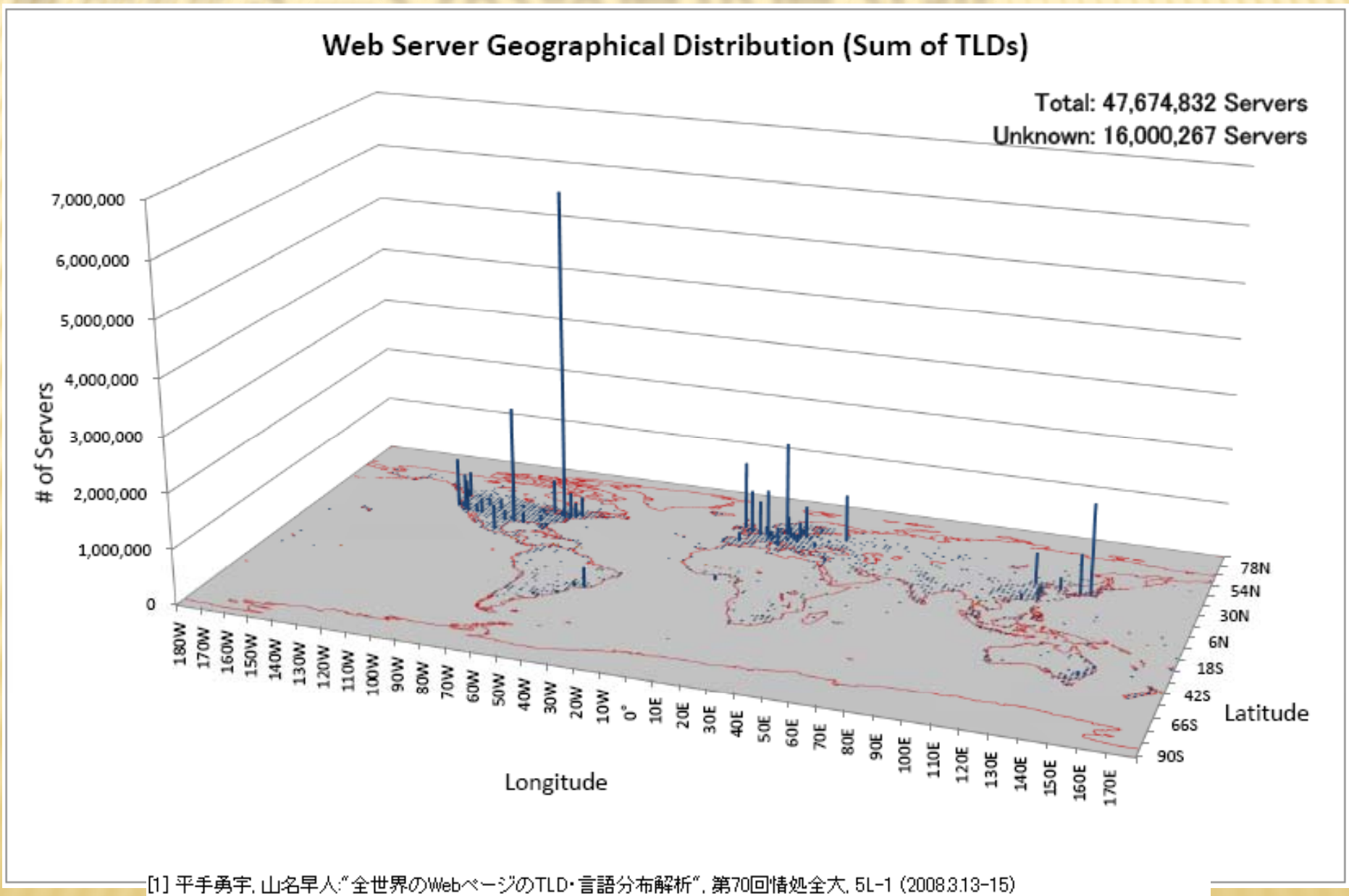


WEBページのTLD分布



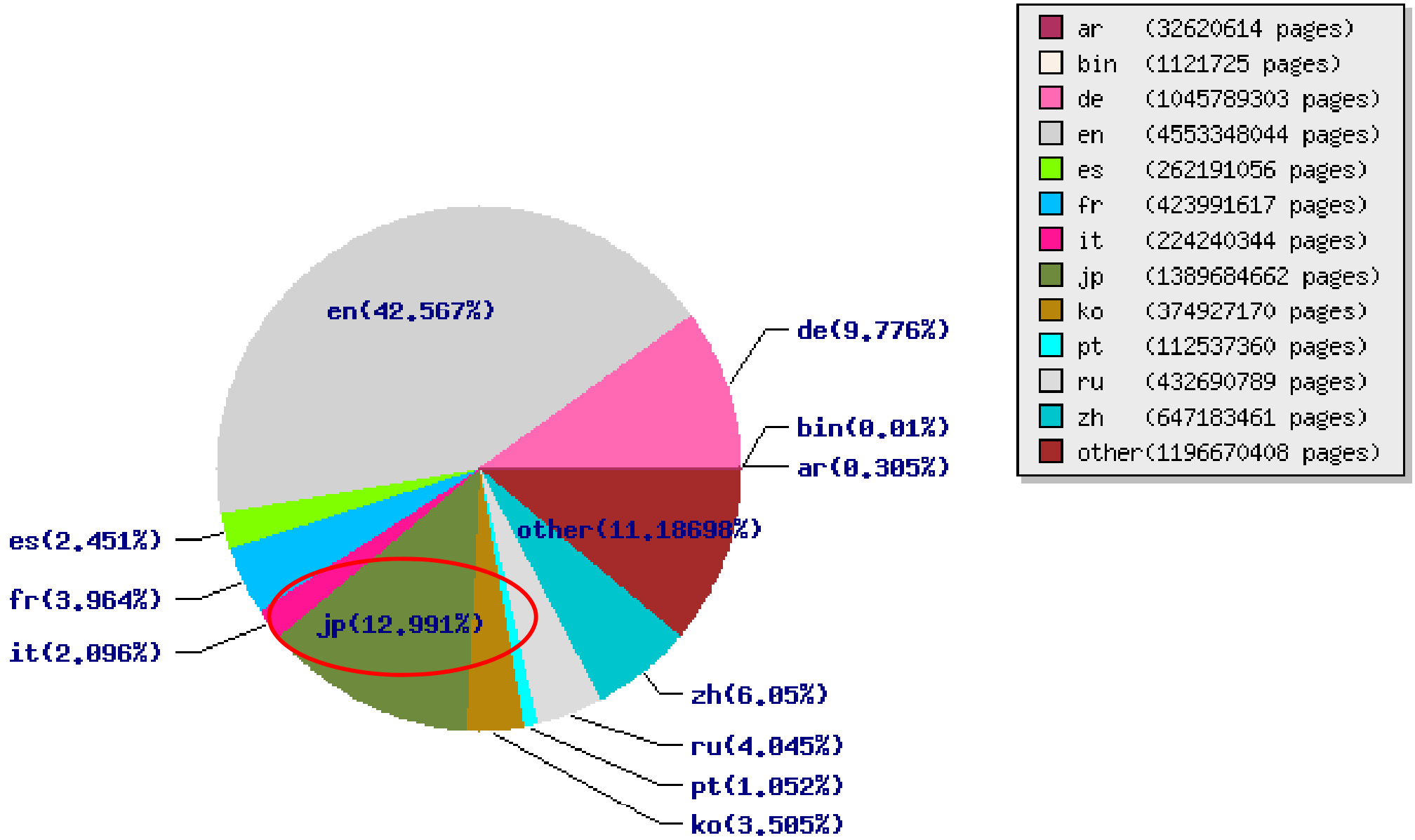
[1] 平手勇宇, 山名早人:"全世界のWebページのTLD・言語分布解析", 第70回情処全大, 5L-1 (2008.3.13-15)
[2] 童　芳, 平手勇宇, 山名早人:"全世界のWebサイトの言語分布と日本語を含むWebサイトのリンク・地理的位置の解析", DEWS2008, A2-3 (2008.3.10-12)

全WEBサーバの設置位置分布



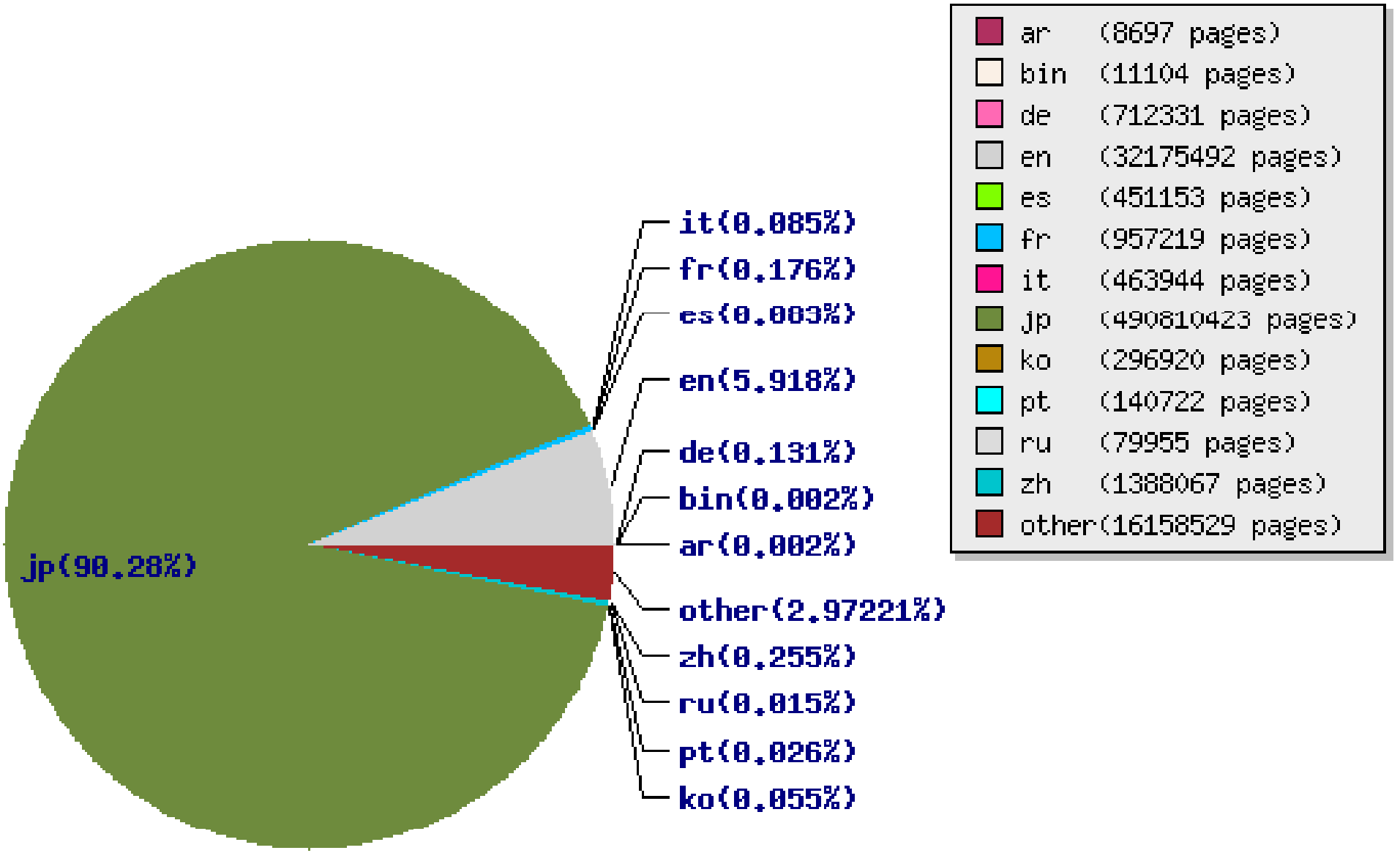
[1] 平手勇宇, 山名早人:"全世界のWebページのTLD・言語分布解析", 第70回情処全大, 5L-1 (2008.3.13-15)
[2] 童　芳, 平手勇宇, 山名早人:"全世界のWebサイトの言語分布と日本語を含むWebサイトのリンク・地理的位置の解析", DEWS2008, A2-3 (2008.3.10-12)

WEBページの言語分布(107億ページ)



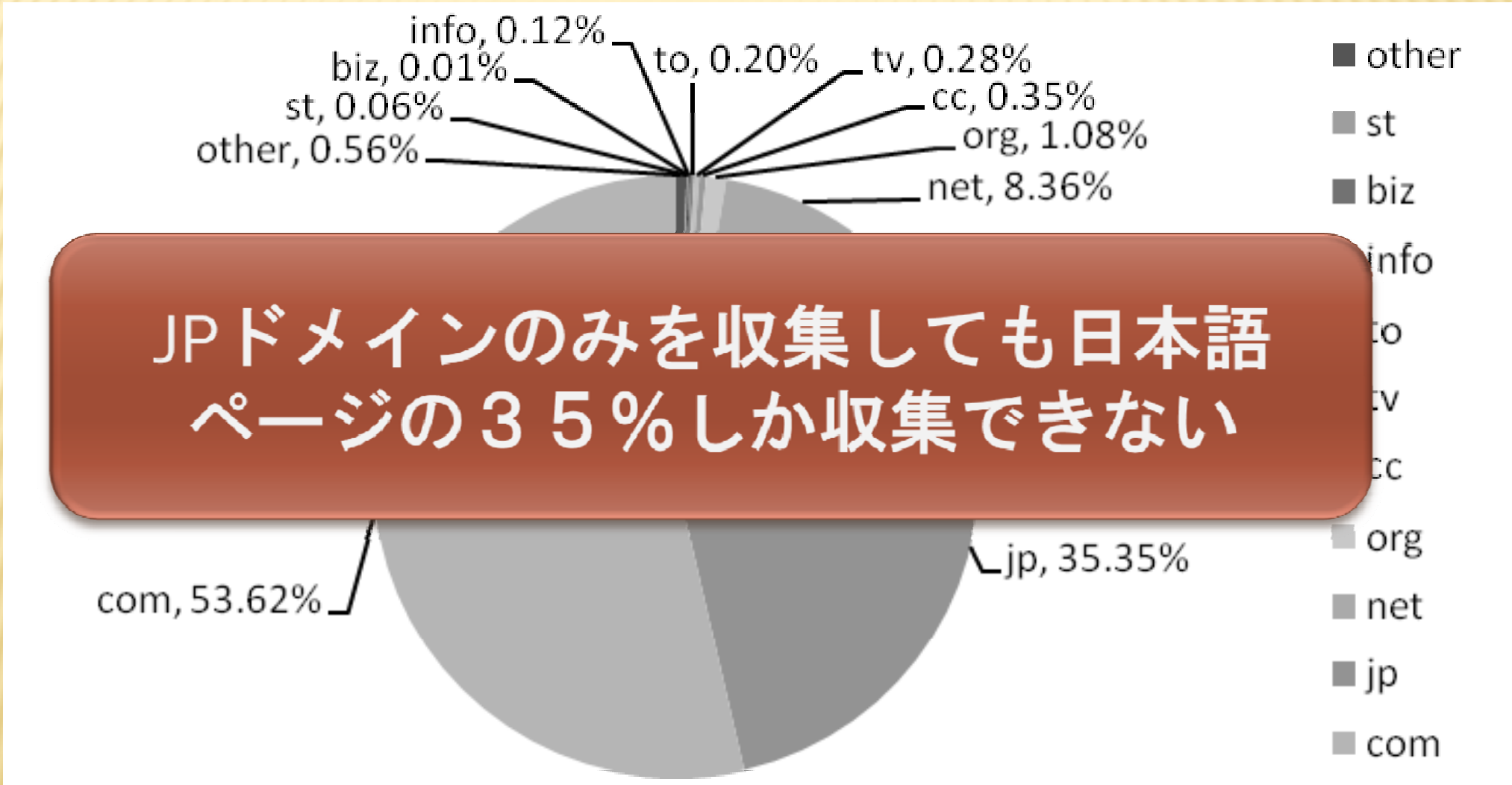
[1] 平手勇宇, 山名早人:"全世界のWebページのTLD・言語分布解析", 第70回情処全大, 5L-1 (2008.3.13-15)
[2] 童　芳, 平手勇宇, 山名早人:"全世界のWebサイトの言語分布と日本語を含むWebサイトのリンク・地理的位置の解析", DEWS2008, A2-3 (2008.3.10-12)

JPドメインに属する約5.4億ページの言語分布



[1] 平手勇宇, 山名早人:"全世界のWebページのTLD・言語分布解析", 第70回情処全大, 5L-1 (2008.3.13-15)
[2] 童　芳, 平手勇宇, 山名早人:"全世界のWebサイトの言語分布と日本語を含むWebサイトのリンク・地理的位置の解析", DEWS2008, A2-3 (2008.3.10-12)

日本語で書かれたページのTLD分布



JPドメインのみを収集しても日本語ページの３５％しか収集できない

[1] 平手勇宇, 山名早人:"全世界のWebページのTLD・言語分布解析", 第70回情処全大, 5L-1 (2008.3.13-15)
[2] 童　芳, 平手勇宇, 山名早人:"全世界のWebサイトの言語分布と日本語を含むWebサイトのリンク・地理的位置の解析", DEWS2008, A2-3 (2008.3.10-12)

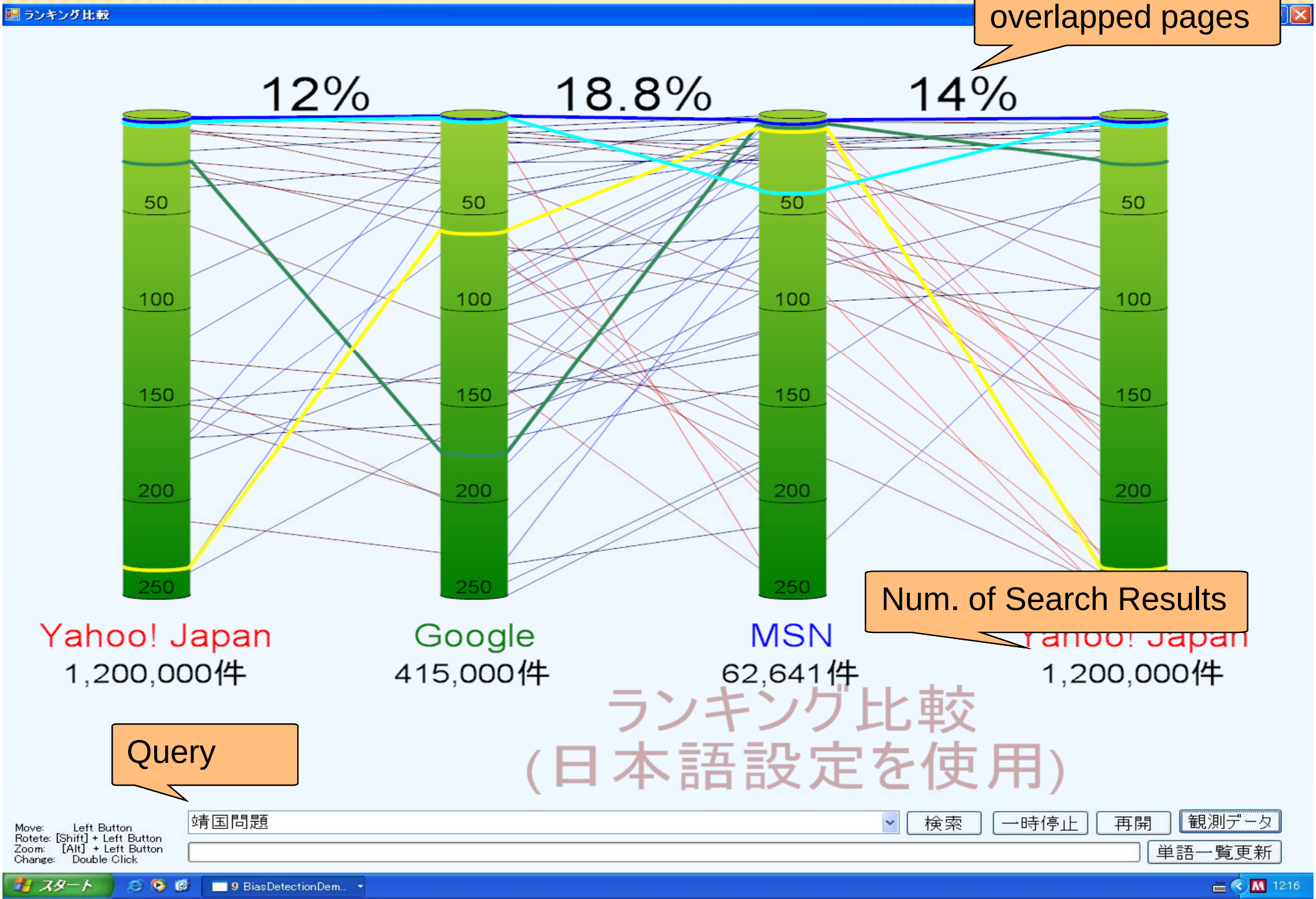
3. 検索エンジンの信頼性
ーランキングー

山名早人: "検索エンジンの信頼性",人工知能学会誌, Vol.23, No.6, pp.752-759 (2008.11)

吉田泰明, 平手勇宇, 山名早人:"商用検索エンジンにランキングされたサイトのランク変動パターンの解析", DEIM2009 (2009.3.8-10)

Yasuaki Yoshida(M1), Takanori Ueda, Takashi Tashiro, Yu Hirate, Hayato Yamana:"What's going on in search engine rankings?", Proc. of the 2008 IEEE International Symposium on Mining and Web (2008.3.25-28)

検索エンジン間のランキング比較



マルチブラウザ

Q－08、靖国問題

Q―08、靖国神社は軍国主義の象徴なのですか？

A―08、違います。どこの国でも国のために戦死した人を祀るのは当然のことです。

国を守るためにまたは国のために戦死した人の慰霊なくして永続的な国の繁栄と発展はありません。それは思想や宗教には関係ないものです。戦前の戦死者の慰霊は、たいてい地元の護国神社でやりその後に檀家寺でやり、さらに東京の靖国神社でもまつられたわけです。それが普通だったのです。つまり戦前の戦死者のほとんどは靖国神社にまつられることを当然と思っていたのですから、私たちはそれを尊重しなければなりませんし、国の指導者が靖国神社に参拝するのも自然なことです。戦後の一部の人の思想によって政府や裁判所がそれを勝手に変えてはいけません。

【参考】
昭和19年10月26日にフィリッピンのセブ島の基地から特攻に志願して戦死した植村大尉の愛児に遺した手紙です。

「･･･私はお前が大きくなつて、立派な花嫁さんになつて、仕合わせになつたのを見届けたいのですが、若しお前が私を見知らぬまゝ死んでしまつても、決して悲しんではなりませ

MIYADAI.com Blog

< 2006-10 >

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Bibliography

MIYADAI.com Blog (Archive) > 靖国問題で即席のコメントを求められました。あくまで即席回答です
«『サイファ 覚醒せよ』文庫版あとがき:「実存的な曖昧さへ」 | 8月13日の思想塾公開イベントは申し込み受付けを終了しました »

靖国問題で即席のコメントを求められました。あくまで即席回答です

投稿者:miyada
投稿日時:2006-08-04 - 13:18:46
カテゴリー:お仕事で書いた文章 - トラックバック(3)

■靖国問題は「東京裁判的手打ち問題」(略して手打ち問題)と「憲法と歴史の捩れ問題」(略して憲法問題)と「天皇陛下の御意思問題」(略して天皇問題)とに大まかに整理できるだろう。整理できると言っても、この三つの問題は複雑に分岐しつつ、絡まり合う。
■第一に、東京裁判的手打ちとは、戦後復興と国際社会復帰のための国際協力を獲得するべく、A級戦犯に戦争責任を帰責することで天皇と国民から免罪する「虚構」のこと。東京裁判は「虚構」のための道具だから不公正は当たり前。この不公正で我々は免罪された。
■この手打ち問題については、東京裁判が公正だとする左は、「虚構」による免罪と引替に犠牲になった者への道義的責任を忘却する点、批判に値する。不公正だとする右は、「虚構」による免罪で戦後社会を自らがぬくぬく生き延びた事実を忘却する点、批判に値する。

靖国

日本共産党
Japanese Communist Party

メールはこちら　サイトマップ

戦後60年

首相の靖国神社参拝

21世紀に、日本がアジアの一員と
生きていけるかどうかの大問題で

小泉首相
志位委員
「考え違う
らの参拝
道理がた
10月17日）

衆院予算委

過去の戦争から教訓をまなんでこそ平和な未来がきずける――戦後60年を迎え、世界は新しい歩みをはじめています。
ところが私たちの国は、首相の靖国神社参拝をめぐってアジア各国からの強い批判にさらされ、外交がゆきづまっています。国内でも、参拝中止をもとめ

靖国問題で即席のコメントを求められました。あくまで即席回答です - MIYADAIcom Blog

ログインまたはアカウント作成

本文　ノート　編集　履歴

ウィキペディア
フリー百科事典

ナビゲーション
- メインページ
- コミュニティ・ポータル
- 最近の出来事
- 最近更新したページ
- おまかせ表示
- アップロード (ウィキメディア・コモンズ)

ヘルプ
- ヘルプ
- 井戸端
- 連絡先
- バグの報告
- 寄付

靖国神社問題

出典: フリー百科事典『ウィキペディア(Wikipedia)』
（首相、大臣の靖国神社参拝問題 から転送）

中立的な観点:この記事は、中立的な観点に基づく疑問が提出されているか、あるいは議論中です。そのため、偏った観点によって記事が構成されている可能性があります。詳しくは、この記事のノートを参照してください。

提案:このページのノートに、このページに関する提案があります。
提案の要約:節『歴史』の2005年10月1日における一弁護士のコメント掲載について

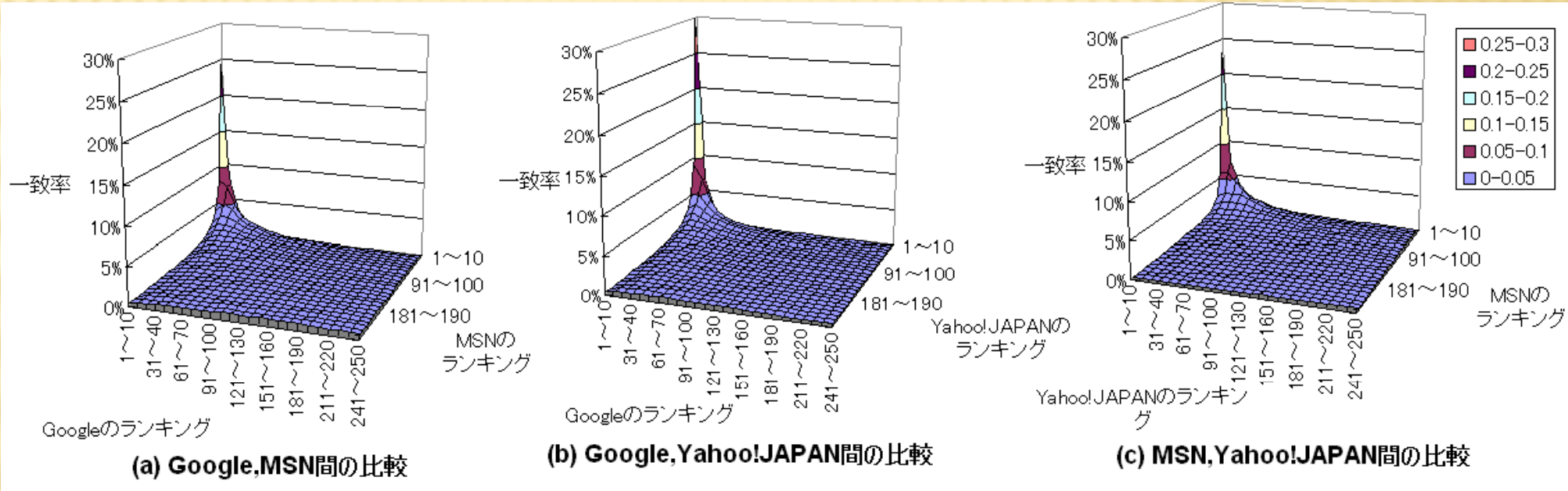
加筆依頼:この項目「**靖国神社問題**」は、**加筆依頼**に出されており、内容をより充実させるために次の点に関する加筆が求められています:中国、韓国以外の国からの

首相の靖国神社参拝問題/JCP特集2005

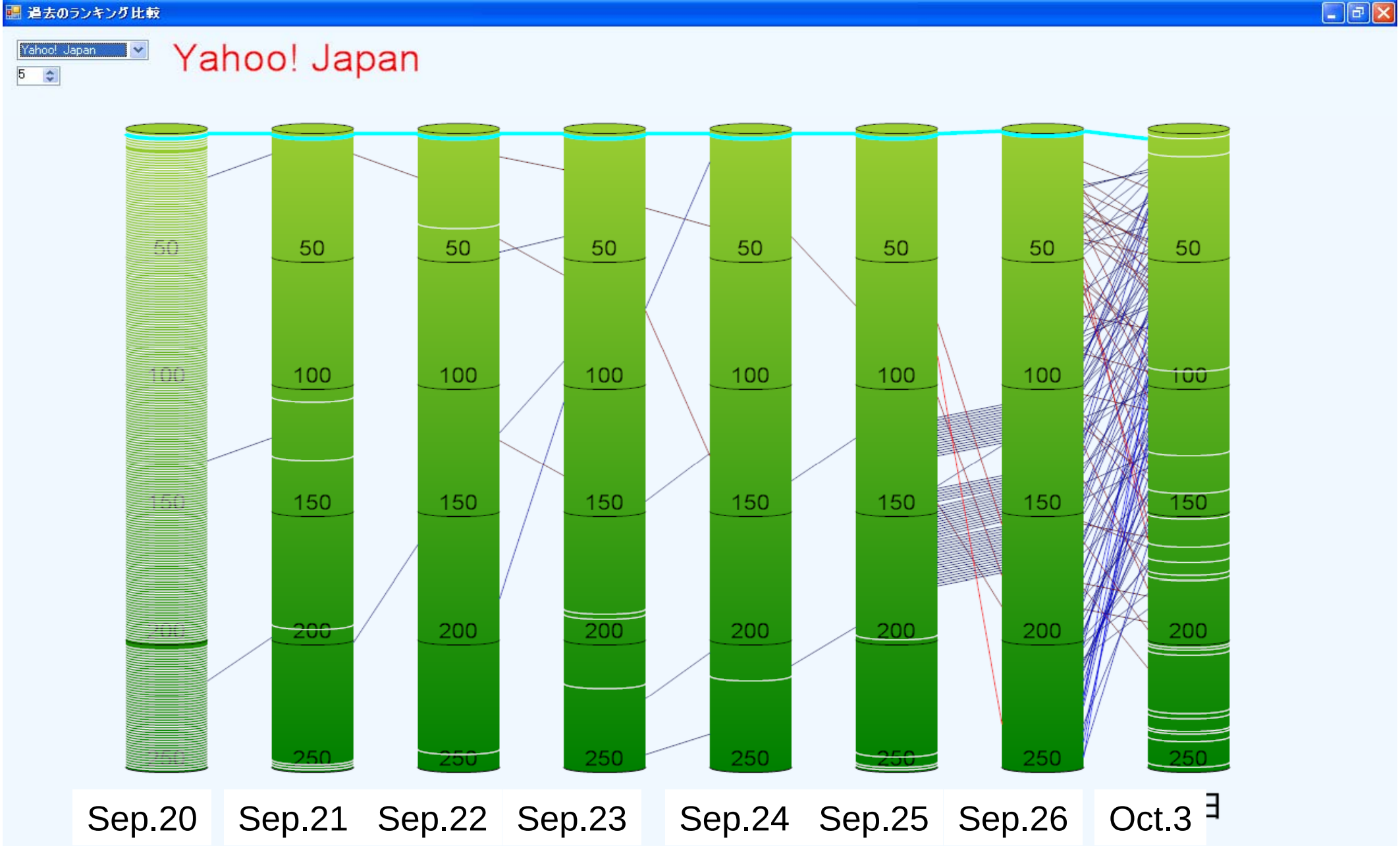
靖国神社問題 - Wikipedia

スタート　9 BiasDetectionDem...　12:20

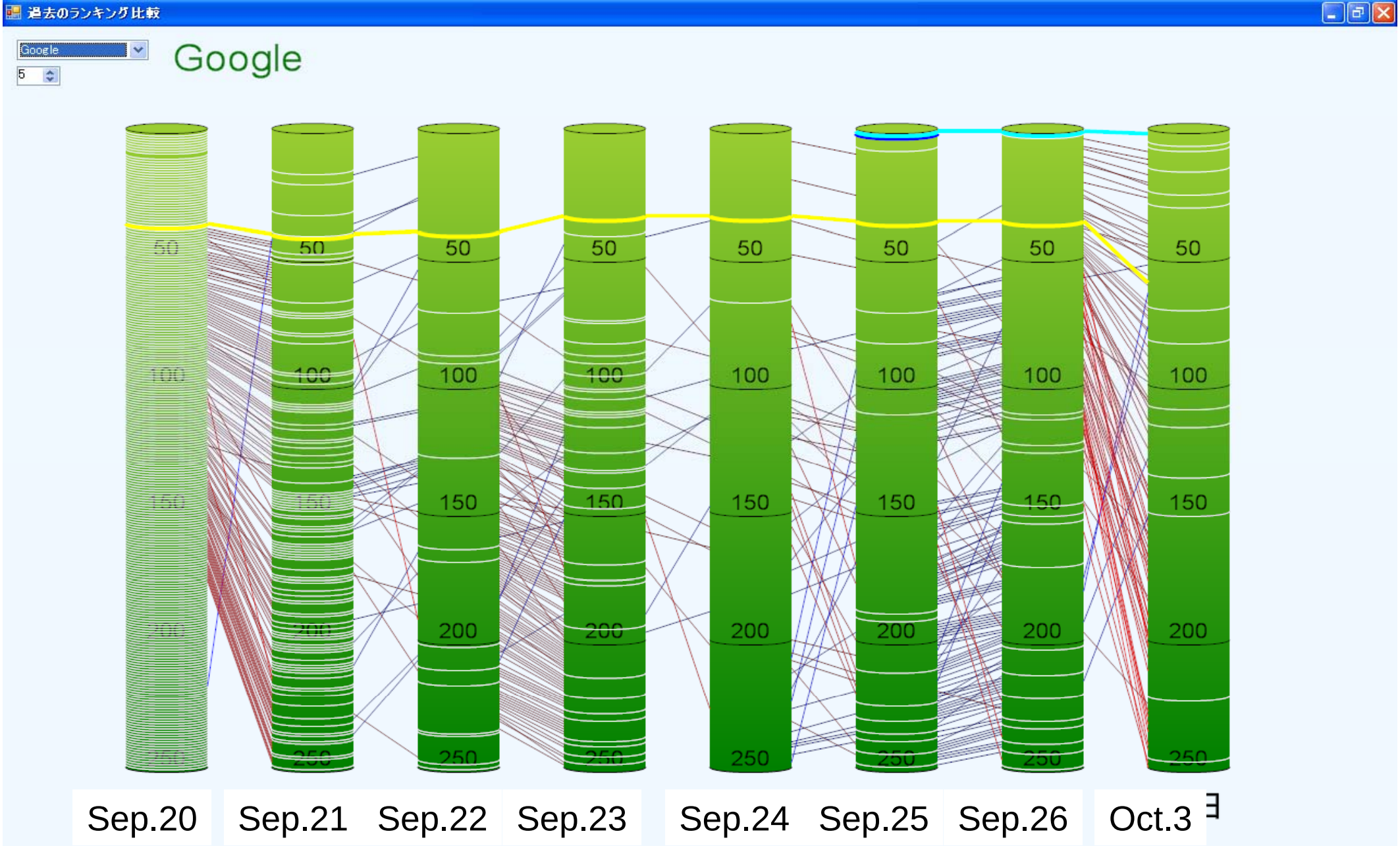
検索エンジン間のランキング一致度



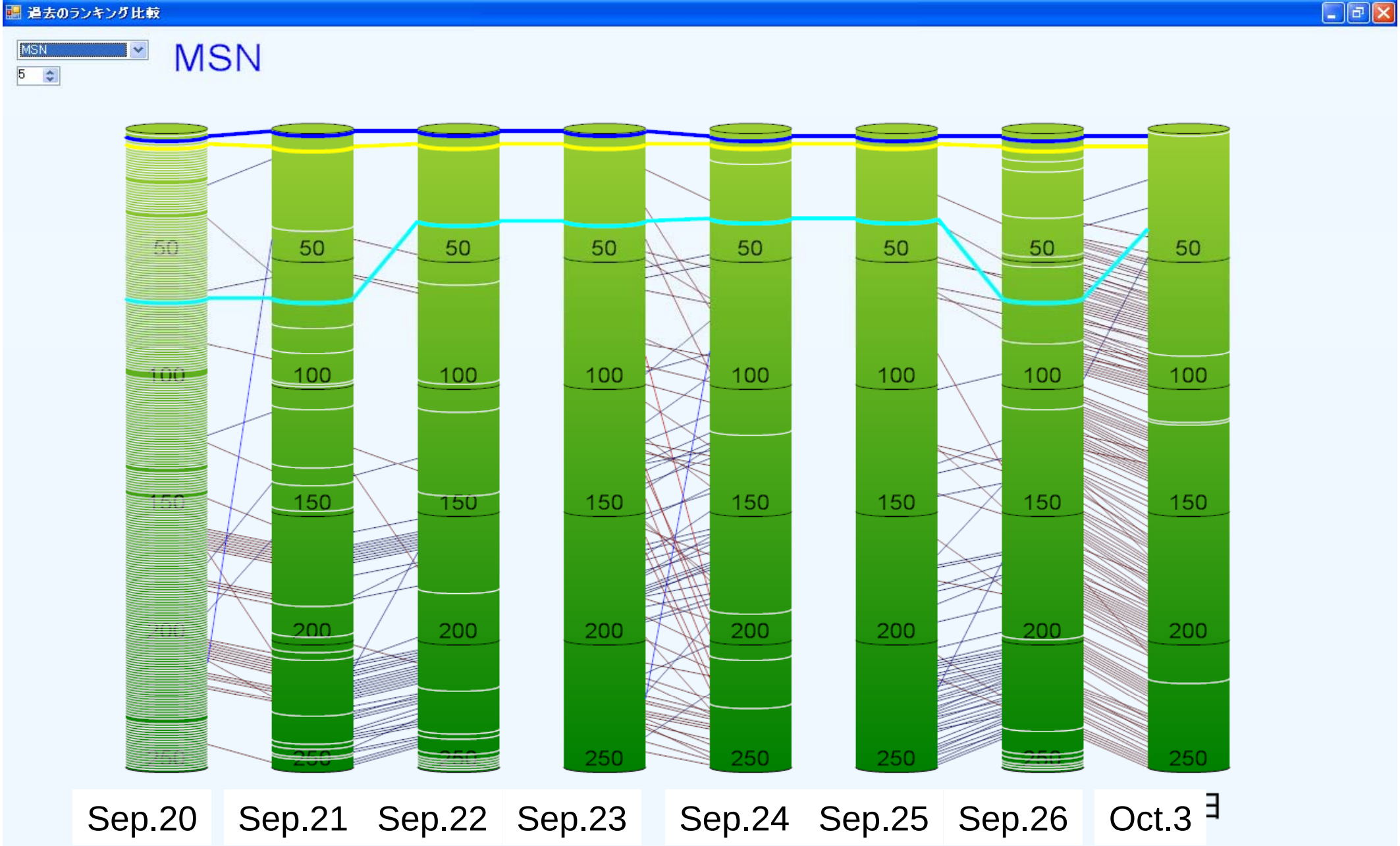
Yasuaki Yoshida(M1), Takanori Ueda, Takashi Tashiro, Yu Hirate, Hayato Yamana:"What's going on in search engine rankings?", Proc. of the 2008 IEEE International Symposium on Mining and Web (2008.3.25-28)



Transition of the ranking



Transition of the ranking



Transition of the ranking

Google における解析グラフ

Google

出現回数

50

Word appearance frequency

Google

出現回数

0

0

ランキング

Yahoo! Japan における解析グラフ

Yahoo! Japan

出現回数

50

Word appearance frequency

Yahoo! Japan

出現回数

0

0

ランキング

MSN における解析グラフ

MSN

出現回数

50

Word appearance frequency

MSN

出現回数

0

0

ランキング

ページ分類

Page Categorization

Google

Yahoo! Japan

MSN

ニュース

ブログ

PDF

動的ページ

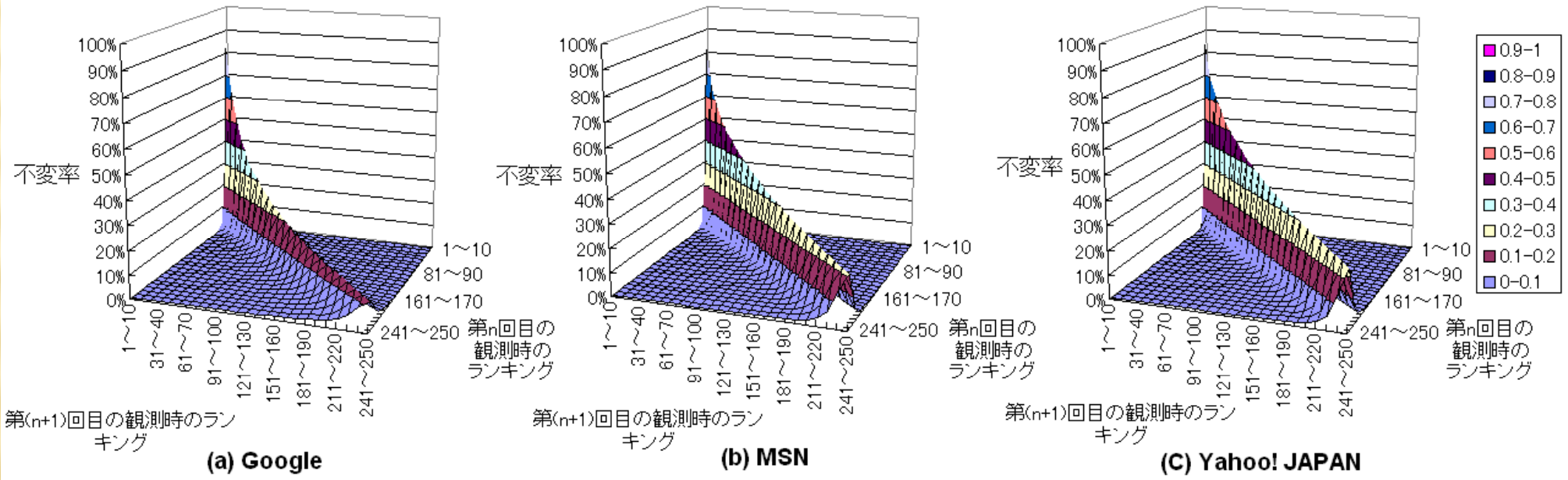
静的ページ

スタート

9 BiasDetectionDem...

12:23

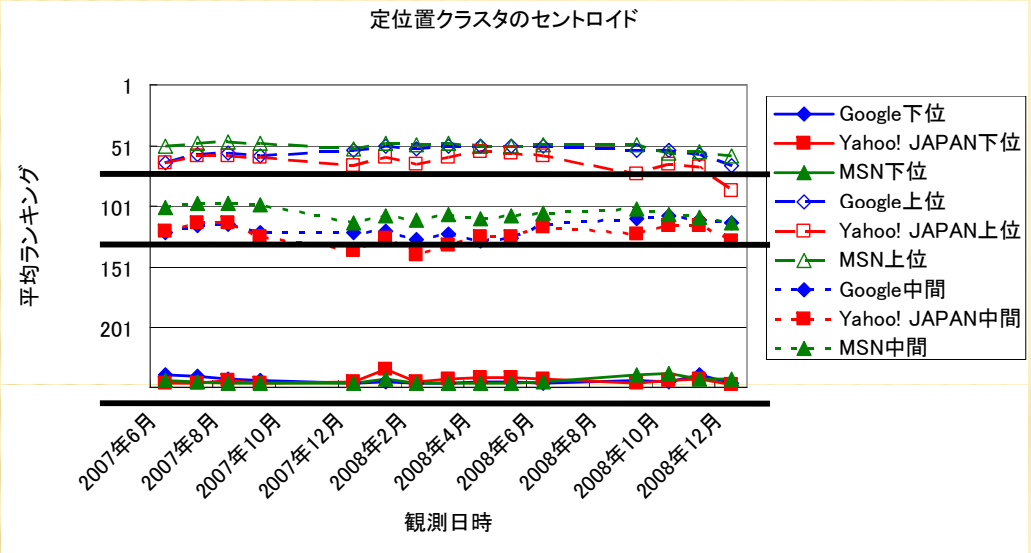
約１０日後のランキング変動



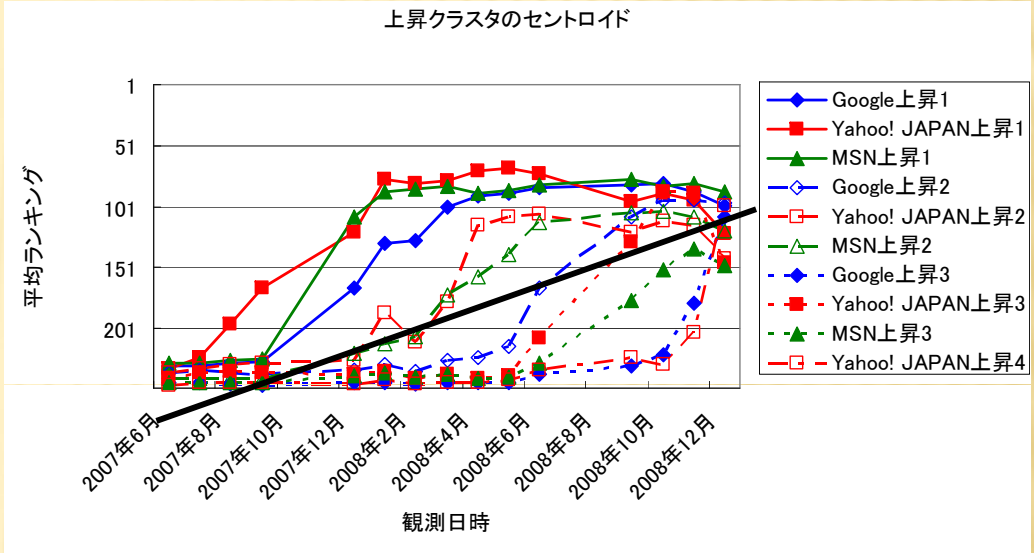
Yasuaki Yoshida(M1), Takanori Ueda, Takashi Tashiro, Yu Hirate, Hayato Yamana:"What's going on in search engine rankings?", Proc. of the 2008 IEEE International Symposium on Mining and Web (2008.3.25-28)

サイトのランク変動パターン

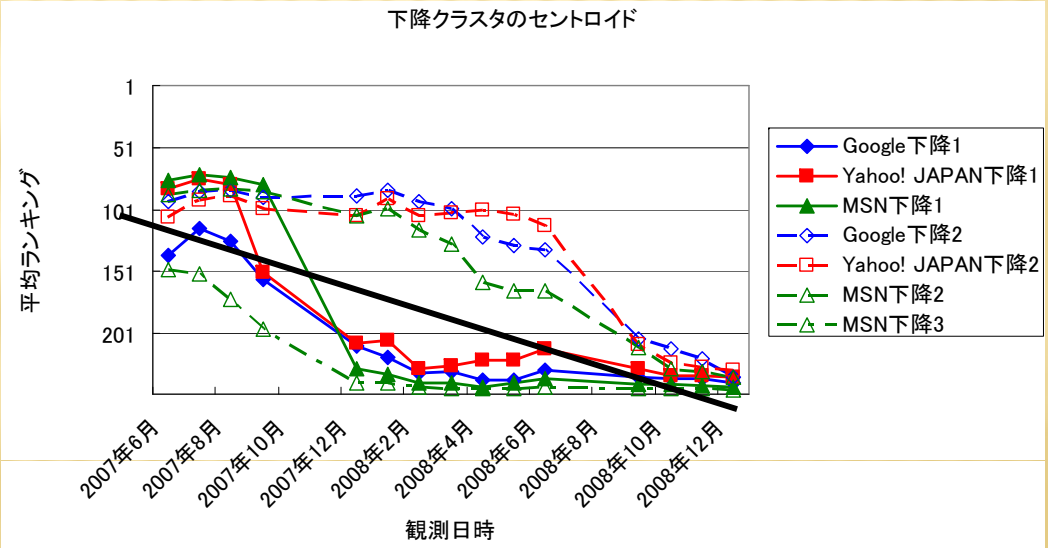
- 定位置パターン



- 上昇パターン



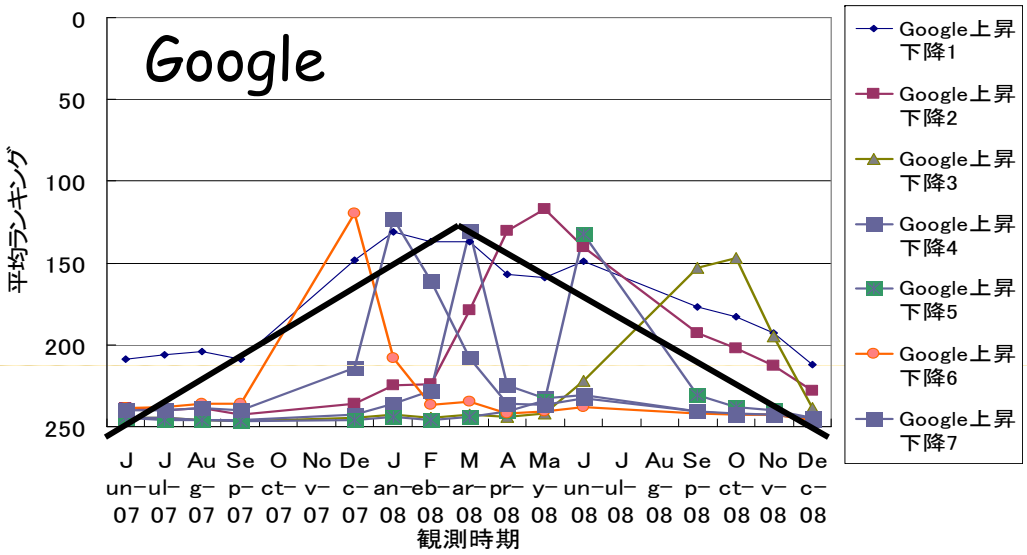
- 下降パターン

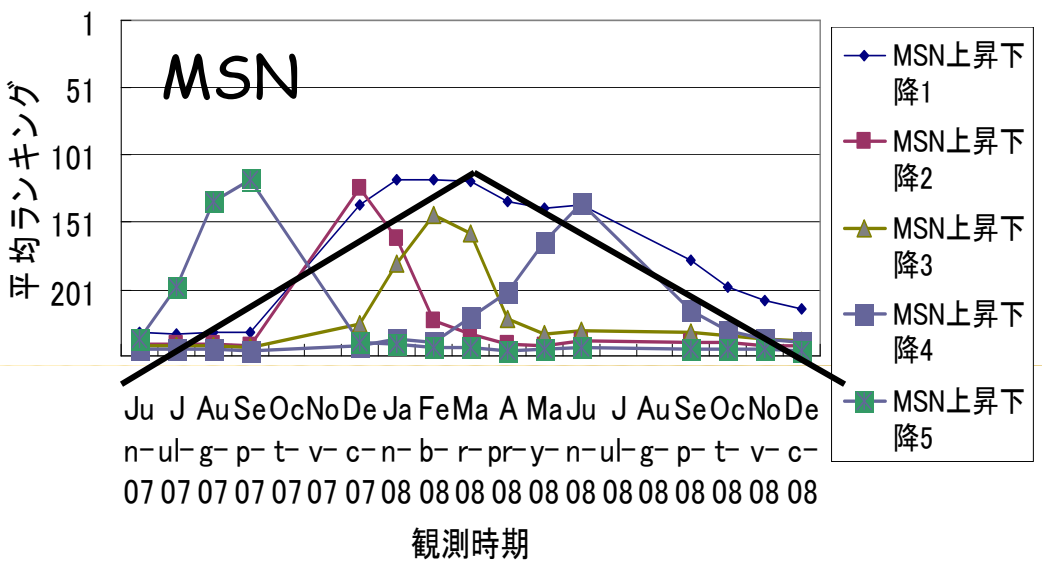


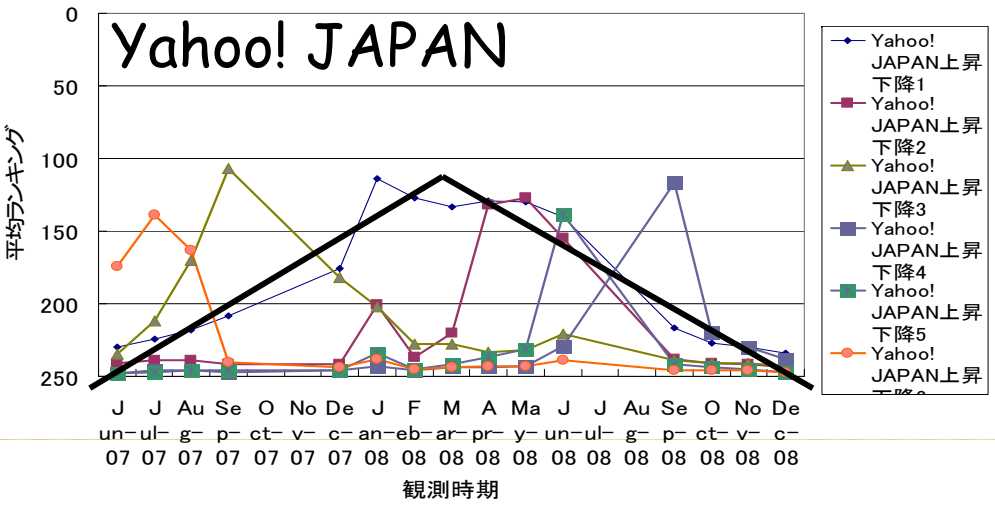
吉田泰明, 平手勇宇, 山名早人:"商用検索エンジンにランキングされたサイトのランク変動パターンの解析", DEIM2009 (2009.3.8-10)

サイトのランク変動パターン

- 上昇下降パターン







下降上昇クラスタは出てこない
→　一度落ちたら上がってこない

吉田泰明, 平手勇宇, 山名早人:"商用検索エンジンにランキングされたサイトのランク変動パターンの解析", DEIM2009 (2009.3.8-10)

4. 検索エンジンの信頼性 ー検索結果数ー

山名早人: "検索エンジンの信頼性",人工知能学会誌, Vol.23, No.6, pp.752-759 (2008.11)

舟橋卓也、曽根広哲、山名早人:"複数キーワードクエリに対する検索ヒット数の信頼性検証 ", 信学技報, Vol.109, No.153, pp.19-24 (2009.7.28)

舟橋卓也, 上田高徳, 平手勇宇, 山名早人: "商用検索エンジンのヒット数に対する信頼性の検証", 日本データベース学会論文誌, Vol.7, No.3, pp.31-36 (2008.12)

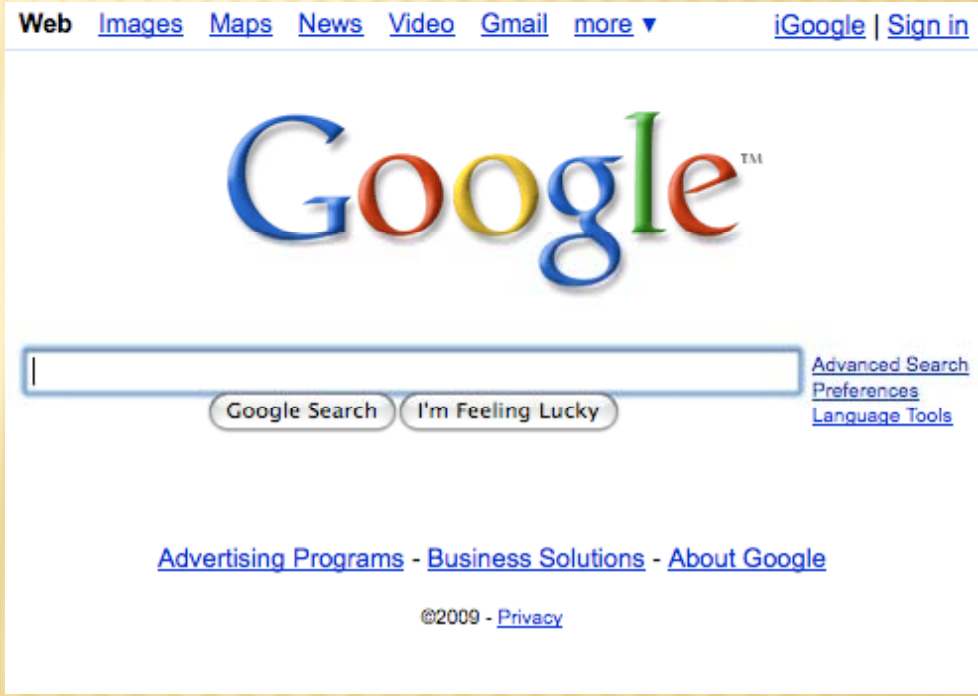
オバマ大統領とGOOGLEはどちらが有名か？

Barack Obama

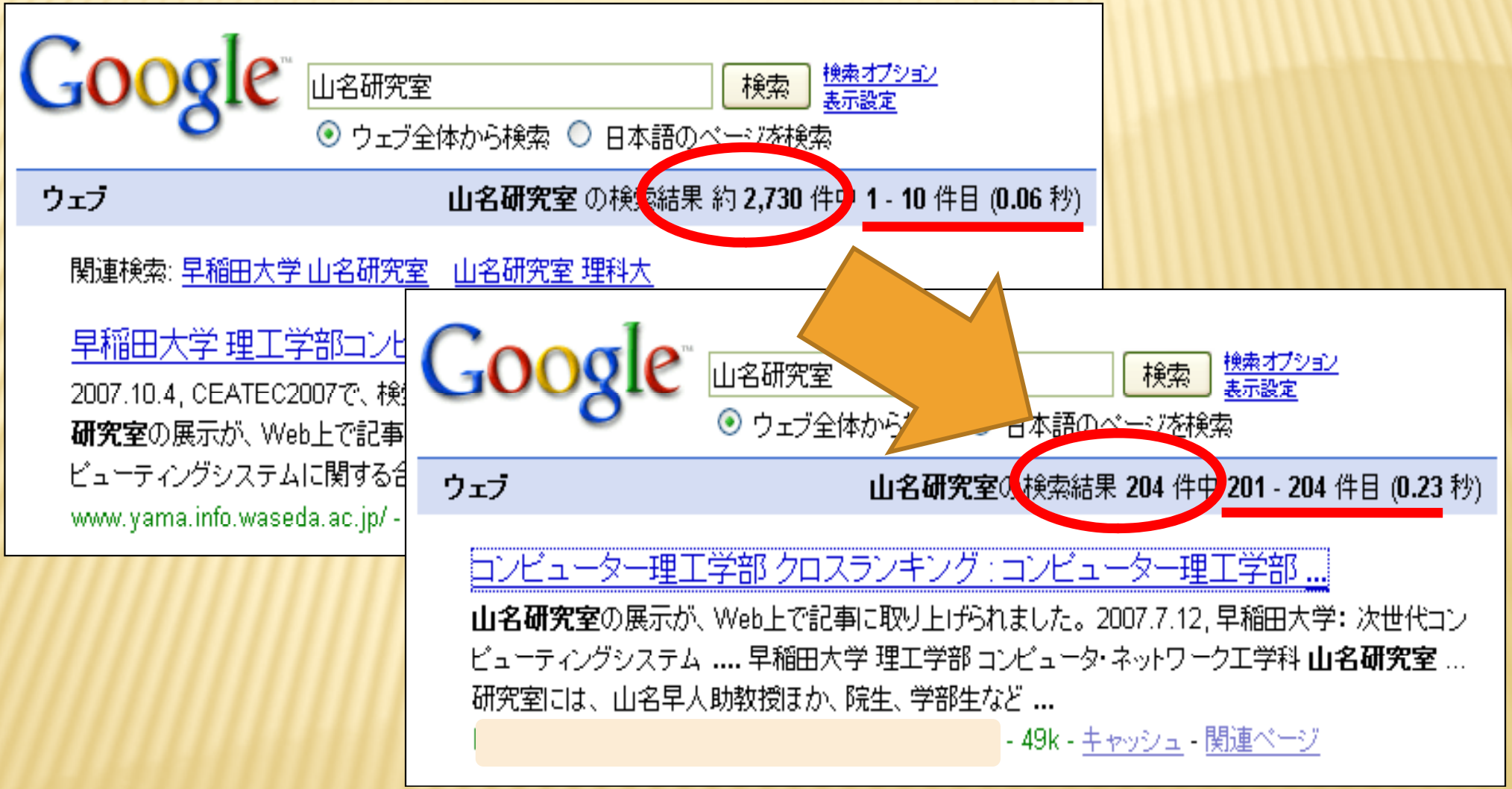
Results **1** - **10** of about **104,000,000** for **Barack Obama**. (**0.10** seconds)

Google

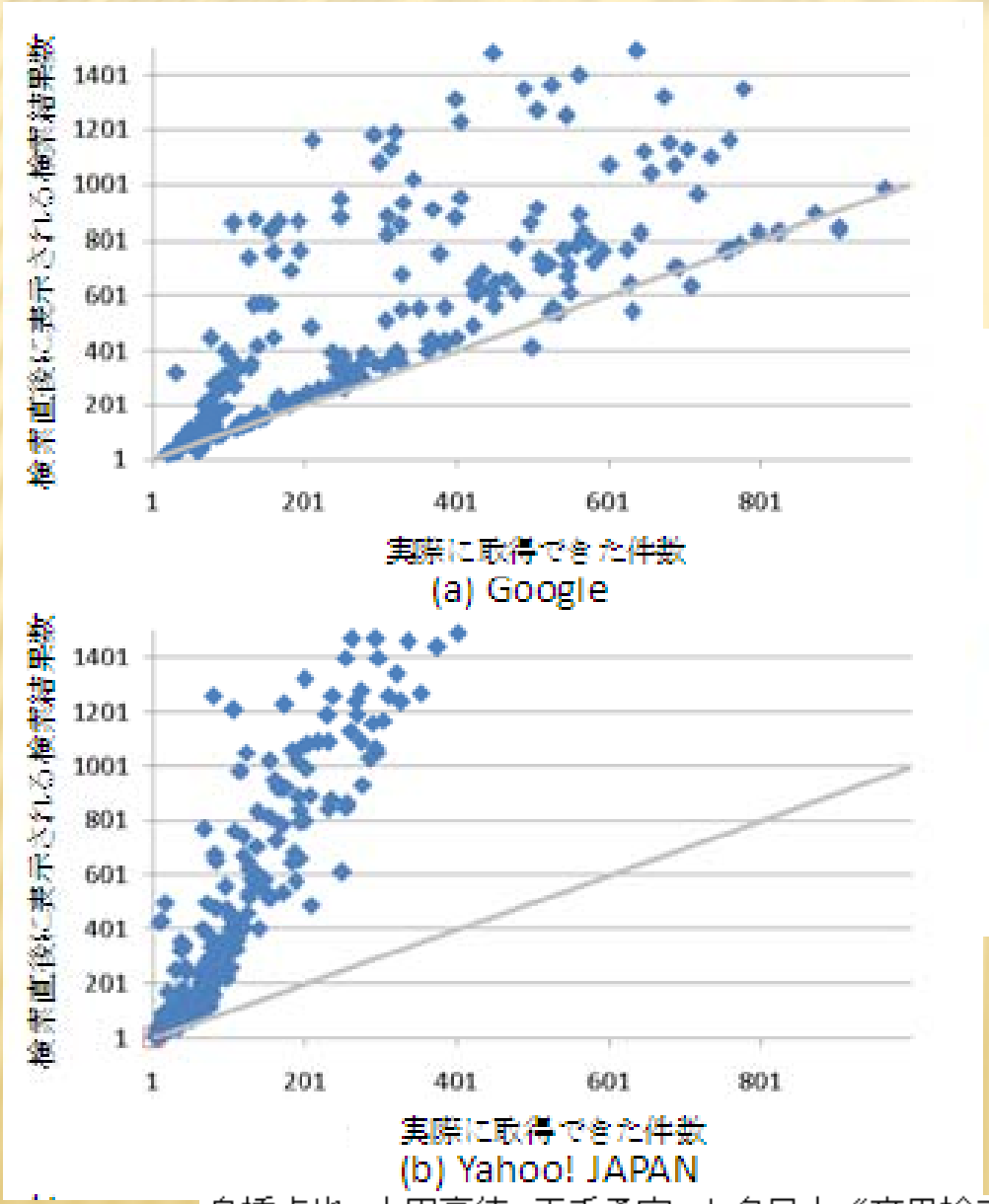
Results **1** - **10** of about **2,640,000,000** for **Google**. (**0.09** seconds)

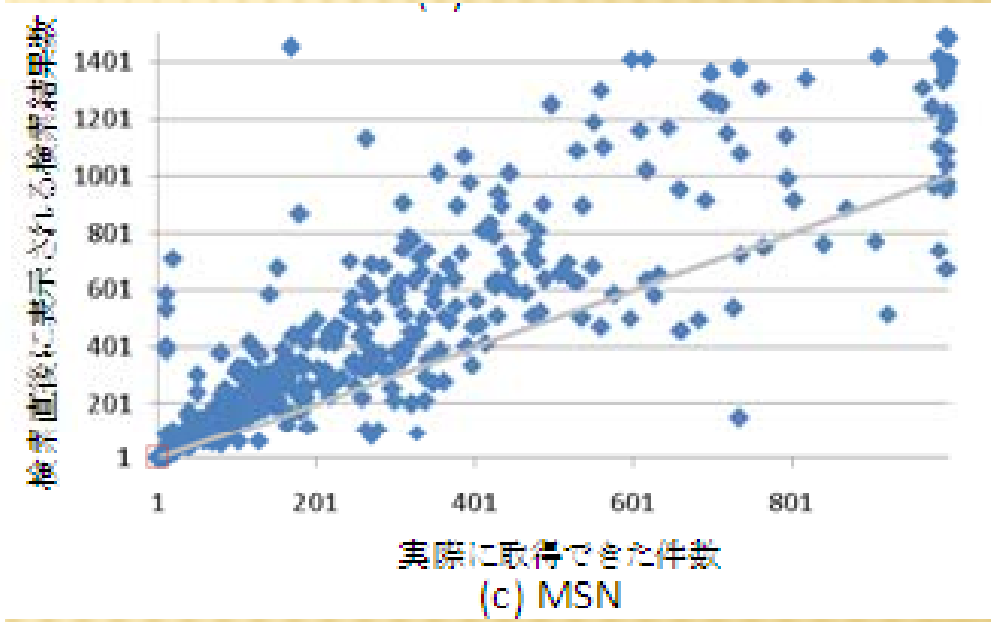


しかし、検索結果ヒット数は変動する…



検索結果ヒット数と実際に得られる数





舟橋卓也、上田高徳、平手勇宇、山名早人:"商用検索エンジンのヒット数に対する信頼性の検証",情処研報(DBS)/iDB2008, Vol.2008, No.88, pp.139-144 (2008.9)

ヒット数変動

- 検索エンジンが変化する3ケースについて検証
- 「どうすれば信頼できるヒット数が得られるのか」を示す

ヒット数が変化する3つのケース

Case1:「検索」ボタンを何度も押した場合

Case2:「次へ」ボタンを何度も押した場合

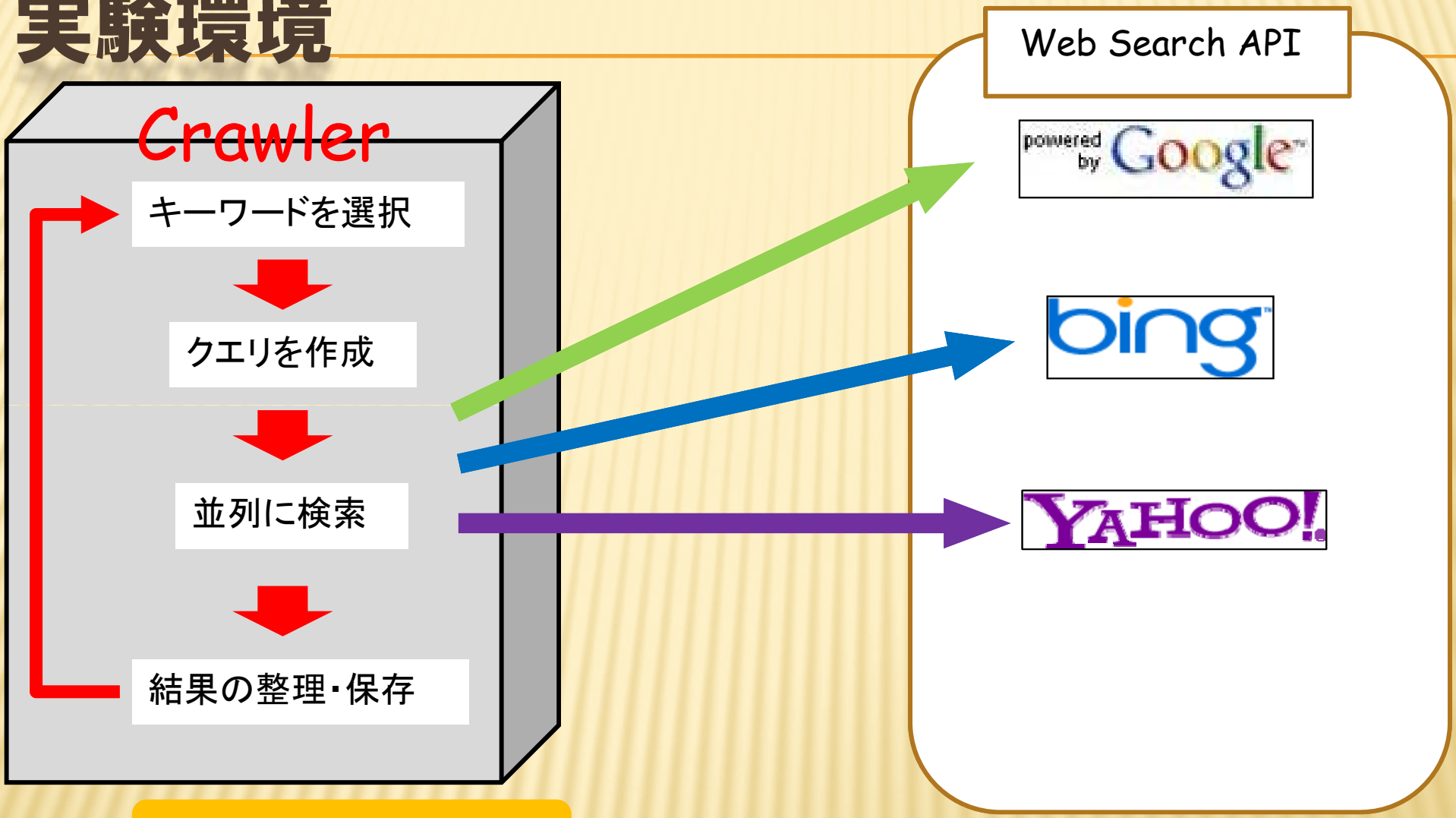
Case3: 検索する日時を変えて検索を行った場合

- 本研究の意義

  ヒット数を利用した研究に,「信頼できるヒット数」の取得方法を提供する

舟橋卓也、曽根広哲、山名早人:"複数キーワードクエリに対する検索ヒット数の信頼性検証", 信学技報, Vol.109, No.153, pp.19-24 (2009.7.28)

実験環境



並列検索のためのクエリ

キーワードAに対して. . .

A : 1-10	A : 11-20	A : 21-30	・・・	A : 991-1000
オフセット1	オフセット11	オフセット21		オフセット991

検証事項

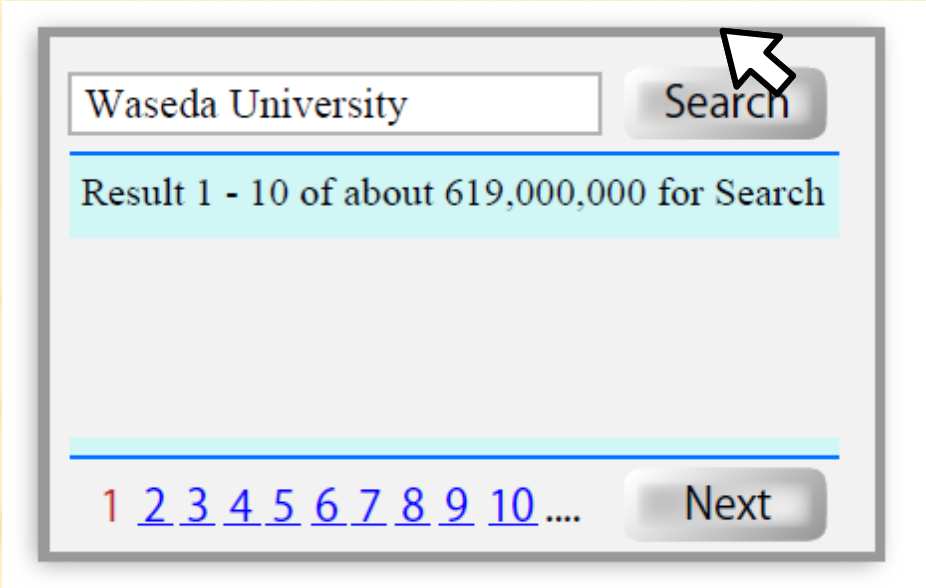
- Google , bing , YAHOO! について検証

- 各社が提供しているAPIを経由して，ヒット数を取得
  - 検索設定は，各検索エンジンのデフォルト設定を利用

- 検索に使用するクエリ
  - 2007年12月のYahoo! JAPANクエリログ（検索頻度の上位10,000件のクエリログ）
  - 「情報爆発」[13]において提供頂いたデータ

[13] 情報爆発時代に向けた新しいIT基盤技術の研究：http://www.infoplosion.nii.ac.jp/info-plosion/

CASE1:「検索」ボタンを何度も押した場合による変動



1. それぞれのクエリに対して，5分以内に100回検索を行う
2. 取得したヒット数から，変動係数を算出する

$$\text{変動係数 cv} = \frac{\text{標準偏差}}{\text{平均}} = \frac{\sqrt{\text{分散}}}{\text{平均}}$$

3. 変動係数でヒストグラムを作成する

CASE1 :検証結果

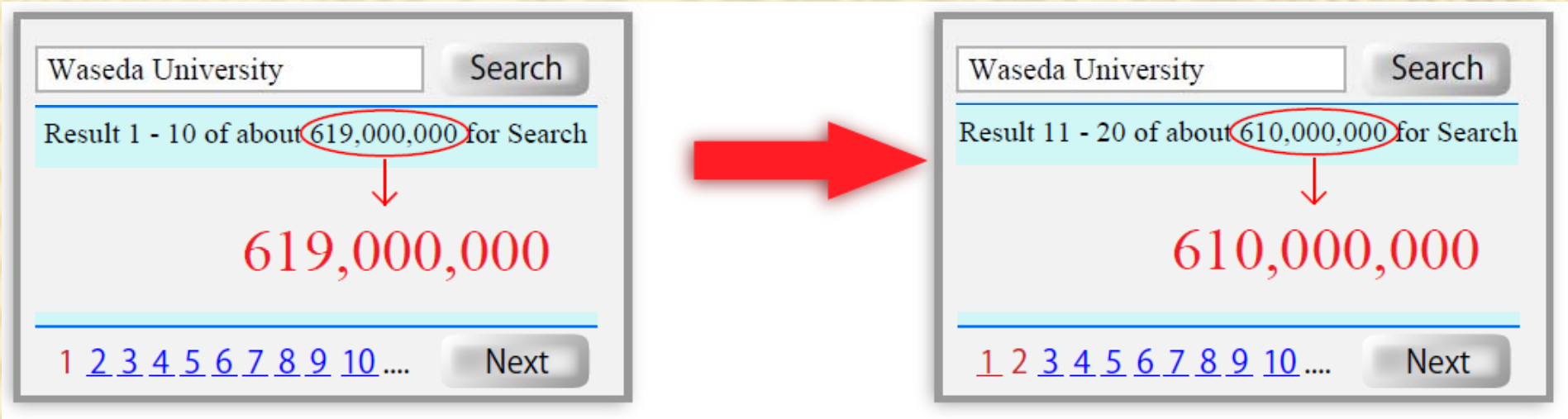
range	Frequency		
	Google	Bing	Yahoo!
cv = 0.0%	9,977	699	9,096
0.0% < cv <= 0.1%	9	2,555	730
0.1% < cv <= 0.5%	0	6,191	46
0.5% < cv <= 1%	0	171	4
1% < cv <= 5%	0	56	1
5% < cv <= 10%	0	12	0
10% < cv <= 20%	0	4	0
20% < cv <= 100%	0	1	0
100% < cv	0	0	0
sum	9,986	9,689	9,877

cv = 変動係数

全ての検索エンジンで，
ヒット数の変動幅はおおむね5%以下

→ 影響小

CASE2：「次へ」ボタンを何度も押した場合の変動

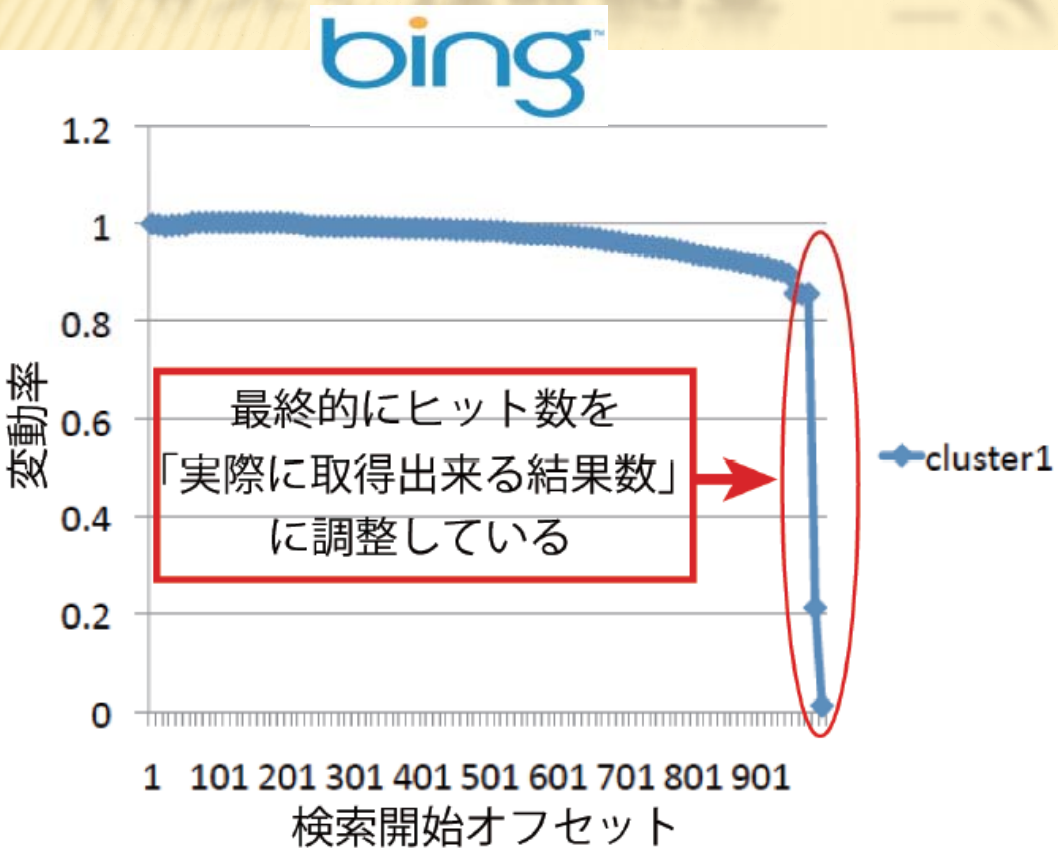


- 検証方法
  - *Deep Hit Count Vector (DV)*を定義

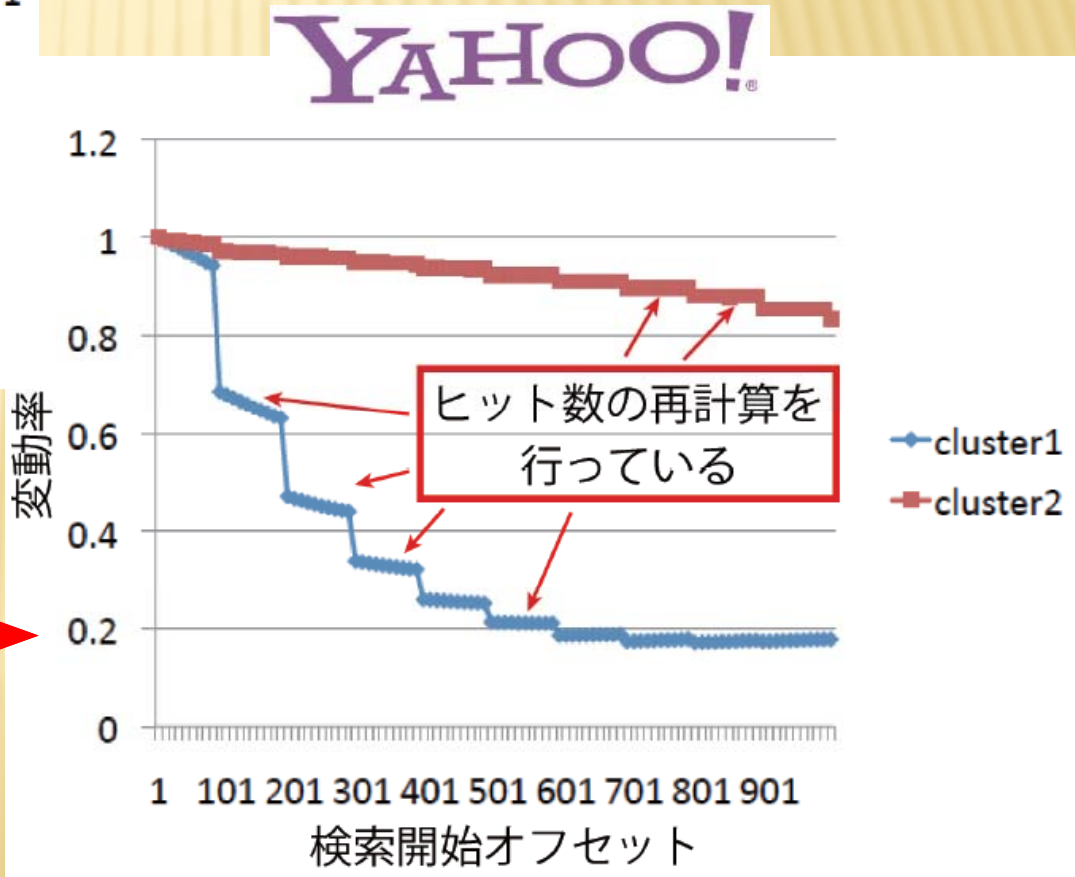
$$DV = \{\frac{HitCount(1,10)}{HitCount(1,10)}, \frac{HitCount(11,20)}{HitCount(1,10)}, \ldots, \frac{HitCount(991/1000)}{HitCount(1/10)}\}$$

    - *HitCount(k, k+9)*　：検索開始オフセットがkのときのヒット数
    - 検索開始オフセット：検索を開始するランキングのオフセット
    - それぞれの要素は，HitCount(1,10)に対する変動率
  - DVを全てのクエリに対して求め，クラスタリング

CASE2: 検証結果　ークラスタリングー



最終的なヒット数が
実際に取得できるWebページ数に
調整されている



ヒット数は，検索開始オフセットが
100変化するごとに大きく変動

CASE 3:検索日時の変化によるヒット数の変動

- 2009　10/11から 12/12の約二ヵ月間,
  10,000個のクエリについて毎日ヒット数を取得

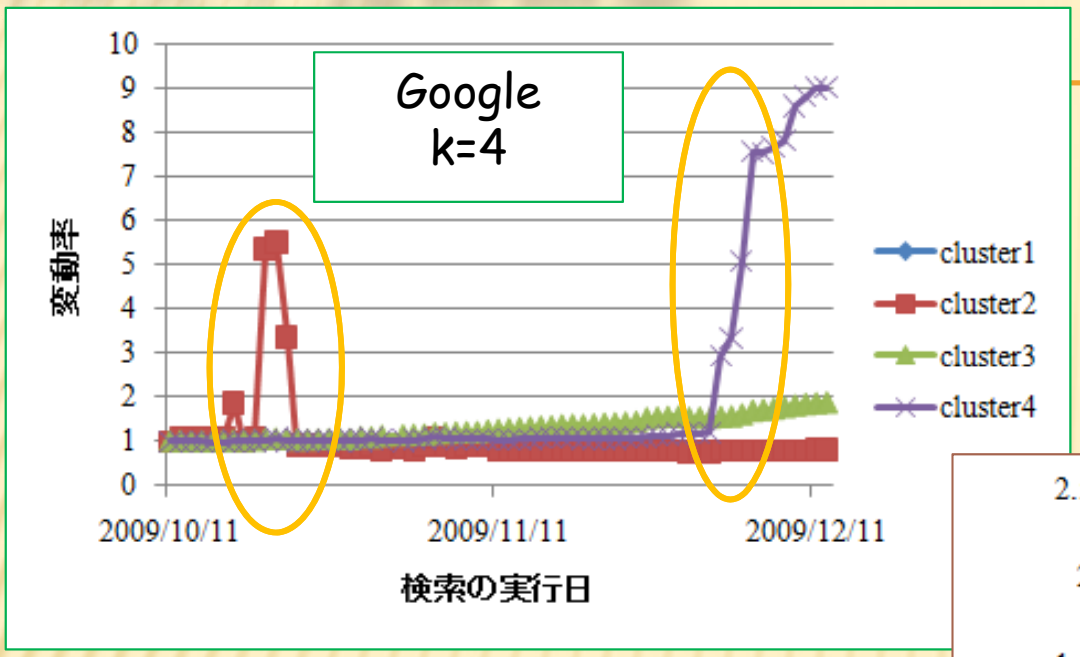
$$\{HitCount(10/11),\ HitCount(10/12),\ ...,\ HitCount(12/12)\}$$

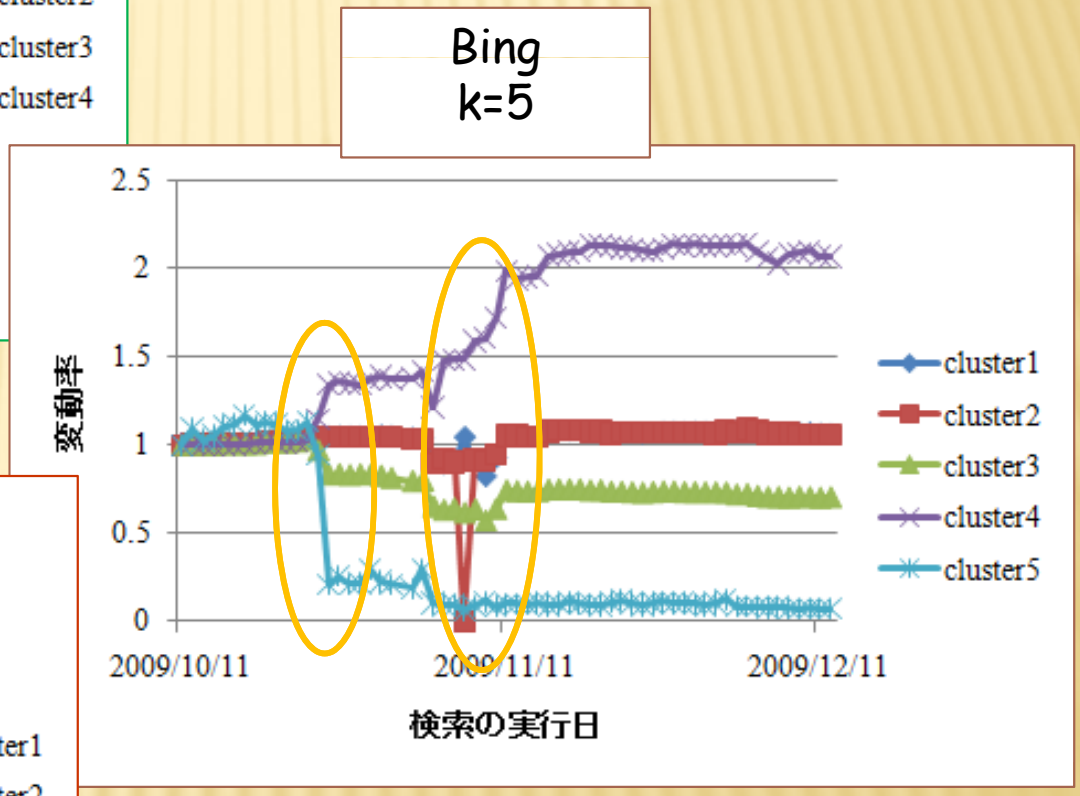
- 10/11のヒット数に対する変動率ベクトルを算出

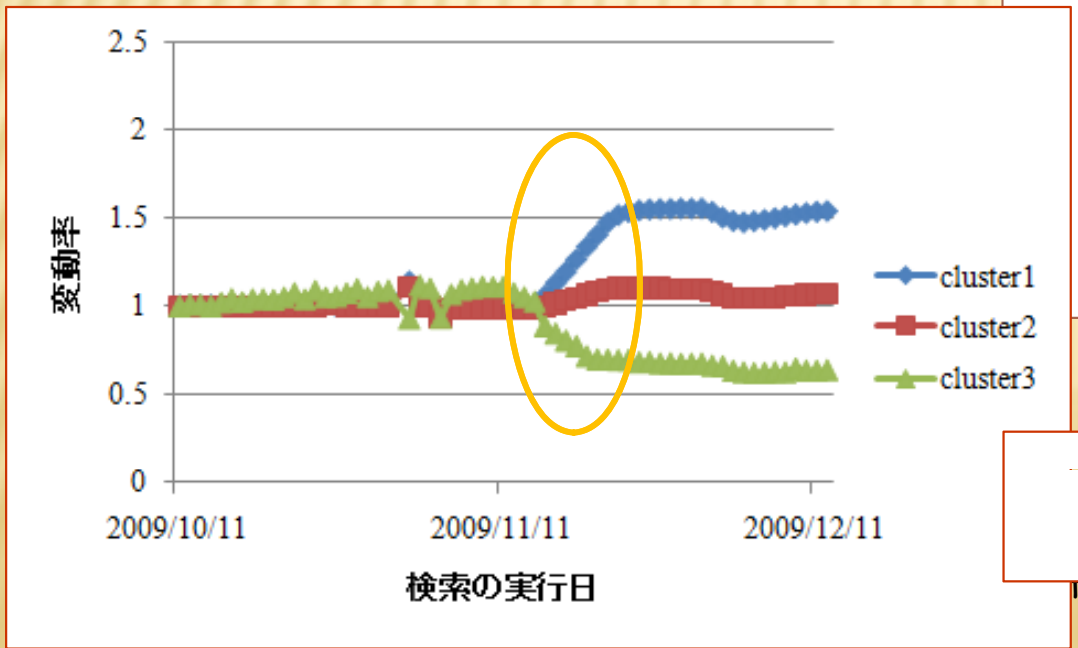
$$\left\{\frac{HitCount(10/11)}{HitCount(10/11)}, \frac{HitCount(10/12)}{HitCount(10/11)}, \ldots, \frac{HitCount(12/12)}{HitCount(10/11)}\right\}$$

- 算出した変動率ベクトルをk-meansクラスタリング
  - 利用する特徴量はコサイン類似度

CASE3 :検証結果







Yahoo!
k=3

まとめ：信頼できるヒット数とは

- 次のような場合に信頼できるヒット数が得られる

検索開始オフセットの変化に伴うヒット数の変動に対し

ヒット数が恣意的に調整されておらず，
かつ，検索開始オフセットが最も大きな値のとき

時間経過に伴うヒット数の変動に対し

時系列によるヒット数の変動が「安定期」（１週間で30％以上変化しない）に入っている場合

検索エンジンの信頼性について

- ユーザの立場としてできること
  - ランキング
    - 調べ物をする際には、必ず複数の検索エンジンを用いる
    - 特にサーベイでは必須
    - ランキングは、検索エンジン間でも大きく異なるのはもちろん、日々変動していることを念頭におく
  - 検索結果数
    - 検索結果数をもって「どちらが有名か」を判断する場合には、必ず「次へ」をクリックして、最後に表示される検索結果数を信頼する
    - 検索結果数は日々変動しているので、できれば１週間程度の期間、調査することが望ましい。

５. 新しい検索エンジンと未来

2020年のSEARCHは？（WSDM2008）

- 従来の検索ボックス
- （ソーシャル　or 人間）パワーサーチ
  - Yahoo!JAPAN 知恵袋
  - OKWave
- Machine Reading
  = Information Extraction + Tractable Interface
  - Text Runner
- 自然言語サーチ
  - セマンティックWeb
    - true knowledge(ケンブリッジ大学)
    - Powerset

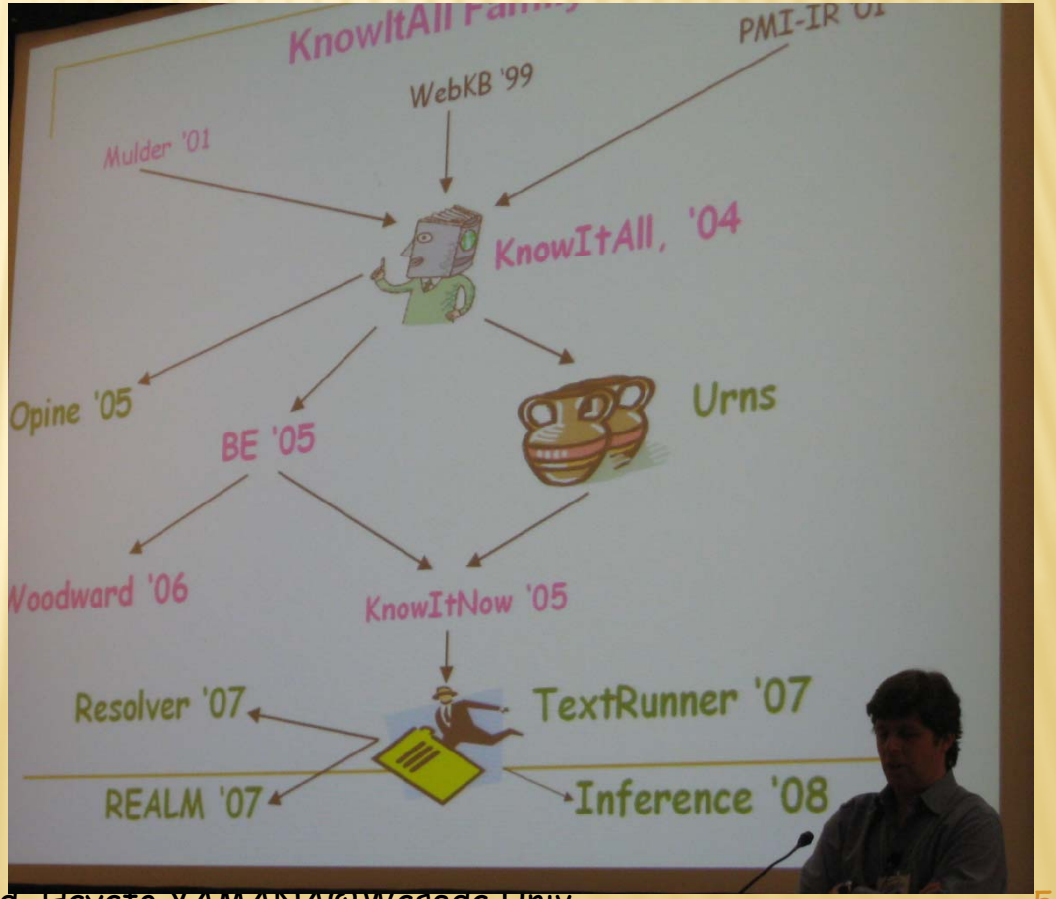
これを区別できますか？

- Book by children
- Book for children
- Book of

①TEXT RUNNER

ワシントン大学教授　Oren Etzioni氏

- 2020年はMachine Readingの時代
  - Machine Reading = Information Extraction + Tractable Interface
- KnowItAllプロジェクト
  - 1億のWebページから5億の文を抽出し利用



HUMAN READING V.S. MACHINE READING

Human Reading	Machine Reading
■ High precision	■ Noisy
■ Broad scope	■ Limited scope
■ Sentence-by-sentence	■ **Corpus-wide statistics**
■ High comprehension	■ Minimal reasoning
■ Background Knowledge.	■ Bottom up
■ Single language	■ **General**
■ Slow	■ **Very Fast!**

Table 1: Human reading and Machine Reading (MR) side-by-side. Despite being much weaker than human reading, MR already exhibits some intriguing capabilities, shown in bold above.

http://turing.cs.washington.edu/papers/SSS07EtzioniO.pdf

TEXTRUNNER SEARCH

TextRunner Search

Retrieved **373** results for **What did Thomas Edison invent?**.

Grouping results by argument 1. Group by: predicate | argument 2

Thomas Edison - 7 results

Thomas Edison invented the light bulb (66), the phonograph (50), the electric light (11), ***31 more...***
Edison did not **invent** the light bulb (10), the electric lightbulb (3)
Thomas Edison had **invented** the electric light bulb (2)
Thomas Edison who **invents** his cosmic rod with technology (2)
Thomas Edison who **invents** his cosmic rod (2)
Edison did n't **invent** the light bulb (2)
Thomas Edison invented travelling the major European powers (2)

Thomas A. Edison - 1 result

Thomas A. **Edison invented** the phonograph (7), the incandescent light (2)

Long before Thomas Edison - 1 result

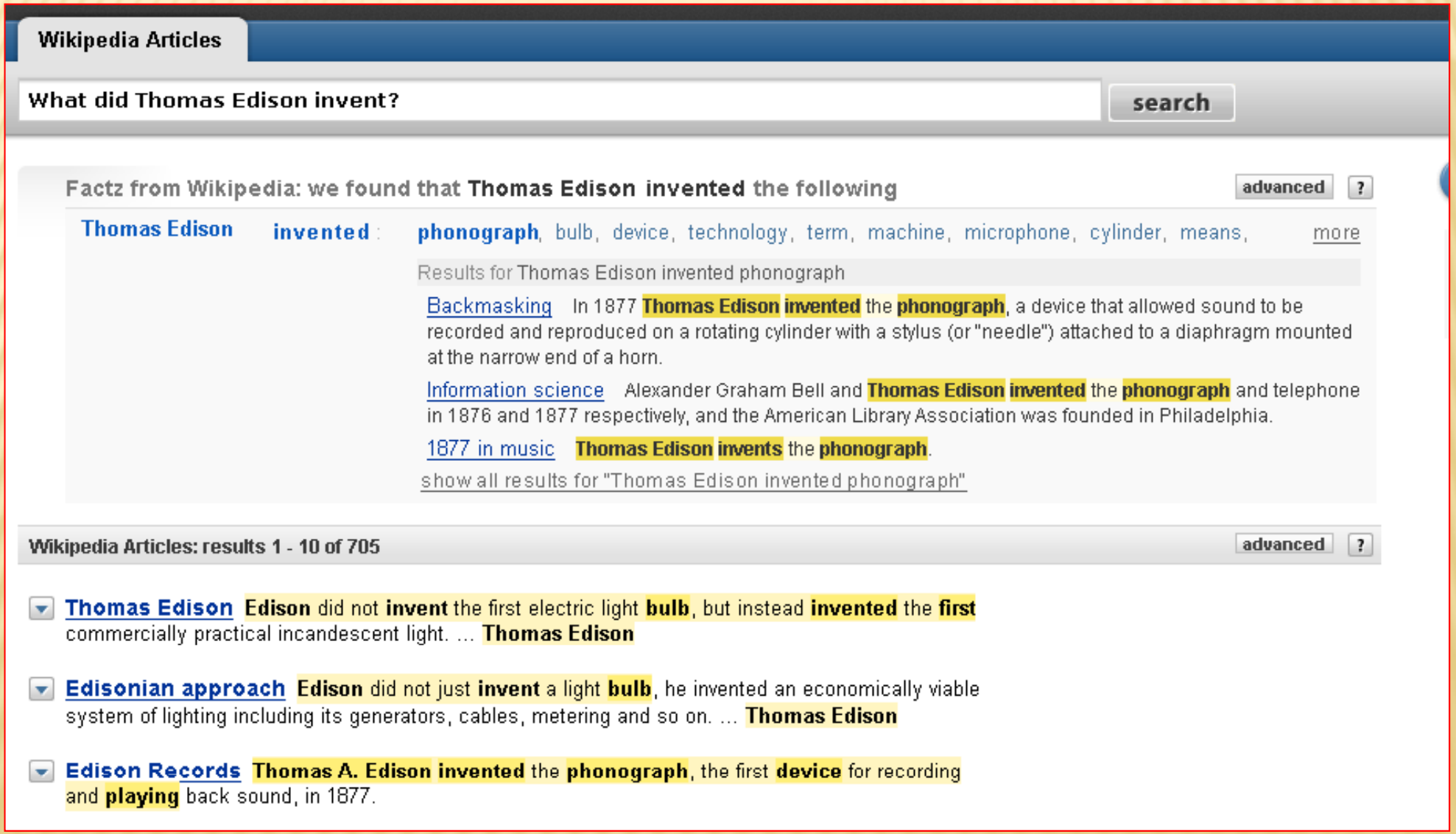
Long before **Thomas Edison invented** film (2)

パターン「Thomas Edison * invent/invented/invents *」を探す

TEXTRUNNERに関する補足

- コーパスはセレクトされたものが重要であり、Webページをそのまま利用することはできない。
- ユーザからのクエリに対してオンザフライで知識を抽出する。
- 事前のタグは不要であり、（arg1, relation, arg2）の関係を自動的に見つける.
- 推論の実現
  - (Turing, born in, London) (london, part of England) → (Tuning born in England)
  - (1,R,2) (1,R,2')(2,R',4),(2',R',4)であれば2=4と推定する。
- 信頼度計算
  - 異なる表現の分布をチェック

②POWERSETの出力結果



POWERSET

Powerset minimize

Website: powerset.com

Location: San Francisco, California, United States

Founded: October, 2006

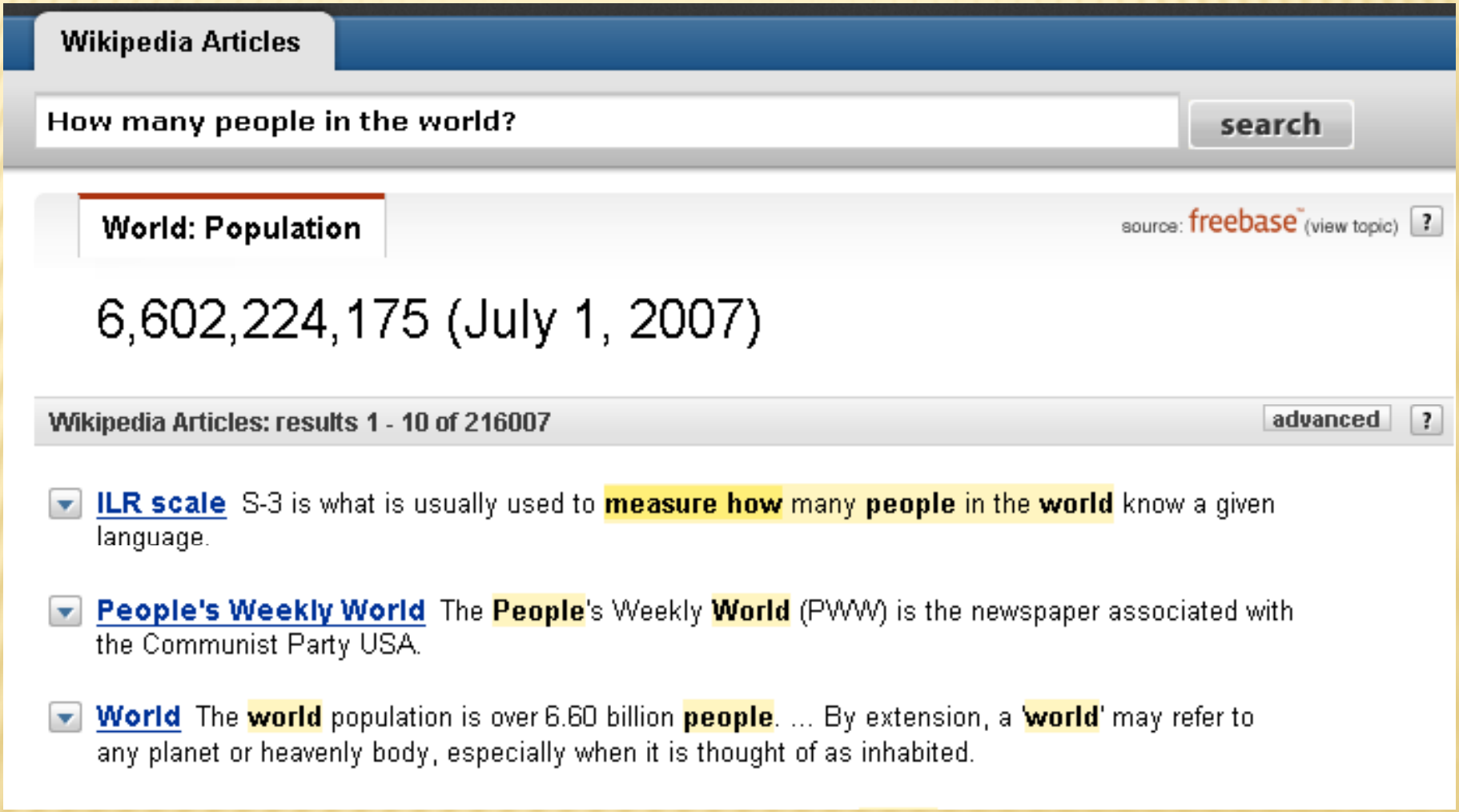
Funding: $12.5M

Powerset is a search engine that focuses on natural language processing. In other words, Powerset will not search based simply on keywords alone, but will try to understand the search phrase as a whole. The goal of the product is to make searching... Learn More

WikipediaとFreebaseのデータを利用

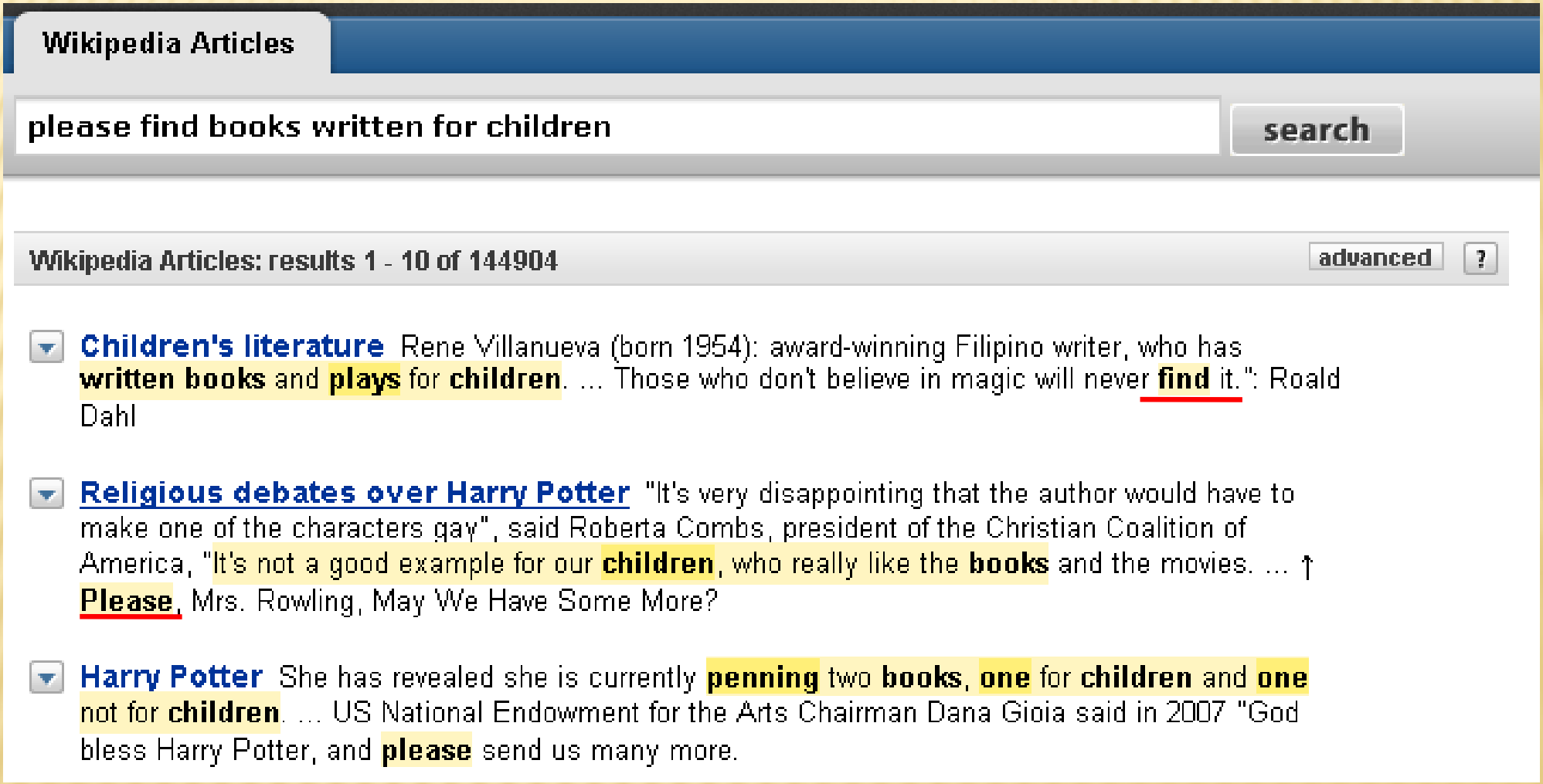
2008／5／11　に一般向けリリース

検索例 1

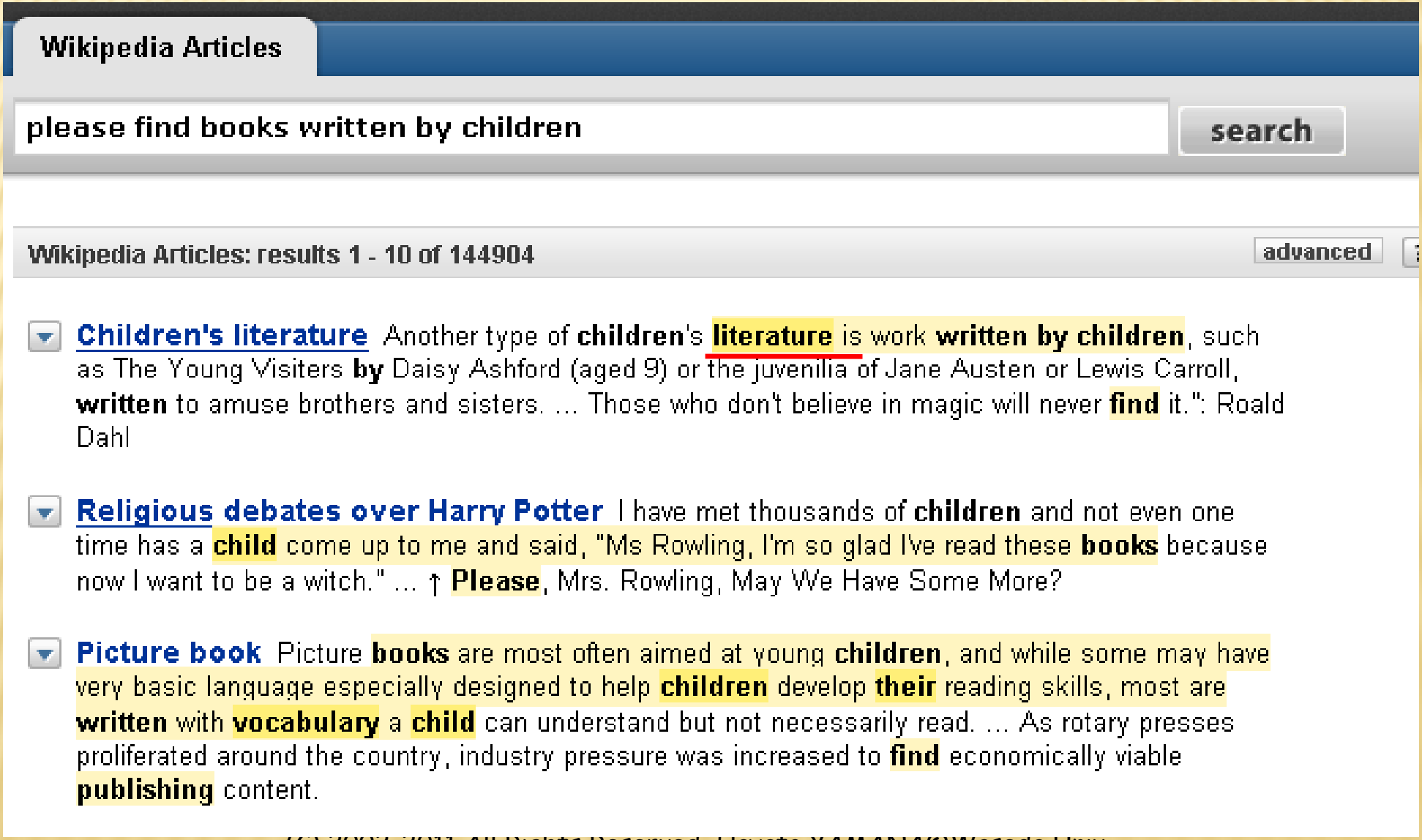


検索例2

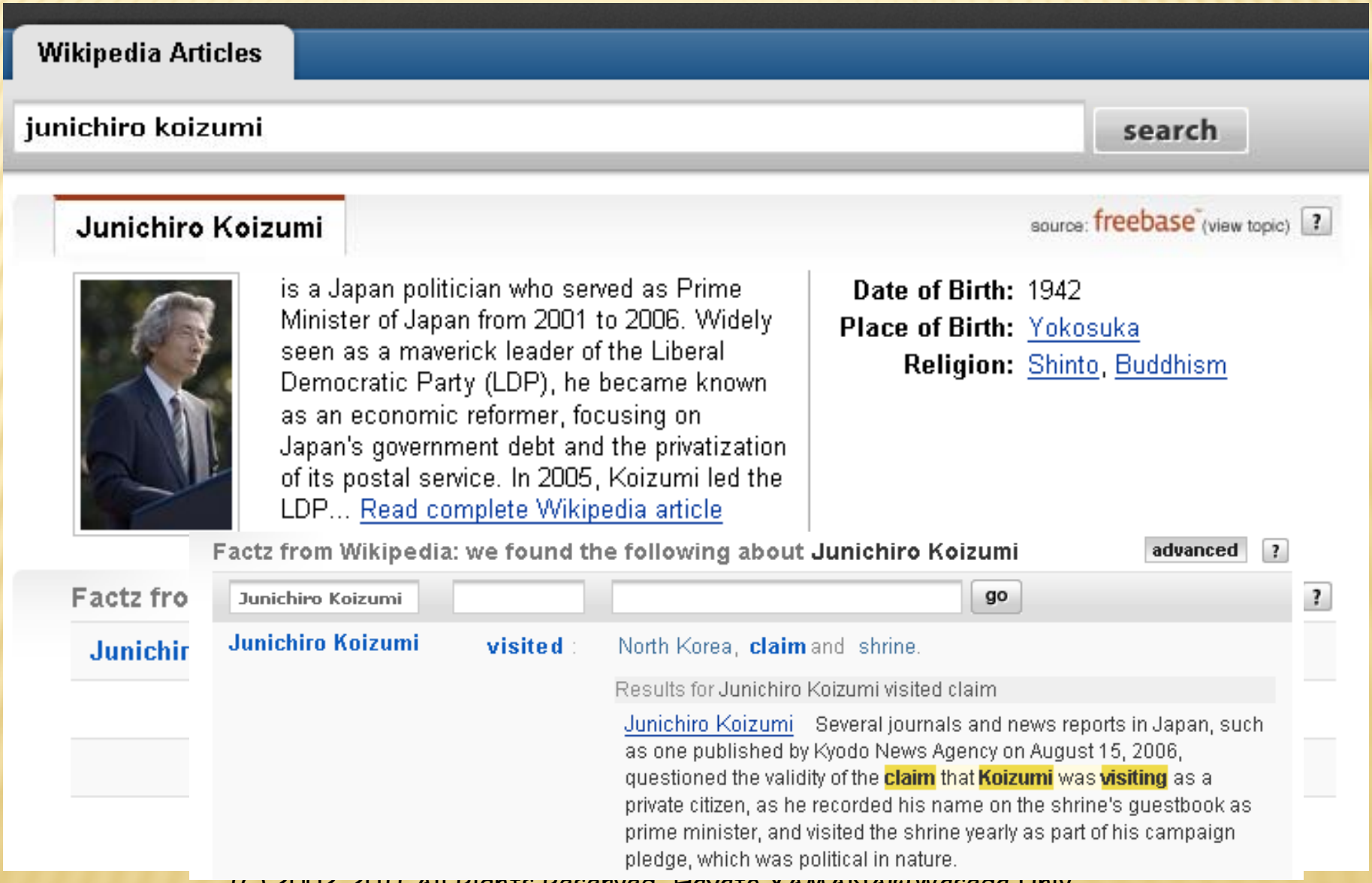
検索例3



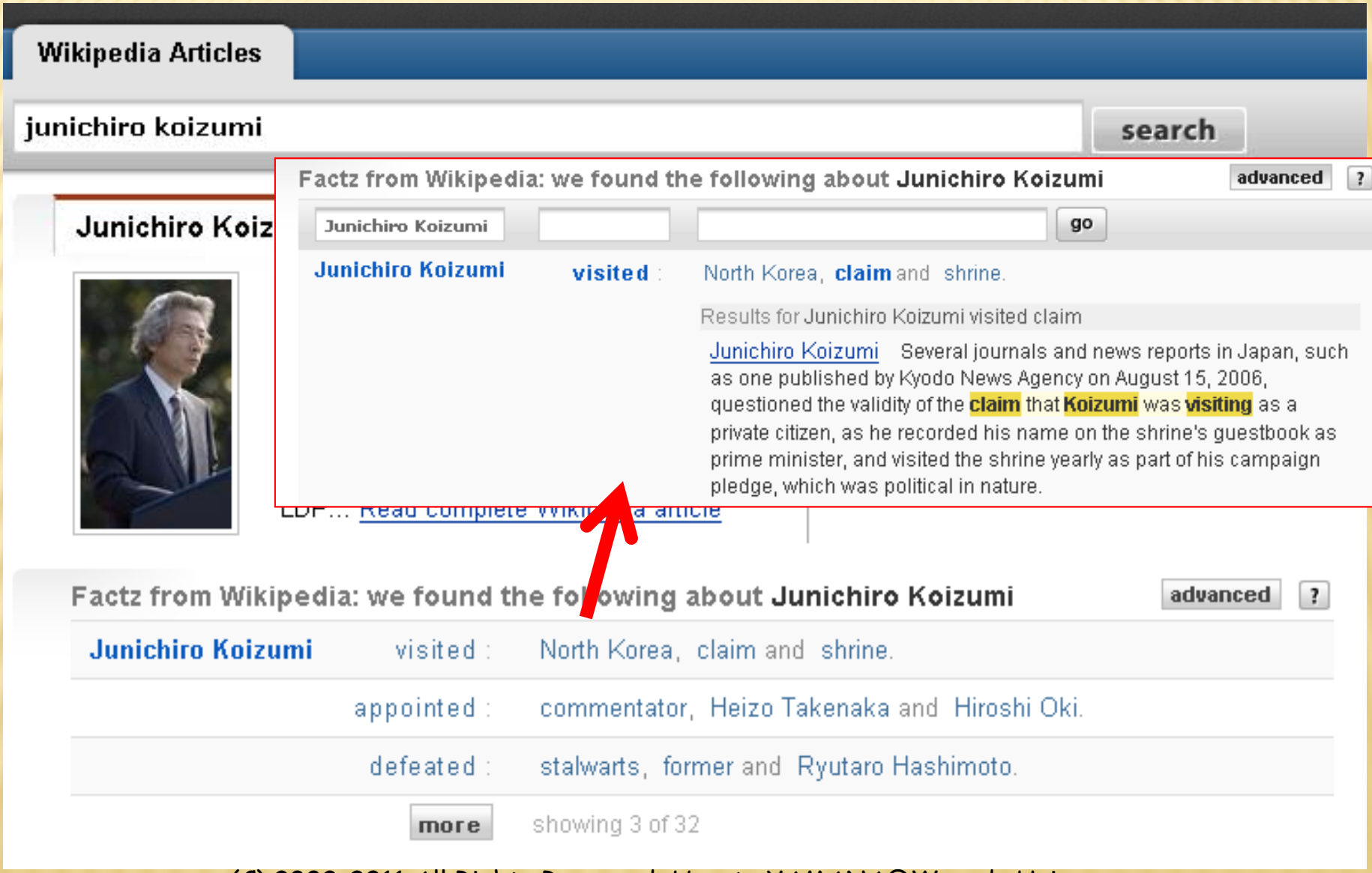
検索例4



検索例５



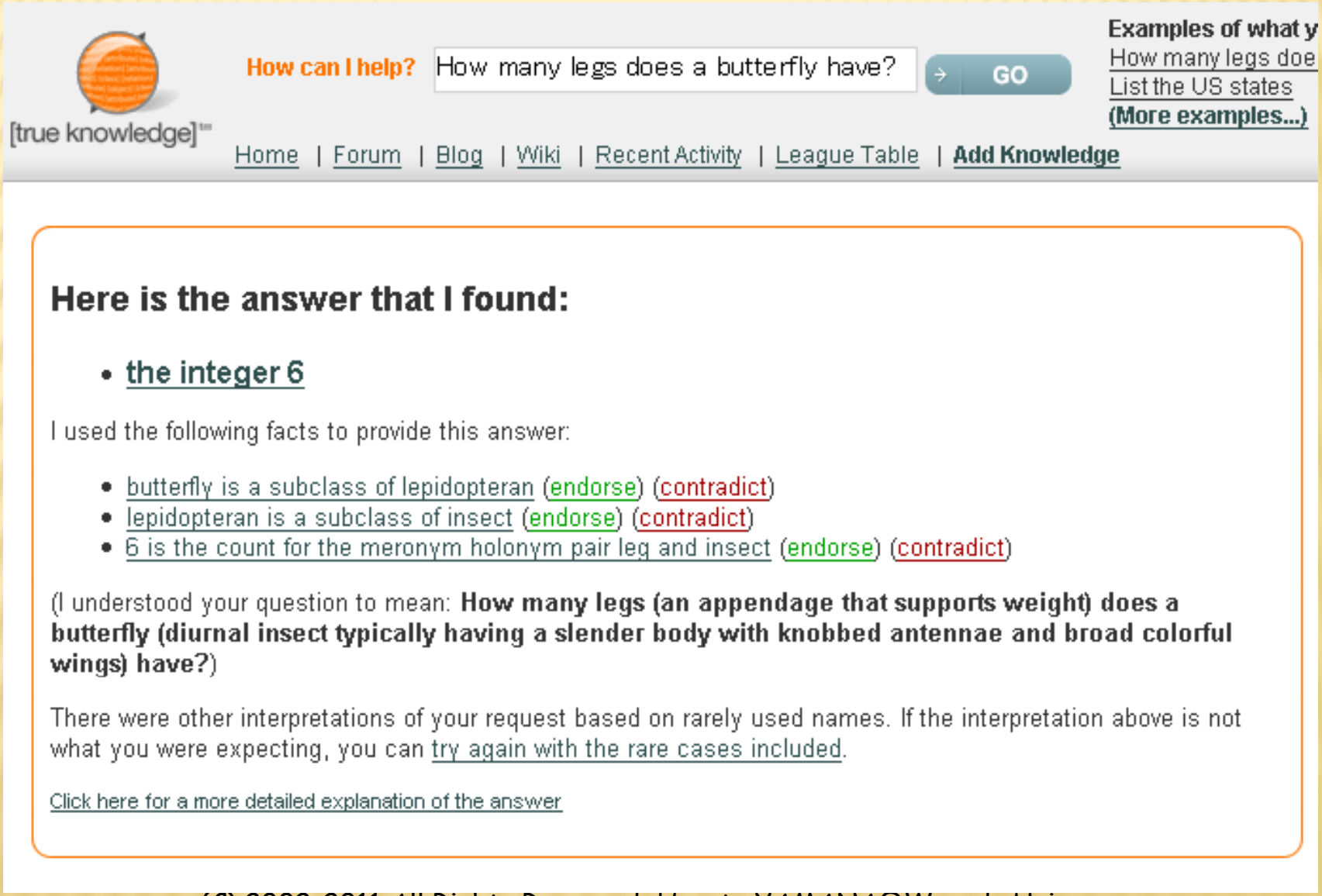
検索例５



POWERSET

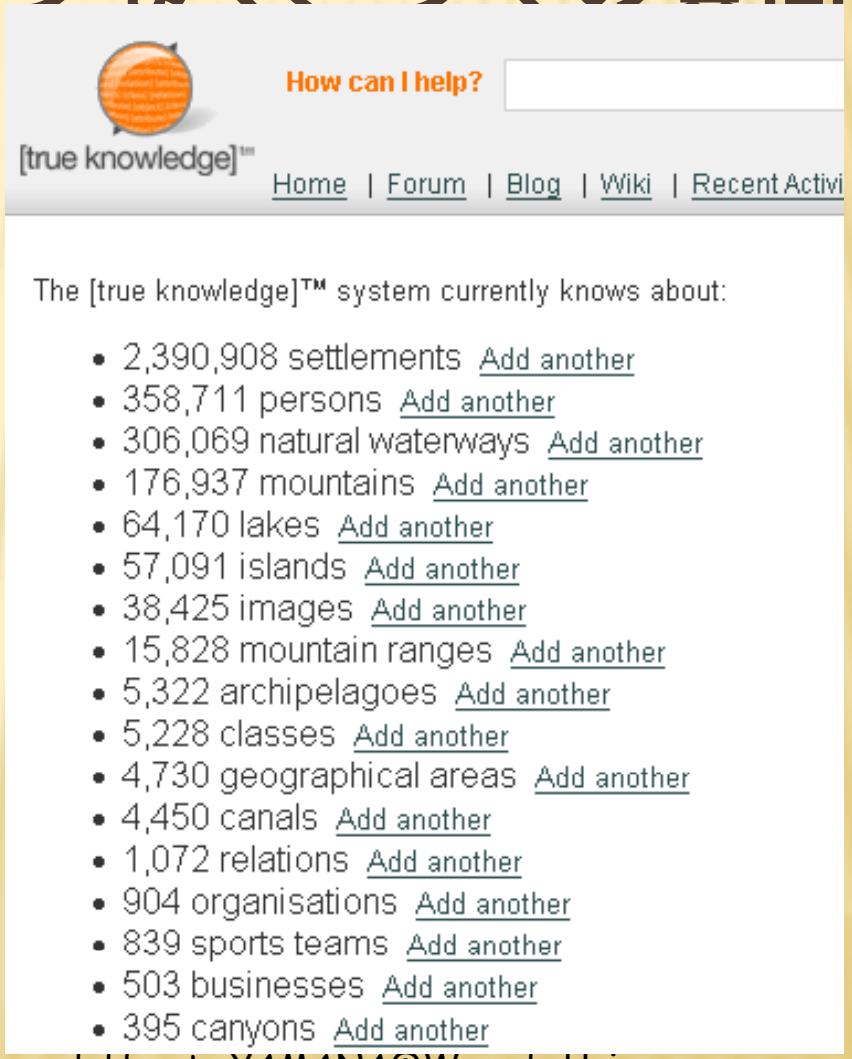
- 自然言語処理をしているようには思えず。
- パターンを抽出しているにすぎないのでは？
  - TextRunnerと似たテクニックの利用

③TRUE KNOWLEDGE



TRUE KNOWLEDGE

- 知識ベースの量が少なく、多くの質問に回答できず。
  - 9907万件のfacts



How can I help? how many people in the world? GO

Examples of what
Is a lemur a prima
new subclass of r
(More examples..

Home | Forum | Blog | Wiki | Recent Activity | League Table | **Add Knowledge**

I found the following interpretations of *how many people in the world*?
Please select the one you meant.

- How many human beings are there?
- How many citizenries (the citizens of a state or country regarded collectively) are there?

There were other interpretations of your request based on rarely used names. If none of the interpretations above are what you were expecting, you can try again with the rare cases included.

Here is the answer that I found:

- **the integer 6677563921**

This conclusion is based on a single fact in the knowledge base:

[fact pattern: [object unspecified]; [is an instance of]; [human being]] is or has been of order 6677563921
(endorse) (contradict)

Thanks to **jonanin** who added knowledge necessary to answer this question.

(I understood your question to mean: **How many human beings are there?**)

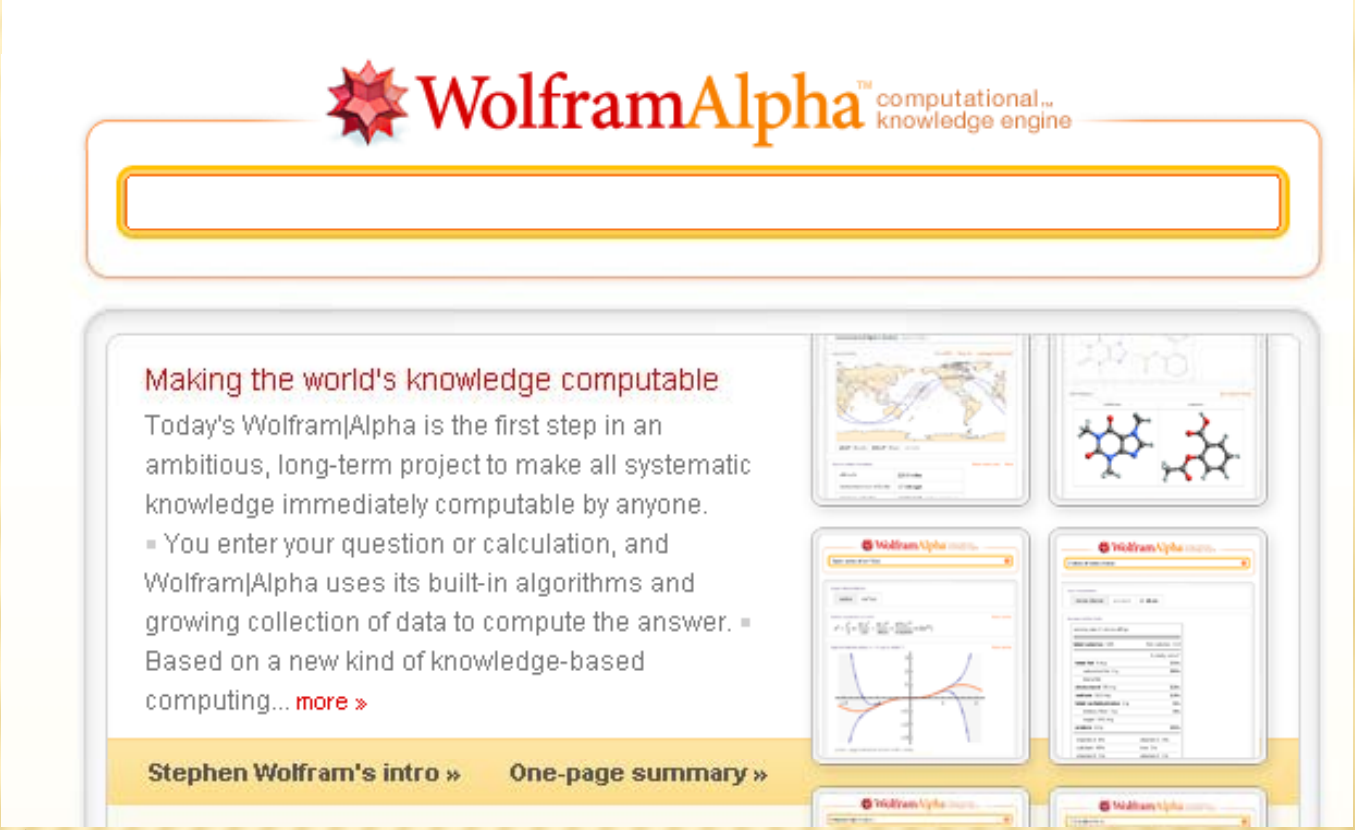
Click here for a more detailed explanation of the answer

④GOOGLE:　JEFFREY DEAN氏による未来（WSDM2009）

- Future Directions
  - Translate all the world's documents to all the world's languages(全世界の言語で全世界のドキュメントを検索！)
    - continuously improving translation quality
    - large-scale systems work to deal with larger and more complex language models
    - (e.g.) to translate one sentence ⇒ ~1M lookups in multi-TB model
  - ACLs(Access Control Lists) in Information Retrieval Systems（アクセスコントロール）
    - best solution for doc shared with 10 people is different than for doc shared with the world
    - sharing patterns of a document might change over time
  - Automatic Construction of Efficient IR Systems（パラメータチューニング）
  - Information Extraction from Semi-structured Data（半構造データ）

http://labs.google.com/papers.html

⑤WOLFARMALPHA(ウルフラムアルファ）



- **計算知能エンジン**（computational knowledge engine)
  - **情報系大学生なら一度は使ったことのある**Mathematica（**数式処理システム**）**を世に送り出した会社である**wolfram research**社が**2009/5/18**から公開**

ウルフラムアルファ

- ３８０００CPU（５つのデータセンターに分散）
- データ
  - 10兆以上のデータ
    - データの一部はWebからとっているが、ほとんどのデータはちゃんとしたデータ提供元から得ている「source information」ボタンで表示。
  - 5万を超えるアルゴリズムとモデル
  - 1000を超えるドメイン（分野）をハンドル
  - 現在は英語のみ（将来は多言語対応）
- プログラム
  - 500万行に及ぶMathematicaのコード
- 開発
  - Stephen Wolframをリーダとした100人程度のチーム
  - 2005年に開発を開始

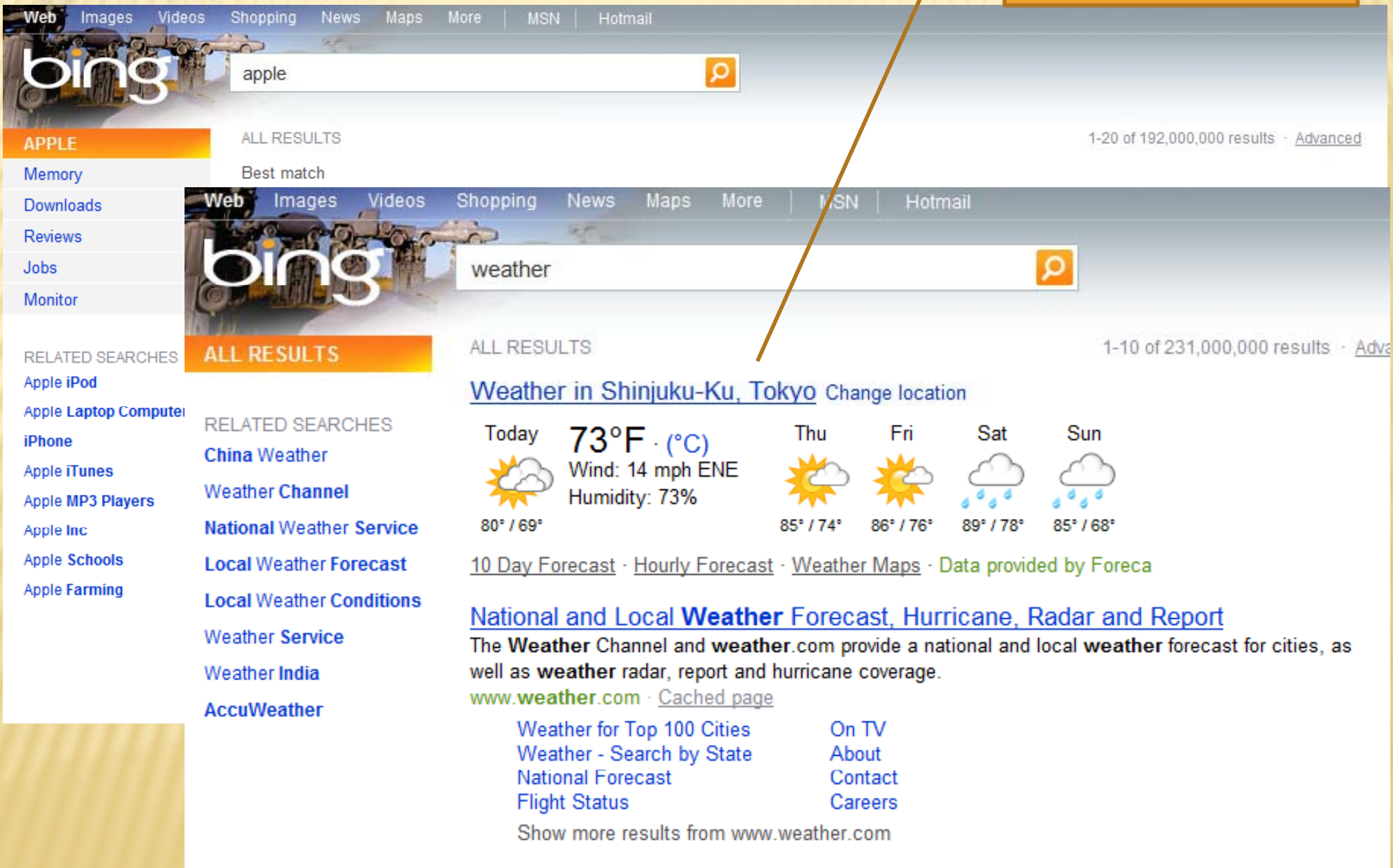
ウルフラムアルファ関連記事

- Wolfram Alpha Launch Starts Tonight at 5pm Pacific: Here is What You Need to Know
  - http://www.readwriteweb.com/archives/wolfram_alpha_launch_starts_tonight.php

- http://wiredvision.jp/news/200905/2009051122.html
  - http://www.wired.com/epicenter/2009/05/how-the-wolfram-alpha-search-engine-could-save-google/

⑥BING

アクセス元を
認識している



⑦NAVER

検索結果数が
表示されない！

各種参考文献は http://www.yama.info.waseda.ac.jp/~yamana/

END