

検索エンジンの「今」とその信頼性

— 未来へ向けて —

早稲田大学 理工学術院
情報理工学科 山名早人

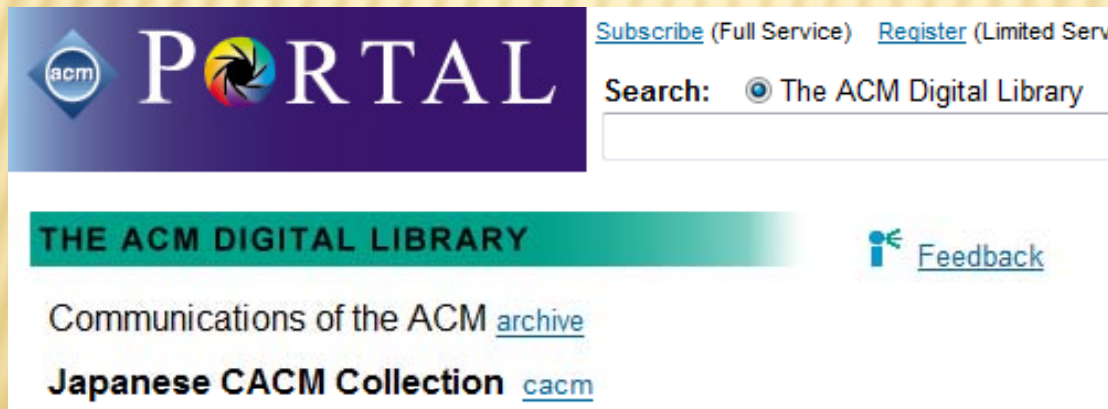
<http://www.yama.info.waseda.ac.jp/~yamana/>

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？

The image displays three browser windows side-by-side, each showing search results for '早稲田大学' (Waseda University). The windows are labeled '早稲田大学 - Bing', '「早稲田大学」の検索結果 - Yahoo!検索', and '早稲田大学 - Google 検索'. Red and green circles highlight specific search results across the three engines, connected by lines to show consistency.

- Bing Search:** The search results include '早稲田大学' (circled in red), '早稲田大学工学部' (circled in green), and '早稲田大学野球部' (circled in red).
- Yahoo! Search:** The search results include '早稲田大学ホームページ' (circled in red), '早稲田大学野球部' (circled in red), and '早稲田大学 - Wikipedia' (circled in green).
- Google Search:** The search results include '早稲田大学' (circled in red), '早稲田大学 - Wikipedia' (circled in green), and '早稲田大学野球部' (circled in red).

The connections between the highlighted results across the three engines are as follows:

- The red circle on '早稲田大学' in Bing connects to the red circle on '早稲田大学' in Google.
- The green circle on '早稲田大学工学部' in Bing connects to the green circle on '早稲田大学 - Wikipedia' in Yahoo! and Google.
- The red circle on '早稲田大学野球部' in Bing connects to the red circle on '早稲田大学野球部' in Yahoo! and Google.

そして、こんな経験も？

ウェブ | 画像 | 動画 | ブログ | 辞書 | 知恵袋 | 地図 | 一覧 ▾

早稲田大学 山名早人 [条件を指定して検索](#)
[検索設定](#) **YAHOO! JAPAN**

ウェブ検索結果 早稲田大学 山名早人 で検索した結果 1~10件目 / 約9,940件 - 0.05秒

ウェブ | 画像 | 動画 | ブログ | 辞書 | 知恵袋 | 地図 | 一覧 ▾

早稲田大学 山名早人 [条件を指定して検索](#)
[検索設定](#) **YAHOO! JAPAN**

ウェブ検索結果 早稲田大学 山名早人 で検索した結果 91~100件目 / 約9,760件 - 0.07秒

ウェブ | 画像 | 動画 | ブログ | 辞書 | 知恵袋 | 地図 | 一覧 ▾

早稲田大学 山名早人 [条件を指定して検索](#)
[検索設定](#) **YAHOO! JAPAN**

ウェブ検索結果 早稲田大学 山名早人 で検索した結果 501~500件目 / 約2,480件 - 0.83秒

[良彰の無料動画はこちらで探して視聴できます](#)

良彰の無料動画の視聴 ... 早稲田大学. 早稲田大学 理工学術院. 早稲田大学 基幹理工学部 情報理工学科. ... 山名 早人. ネットワーク利用技術開発. 情報オブジェクト化技術開発 ... [12] 眞隈 志, 深澤 ...
無料音楽 [.biz/index3.php?key=良彰](#)

新しい「検索エンジン」？

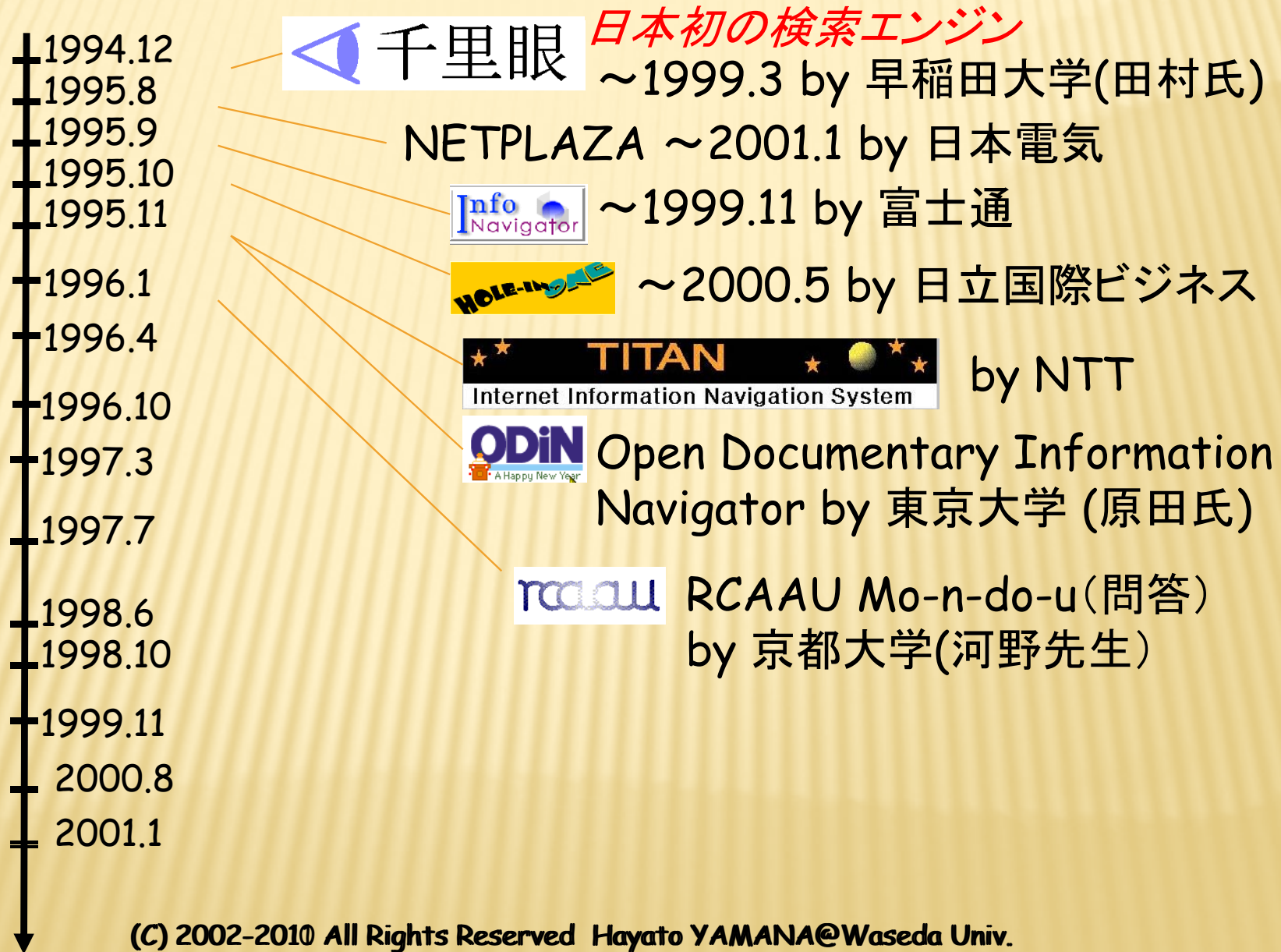
The image displays three overlapping search engine interfaces. At the top is WolframAlpha, featuring its red star logo and the text 'WolframAlpha™ computational knowledge engine'. Below it is True Knowledge, with the text 'True Knowledge® The Internet Answer Engine™ BETA' and 'register sign in' links. At the bottom is Powerset, with the text 'Powerset Live questions: Ready. Powerset. Go.' and a search bar containing 'Enter a topic, phrase or question' and an 'Explore' button. A small yellow sticky note in the bottom right corner of the Powerset interface reads 'Microsoft brings you a better way'.

AGENDA

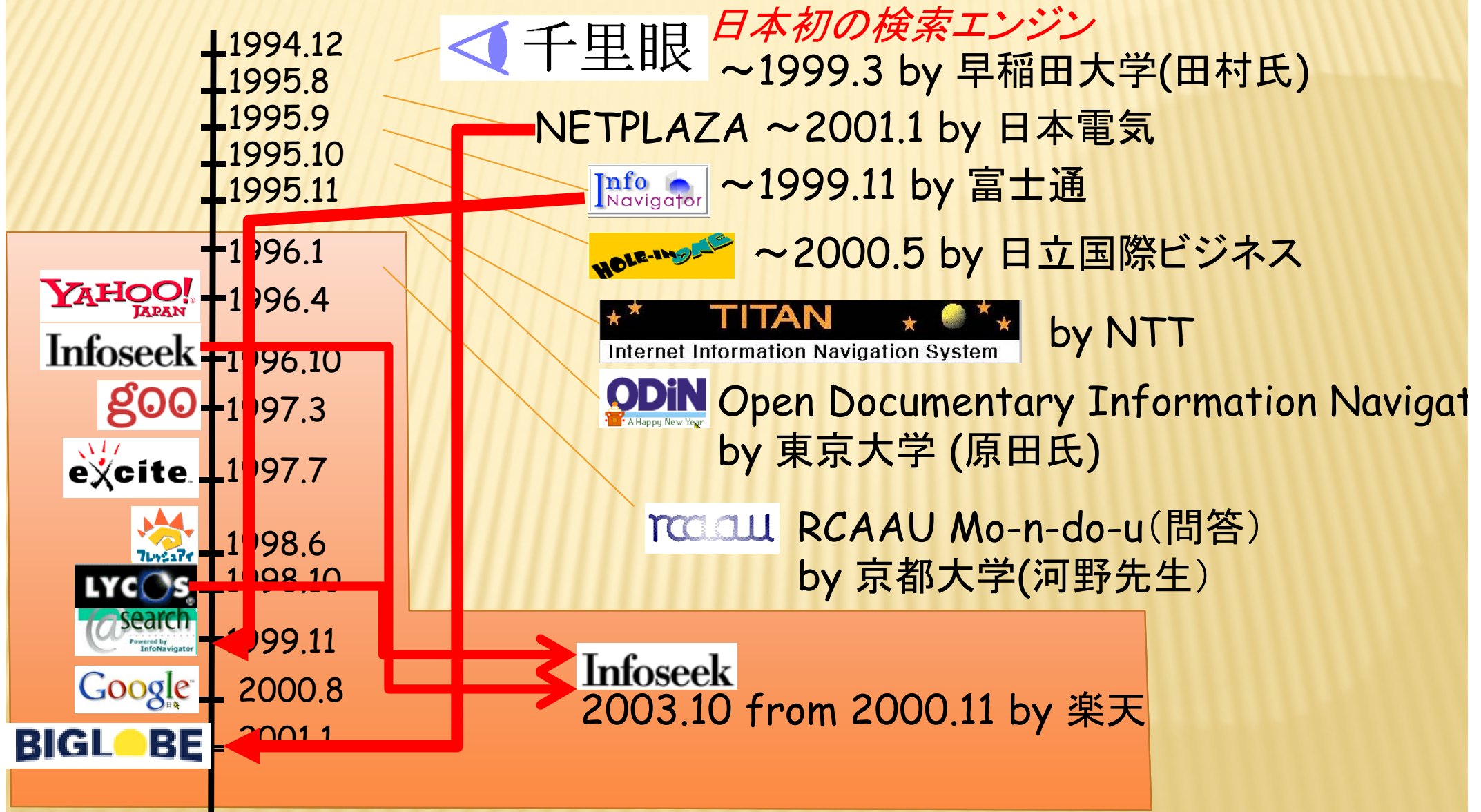
1. 日本における検索エンジンの歴史
2. WEBの規模と現状
3. 検索エンジンの信頼性ーランキンゲー
4. 検索エンジンの信頼性ー検索結果数ー
5. 新しい検索エンジンと未来

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）



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

- × 
 - + 1996.4 powered by Original
 - + 1998.5 powered by 
 - + 2001.4 powered by 
 - + 2004.5 powered by YST
- × NETPLAZA  (NEC)
 - + 1995.9 powered by Original
 - + 2000.11 powered by 
- × InfoNavi/ (Fujitsu)
 - + 1995.10 powered by Original
 - + 2001.4 powered by 
 - + 2009.夏 powered by YST
- ×  (NTT)
 - + 1997.3 powered by Original
 - + 2003.12 
- × 
 - + Original
 - + 2009.5  ^
- Infoseek Japan
 - 1996.10 powered by Original
 - 2003.9 powered 
- Excite Japan
 - 1997.7 powered by Original
 - 2002.1 powered 
- 
 - 2008.1 powered by Original
- 
 - 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

個人情報保護のため削除

千里眼

apan)



poratorie

n)



ほとんどの
研究者は
Googleへ

2. WEBの規模と現状

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

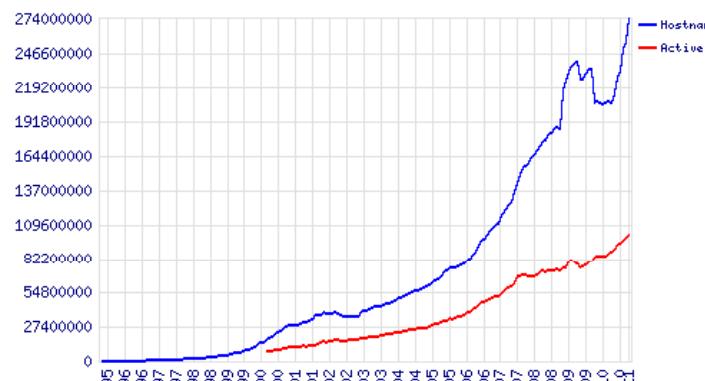
nginx once again saw the largest increase in market share amongst the top web servers, gaining 0.88 percentage points or 3.59M hostnames. nginx now has 43.1M hostnames, up from 39.5M in the previous survey. The top 1 million sites. 2.4M of the additional nginx hostnames are hosted by Ecatel, which appears to have migrated from Apache to nginx this month. nginx now has 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 Secure

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 United States is concentrated in AmerINOC and Softlayer, together accounting for 7.34M hostnames. Germany's growth, on the other hand, is concentrated in 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



世界に

× 1997年

+ 大規模統計

× 1999年

+ 360万Web

× 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

?

なりから

2500の
的に算出。

[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

<http://www.cuil.com/>

cuil

Search

Search 127 billion pages

[About Cuil](#) | [Preferences](#) | [Add Cuil to Internet Explorer](#)

Private

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



🔥 [Highlights and Videos from CrunchGear's CES coverage](#) >>

Cuil Fails to Be Acquired

Michael Arrington

Sep 19, 2010



176



348

16



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 Cuil's assets - particularly Google and Microsoft.

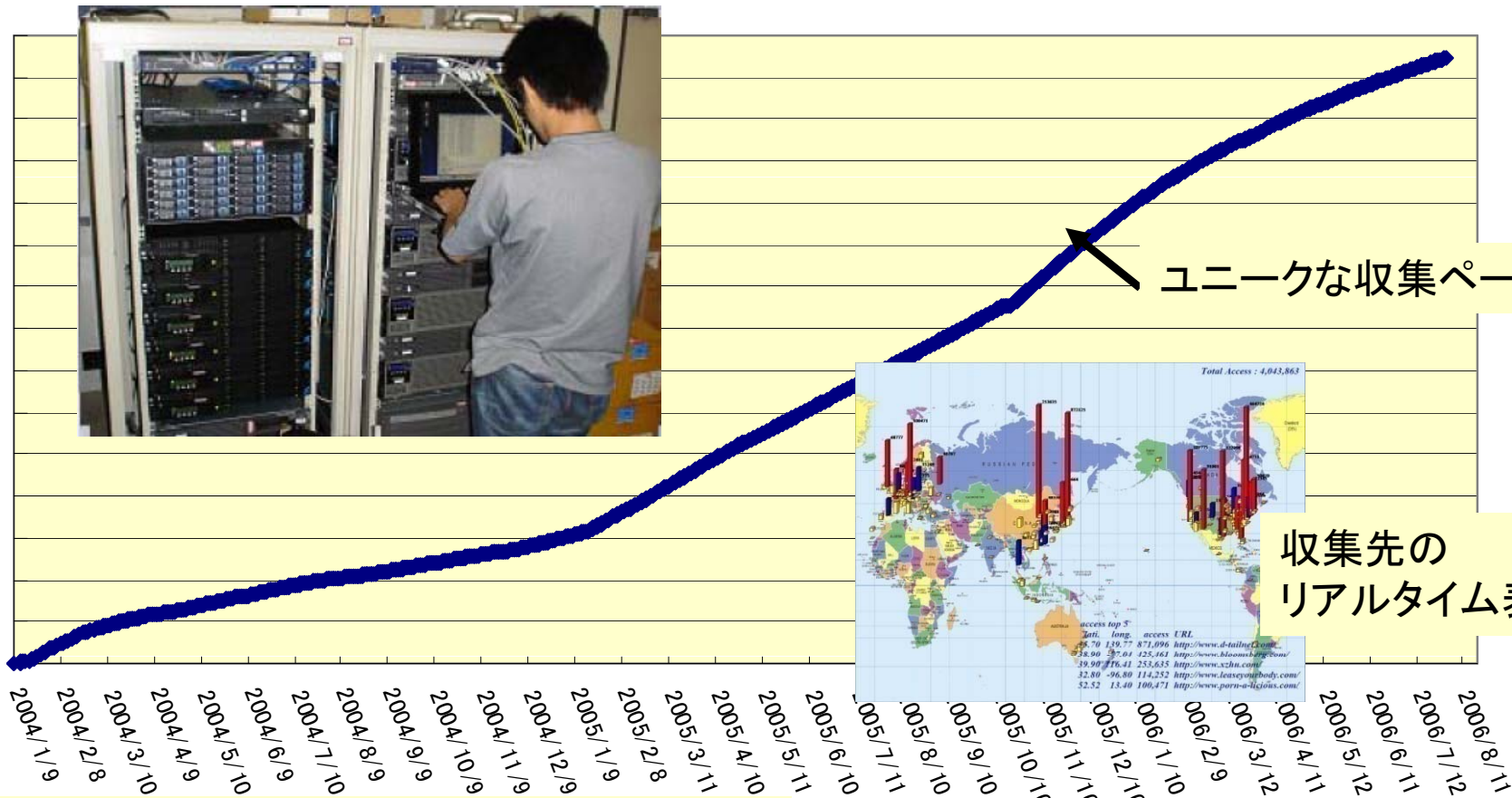
実績- WEBデータ収集

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拠点到マシン追加[合計80CPU]
06/09/01~現在 Japaneseページの再収集開始

ページ数

億

150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0



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

発見したWEBサーバ数：13,468万台

アクセス済：8,116万台

収集済：5,548万台

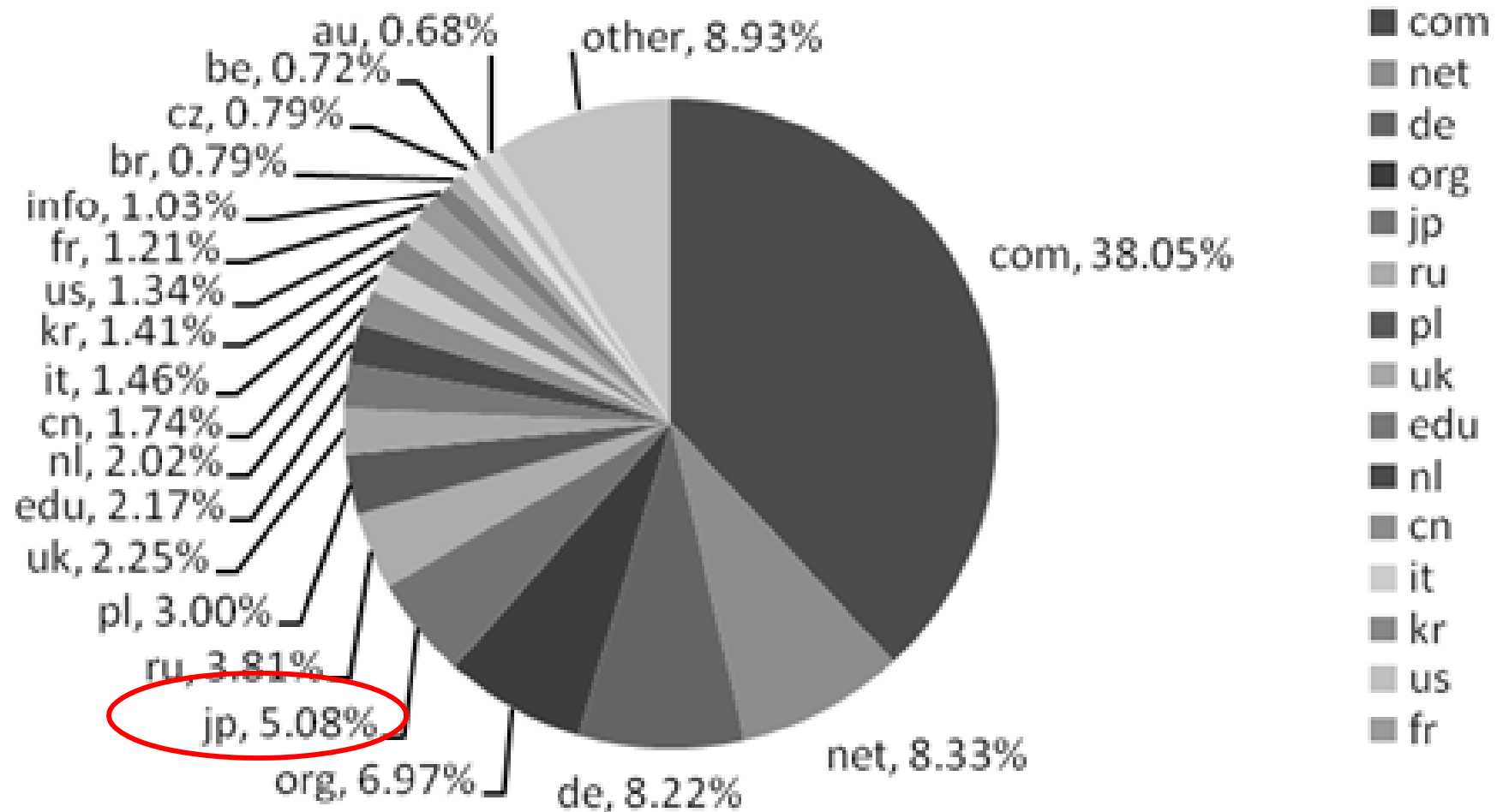
14,456,201,906ページ

アクセス
できず：
2,568万
台

robots.txtにより全
体がアクセス禁止：
256万台

WEBページのTLD分布

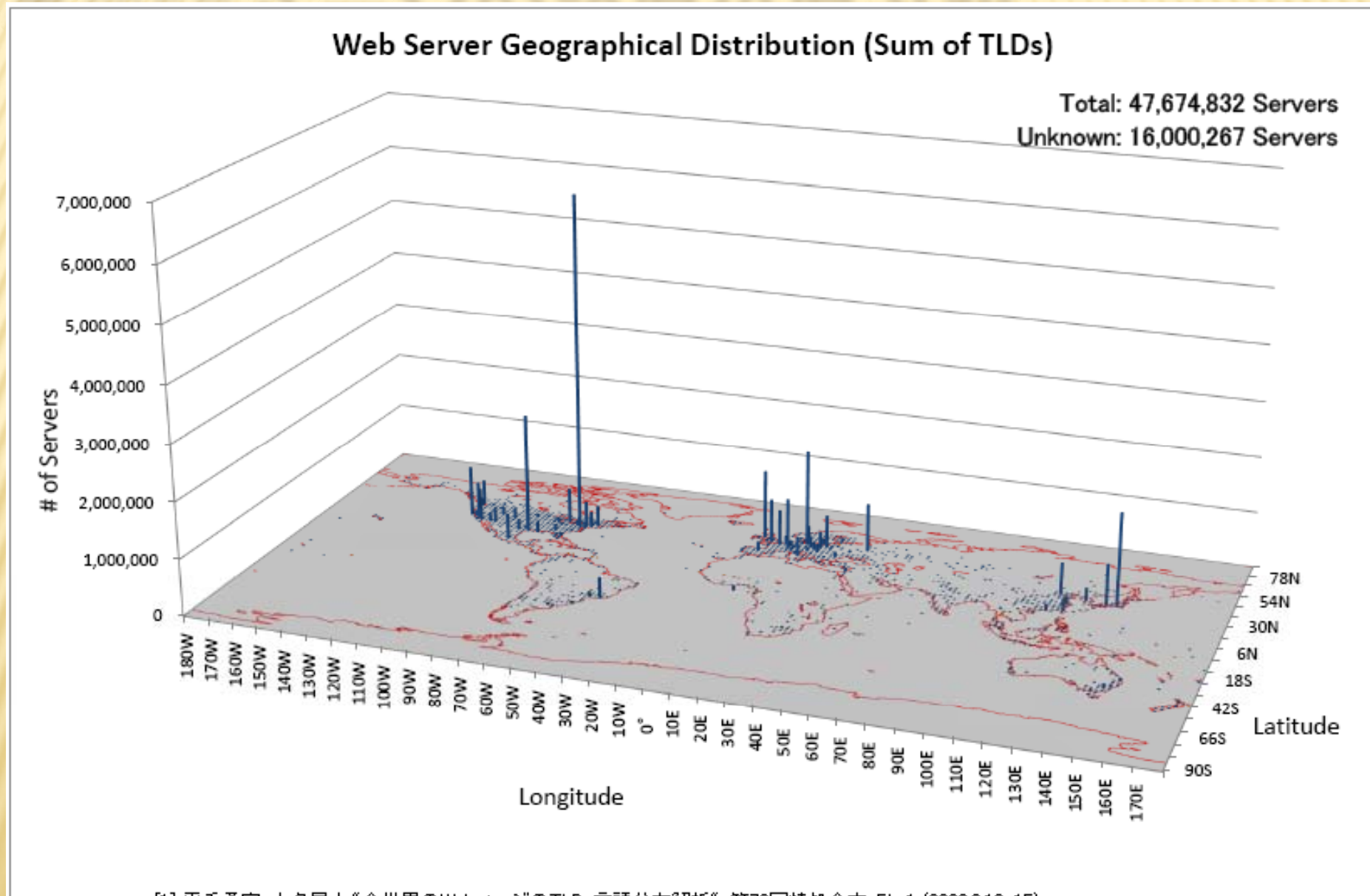
TLD Distribution of Web Pages



[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回情報処全大, 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)

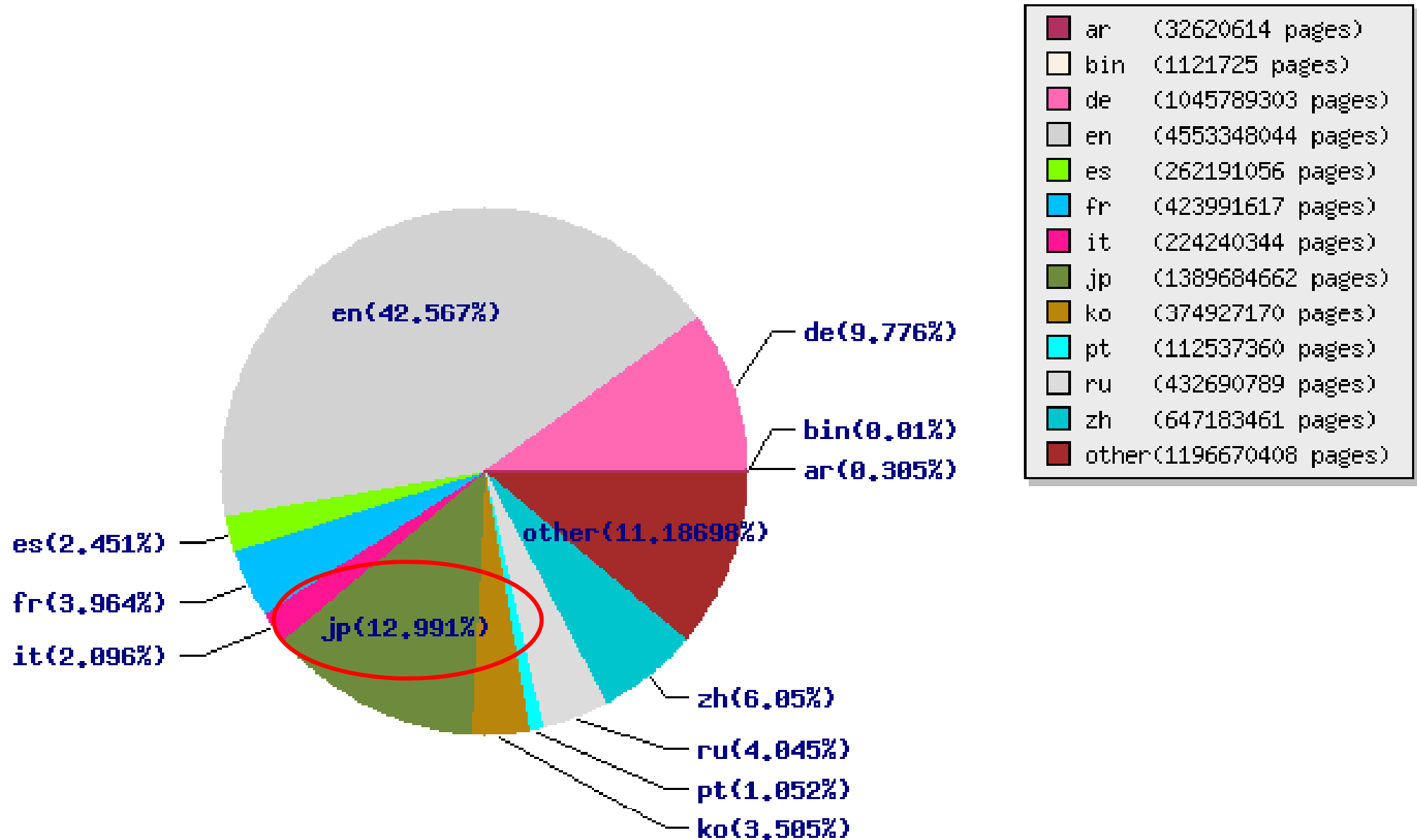
全WEBサーバの設置位置分布



[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] 堂 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)

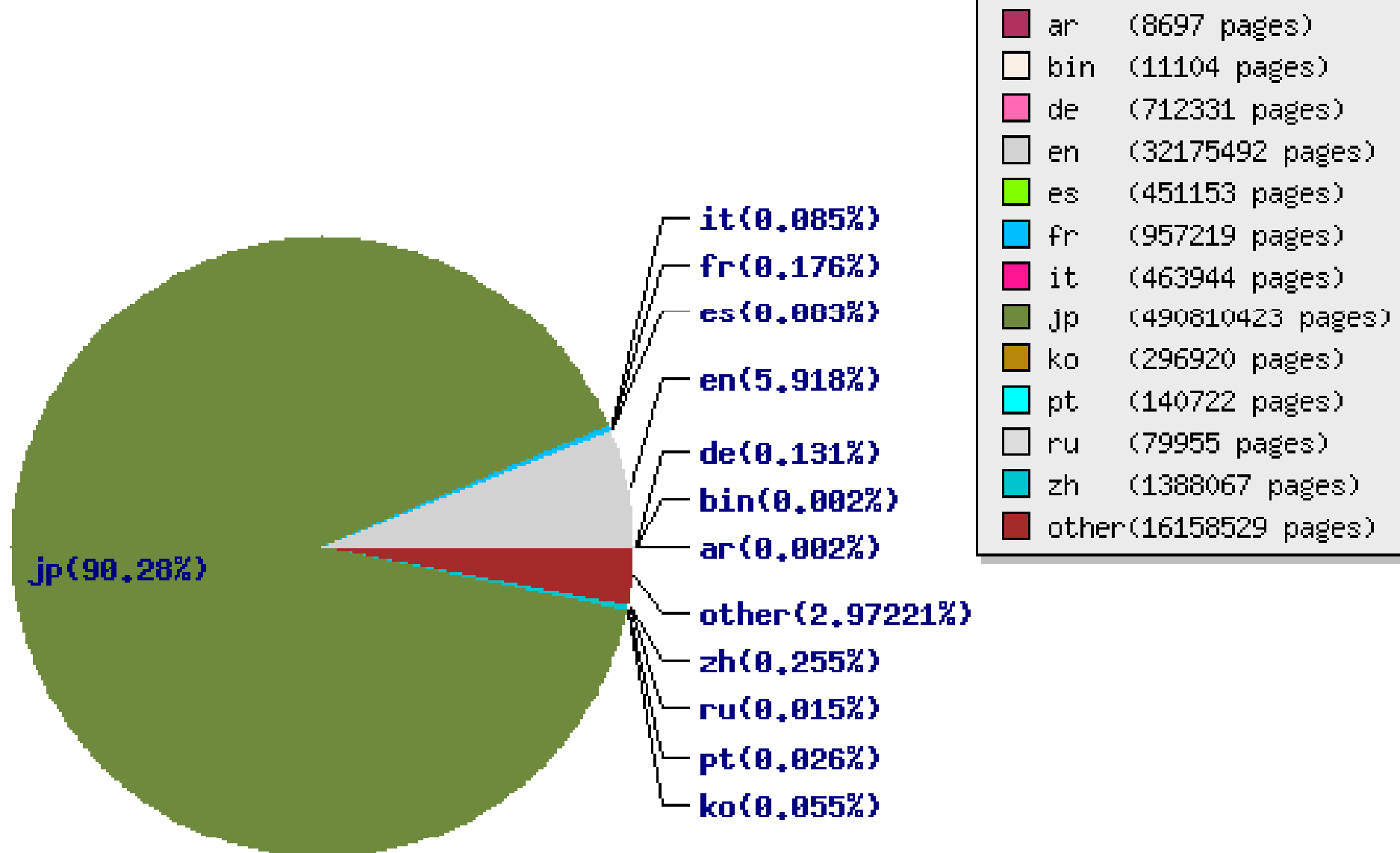
WEBページの言語分布(107億ページ)



[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)

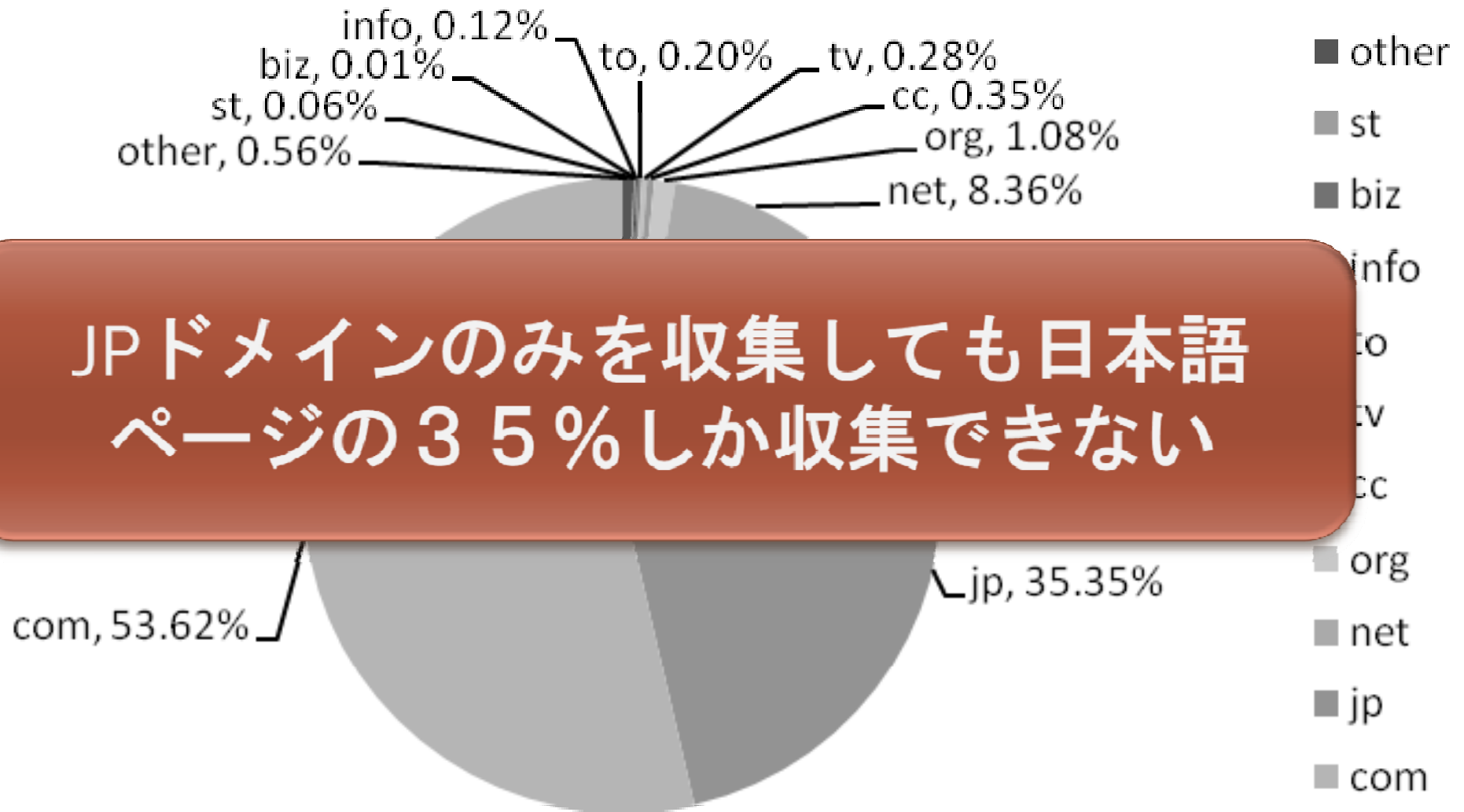
JPドメインに属する約5.4億ページの言語分布



[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] 堂 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)

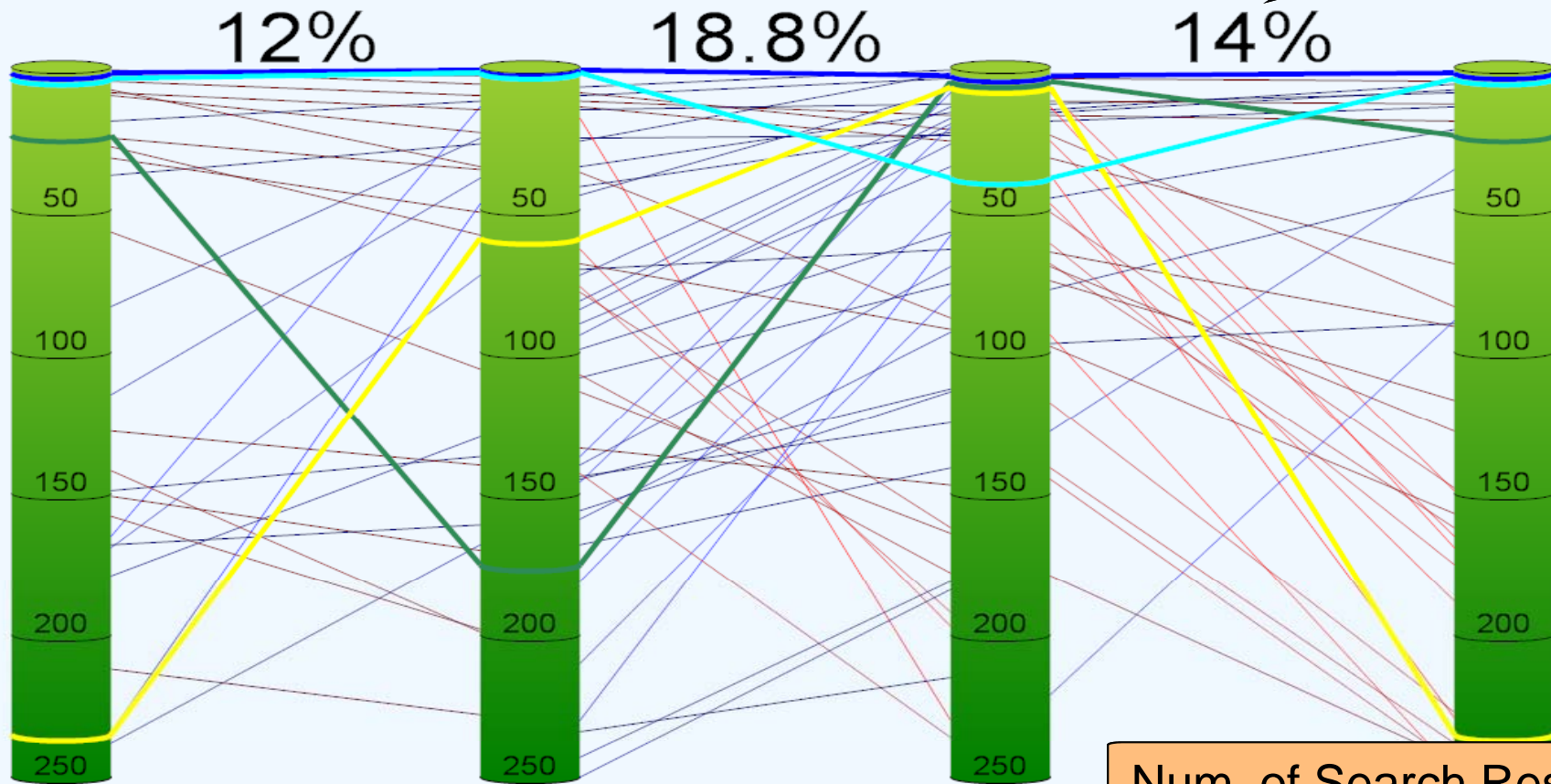
吉田泰明, 平手勇宇, 山名早人: “商用検索エンジンにランキングされたサイトのランク変動パターンの解析”, DEIM2009 (2009.3.8-10)

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

ランキング比較

Percentage of overlapped pages



Yahoo! Japan
1,200,000件

Google
415,000件

MSN
62,641件

Num. of Search Results

Yahoo! Japan
1,200,000件

Query

ランキング比較
(日本語設定を使用)

Move: Left Button
Rotate: [Shift] + Left Button
Zoom: [Alt] + Left Button
Change: Double Click

靖国問題 [検索] [一時停止] [再開] [観測データ]
[単語一覧更新]

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。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日にフィリッピン島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って、立派な花嫁さんになって、仕合わせになつたのを見届けたのですが、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ん」

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して生きていけるかどうかの大問題

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



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

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

Bibliograph



私たちが住みたい都市
伊東豊雄 × 藤田清一
松山善 × 上野千鶴子
八木好美 × 西川裕子
磯崎新 × 宮台真司
山本理顕 編

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け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 - 13:18:44

カテゴリ: お仕事で書いた文章 - トラックバック(3)

■靖国問題は「東京裁判の手打ち問題」(略して手打ち問題)と「憲法と歴史の振れ問題」(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、東京裁判の手打ちとは、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正だと当たり前。この不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 - MIYADAI.com Blog



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

ログインまたはアカウント作成

本文 ノート 編集 履歴

靖国神社問題

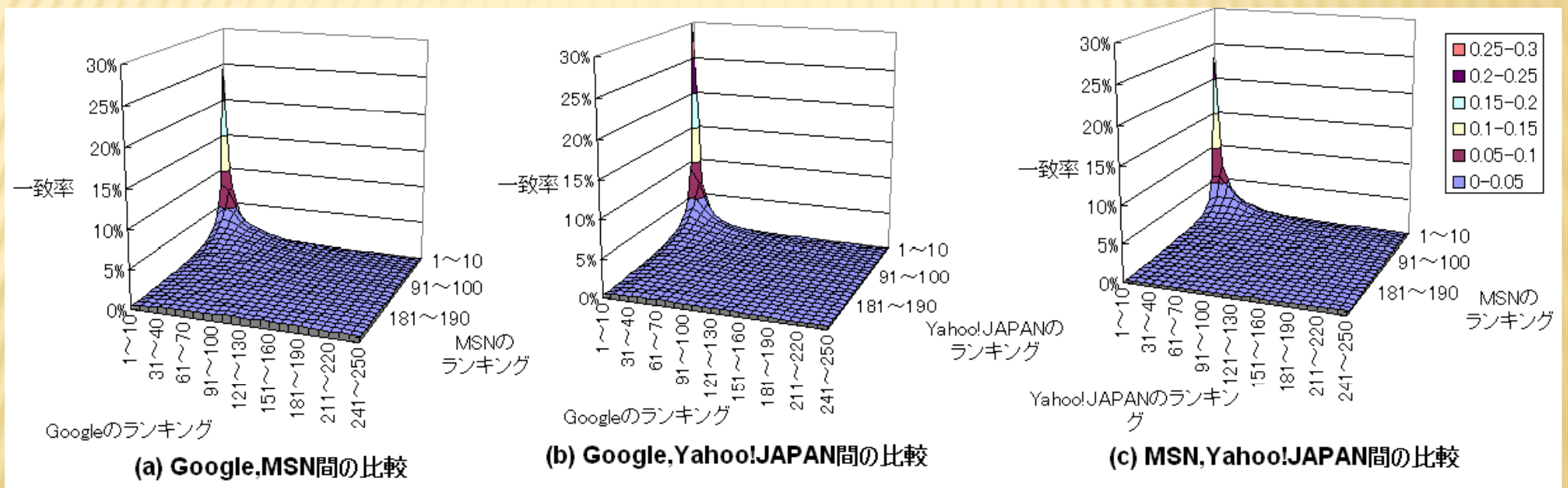
出典: フリー百科事典『ウィキペディア (Wikipedia)』
(首相、大臣の靖国神社参拝問題 から転送)

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反応についての記述

検索エンジン間のランキング一致度



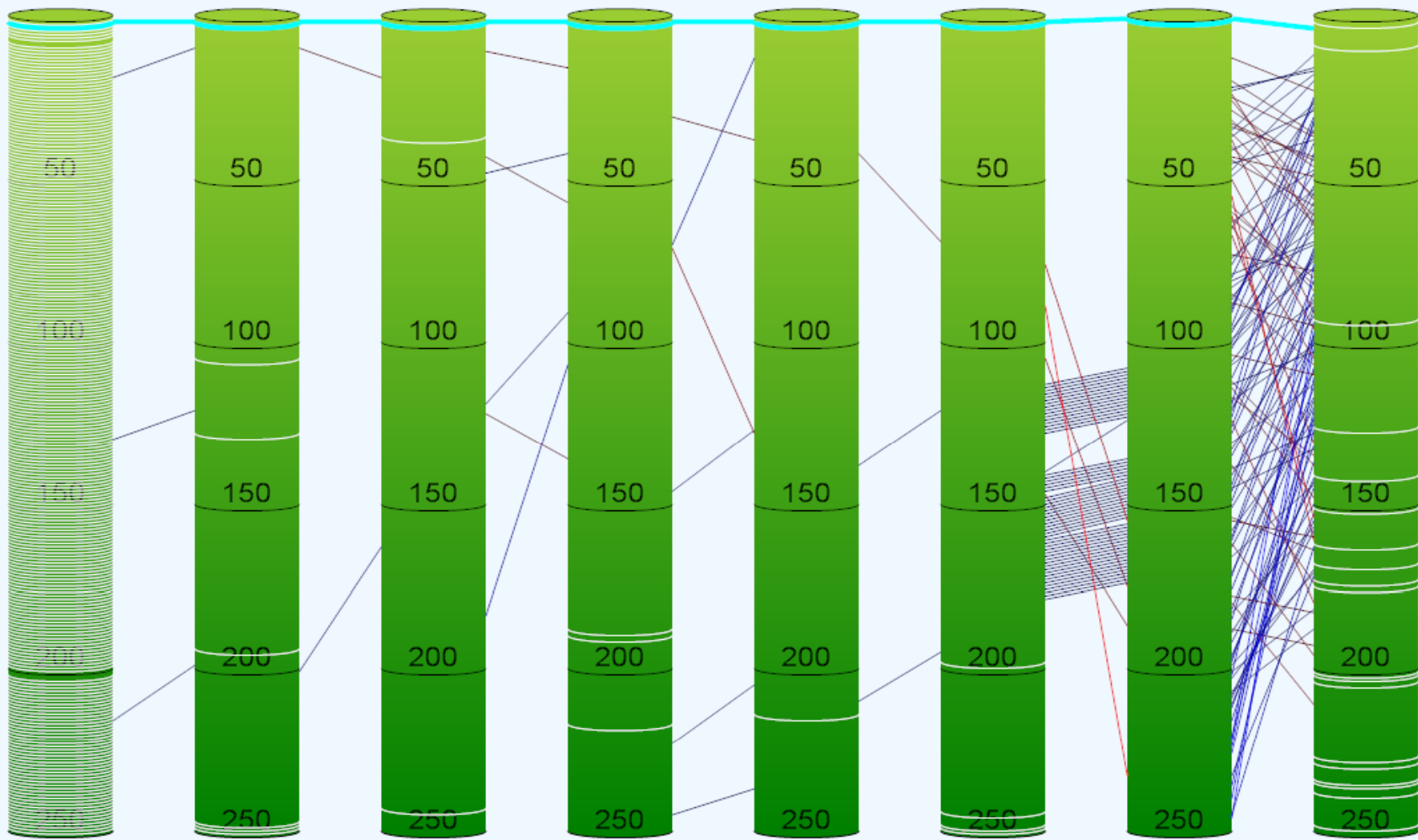
Yasuaki Yoshida(M1), Takanori Ueda, Takashi Tashiro, Yu Hirate, Hayato Yamana: "What's going on in search engine rankings?", Proc. of the 2008 IEEE International Symposium on Mining and Web (2008.3.25-28)

(c) 2002-2011 All Rights Reserved Hayato YAMANA@waseda Univ.

Yahoo! Japan

Yahoo! Japan

5



Sep.20

Sep.21

Sep.22

Sep.23

Sep.24

Sep.25

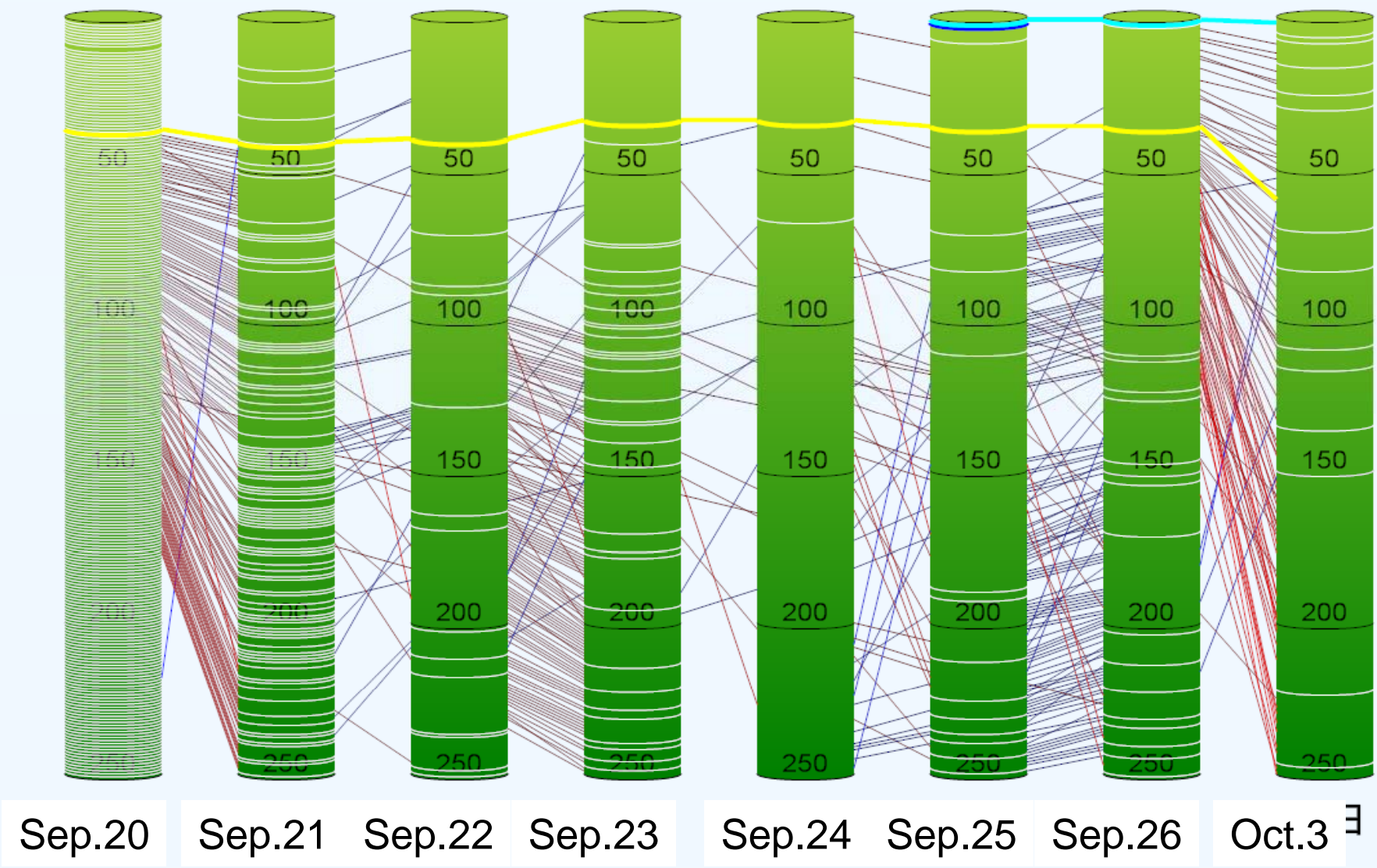
Sep.26

Oct.3 ^日

Transition of the ranking

Google
5

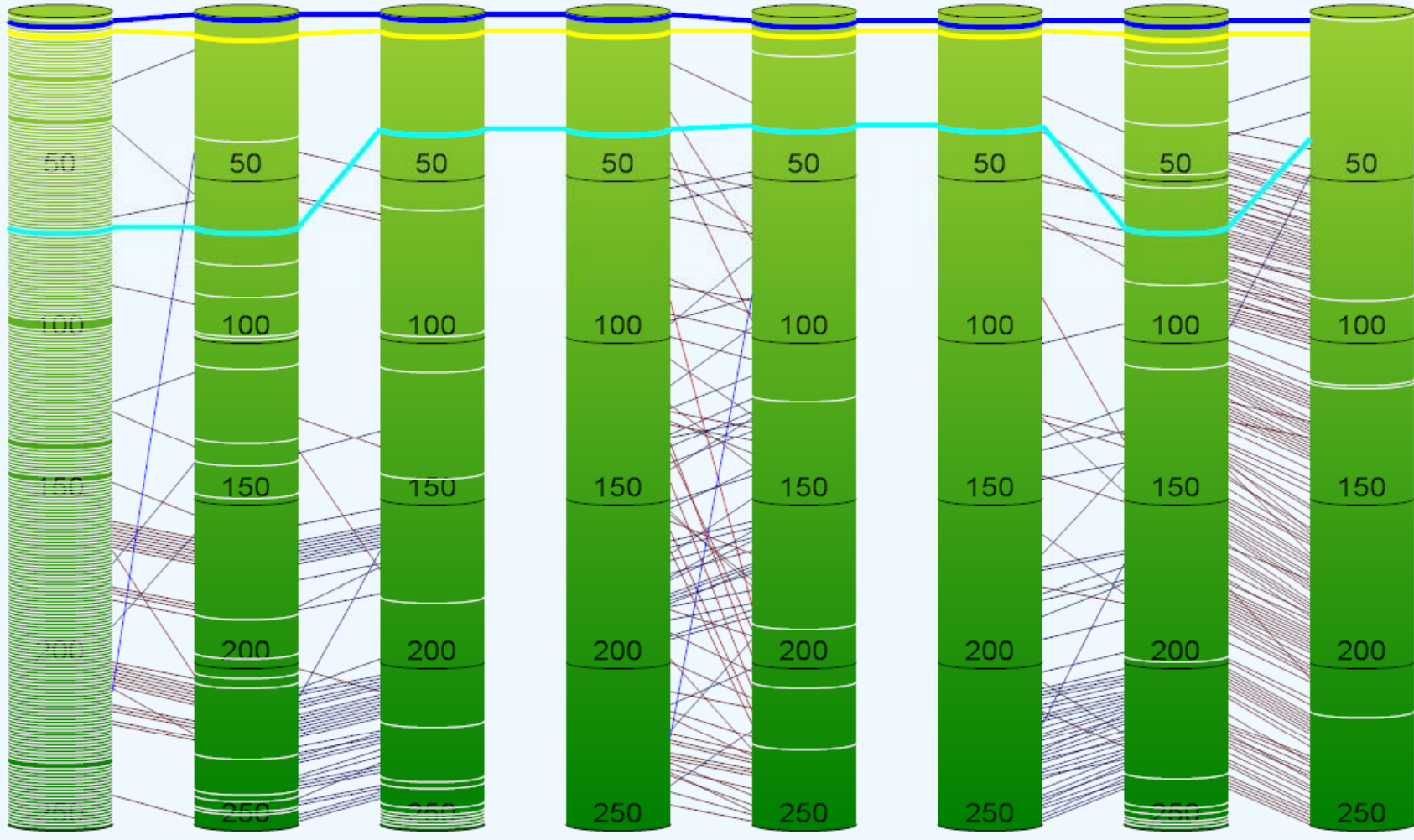
Google



Transition of the ranking

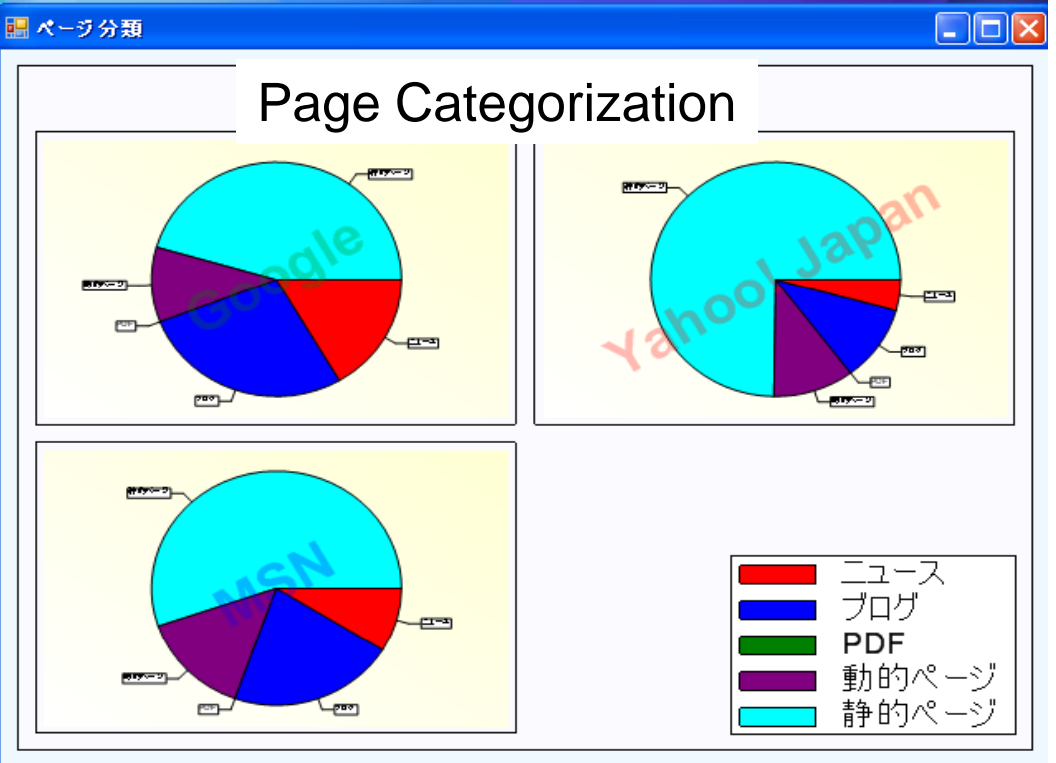
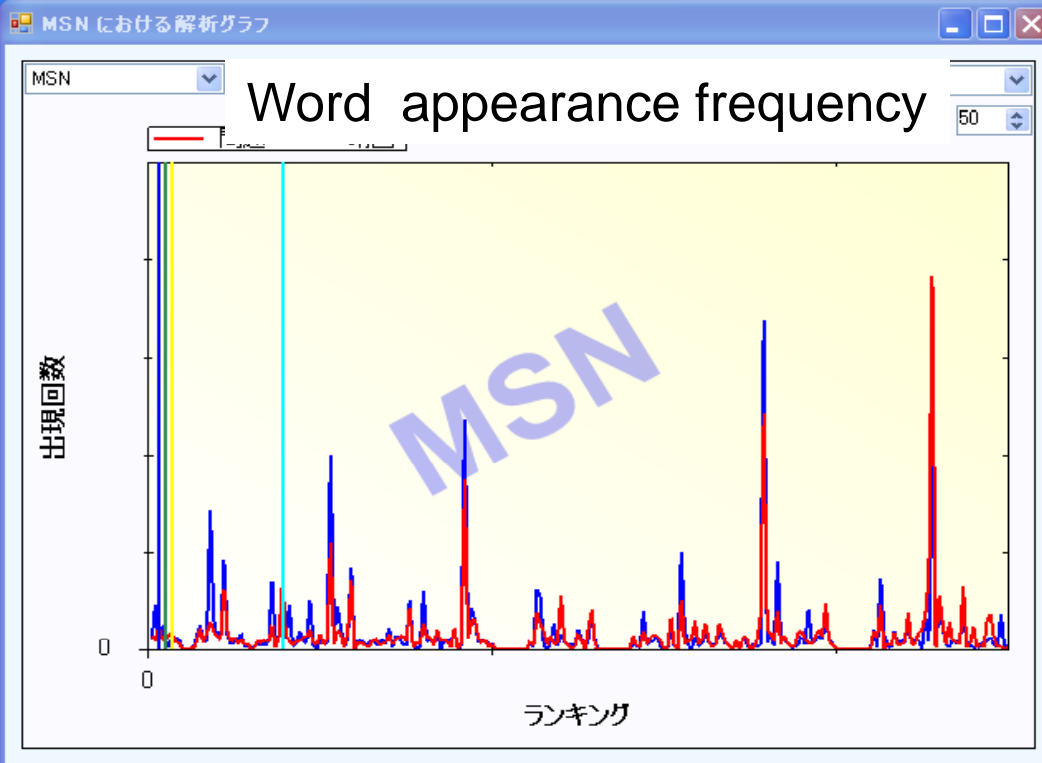
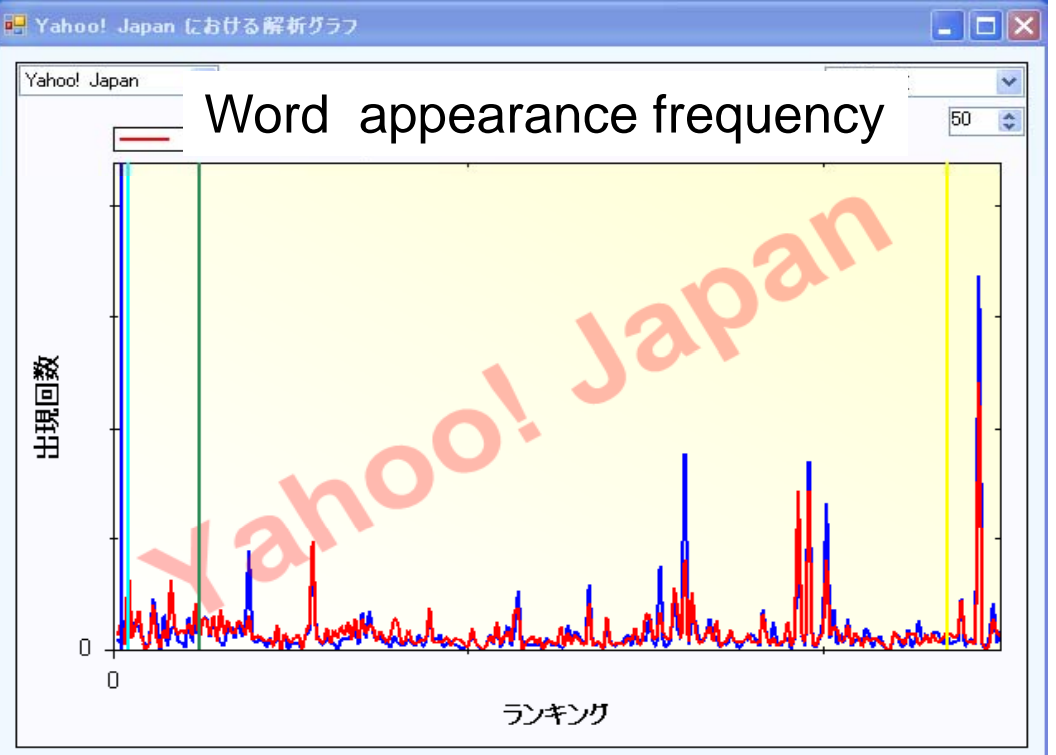
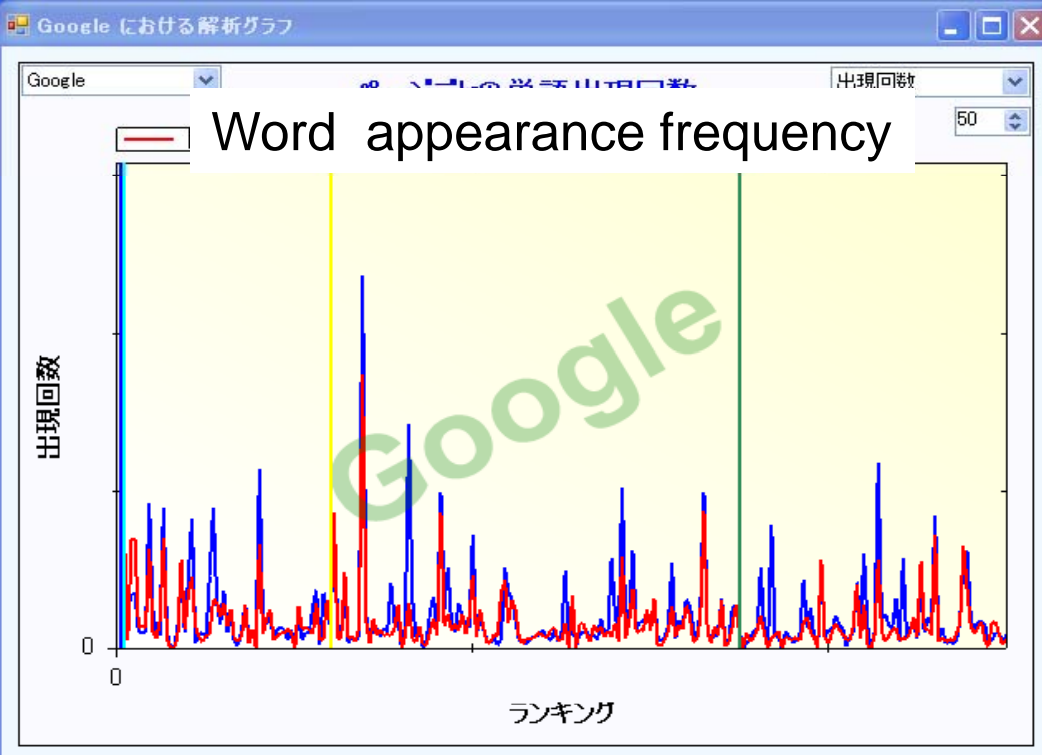
MSN
5

MSN

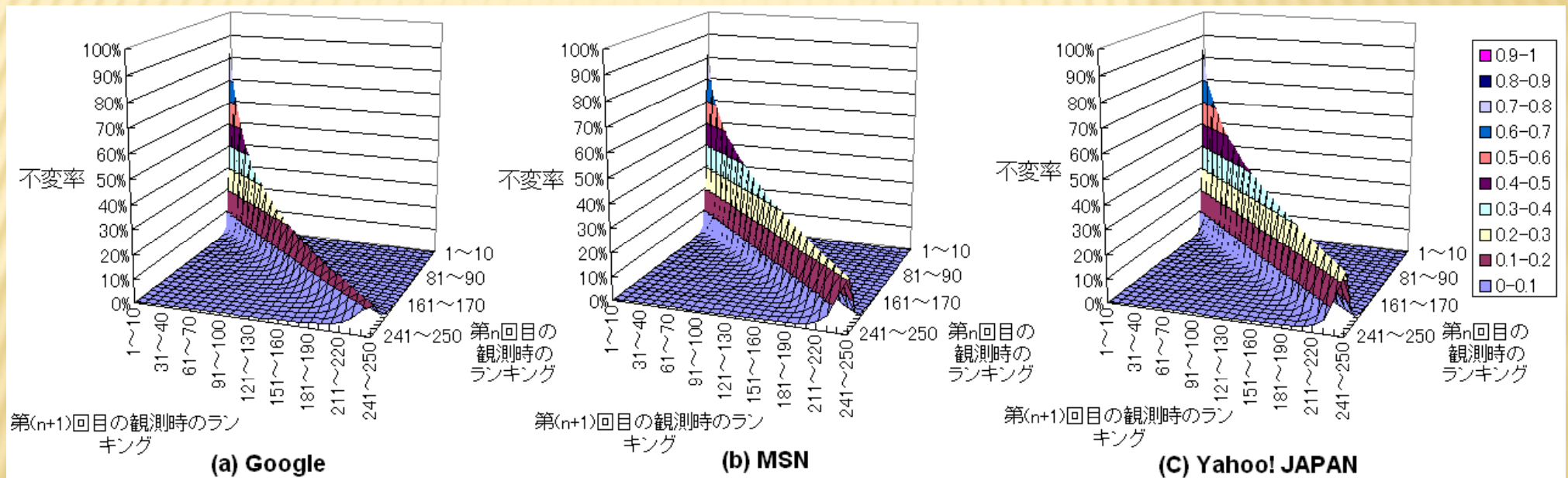


Sep.20 Sep.21 Sep.22 Sep.23 Sep.24 Sep.25 Sep.26 Oct.3 ^日

Transition of the ranking



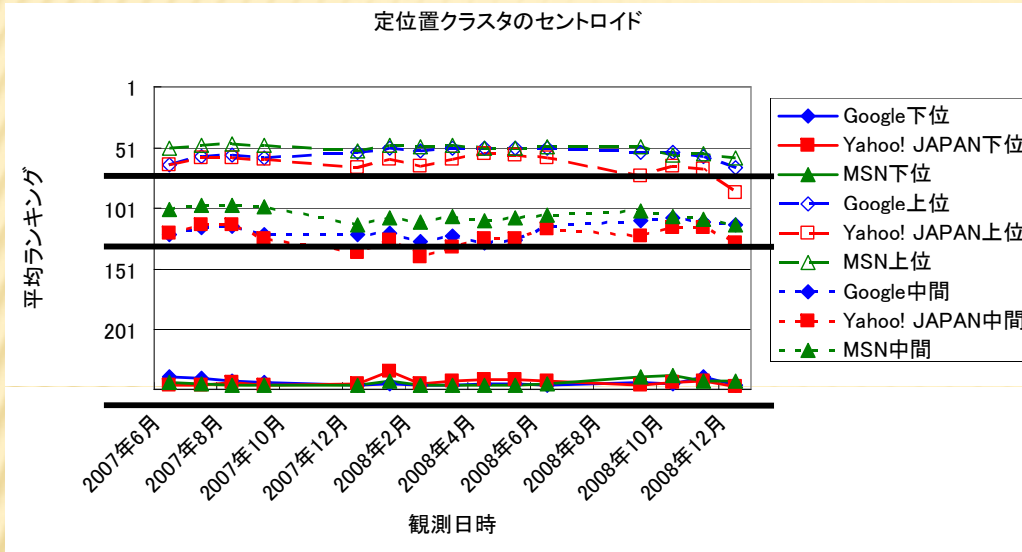
約10日後のランキング変動



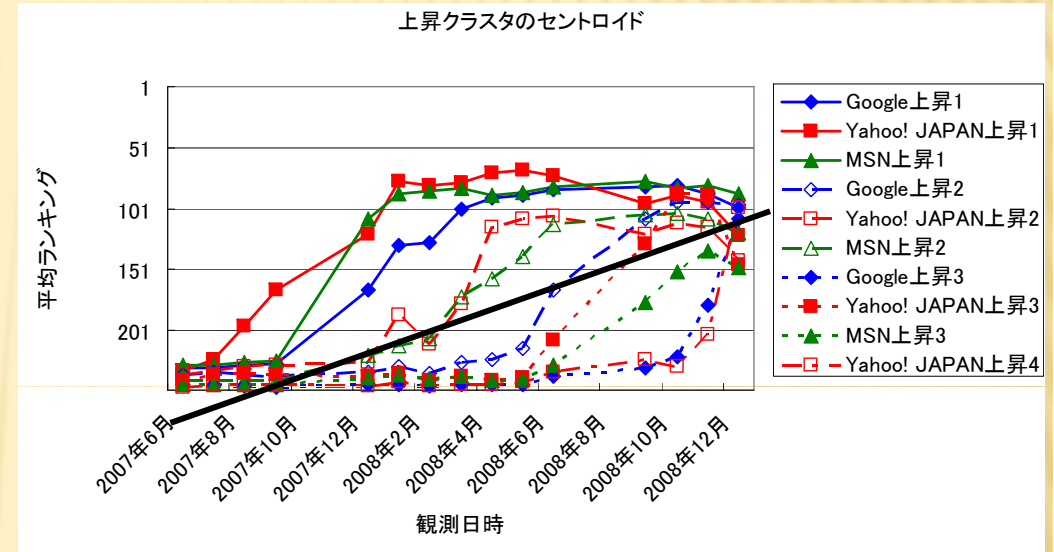
Yasuaki Yoshida(M1), Takanori Ueda, Takashi Tashiro, Yu Hirate, Hayato Yamana: "What's going on in search engine rankings?", Proc. of the 2008 IEEE International Symposium on Mining and Web (2008.3.25-28)

サイトのランク変動パターン

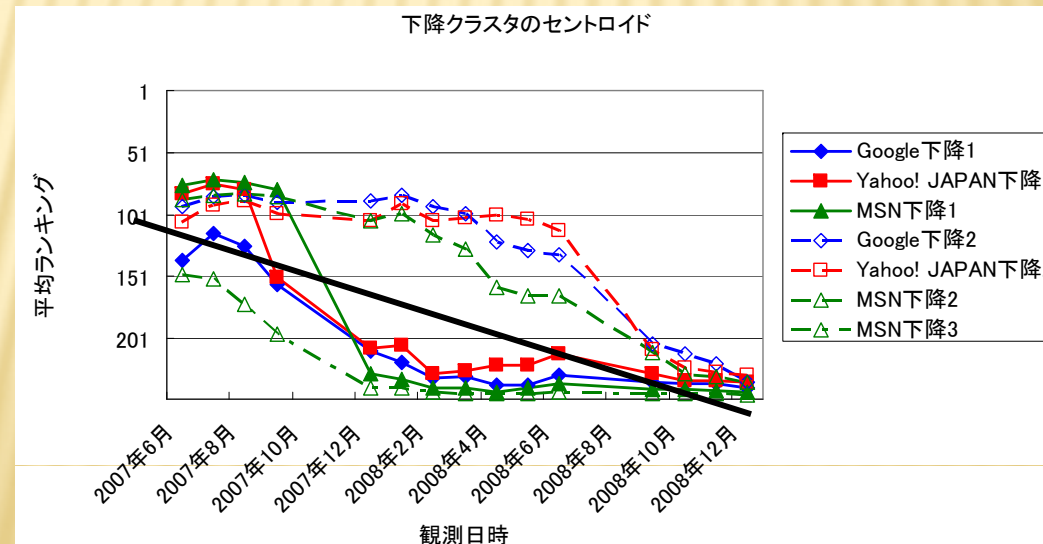
・ 定位置パターン



・ 上昇パターン

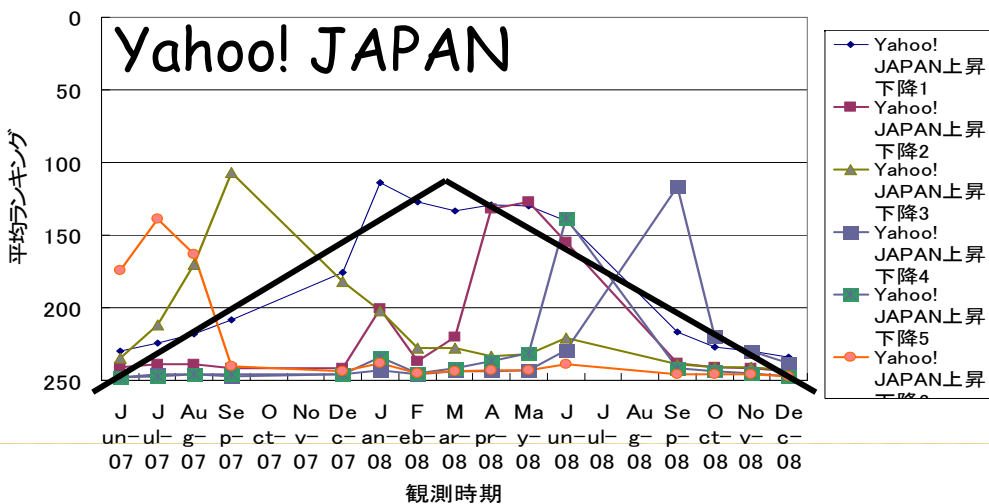
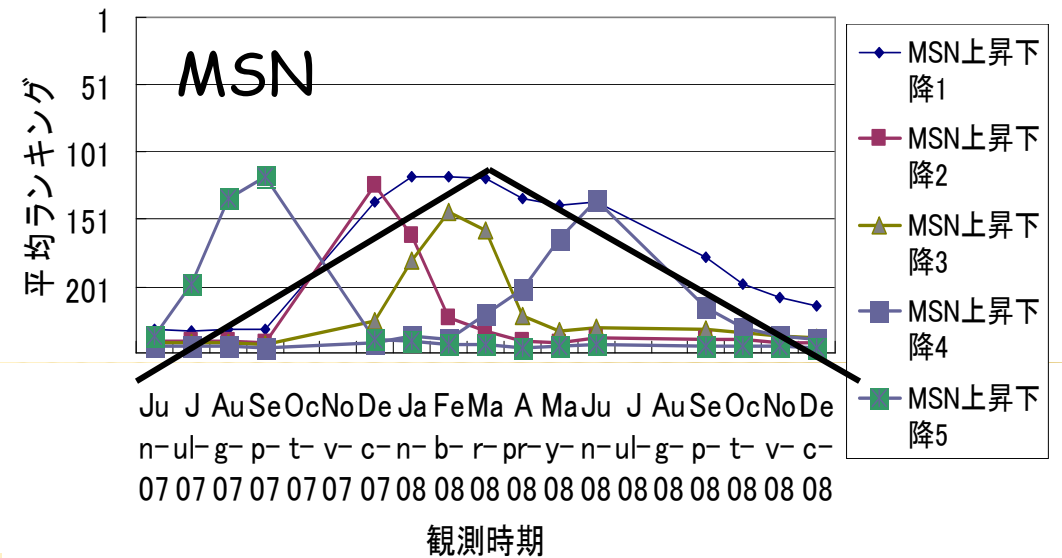
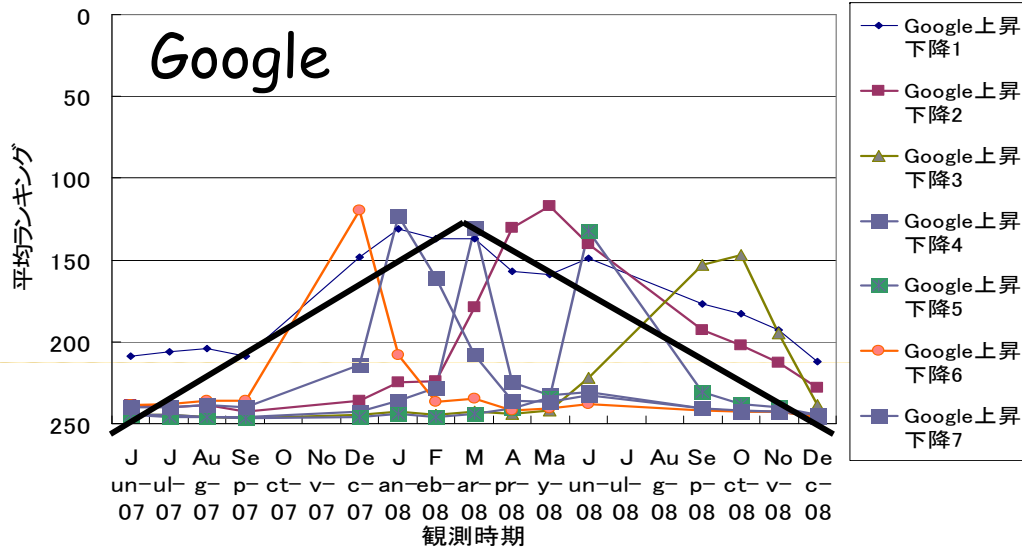


・ 下降パターン



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

4. 検索エンジンの信頼性 — 検索結果数 —

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、曾根広哲、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、上田高德、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？

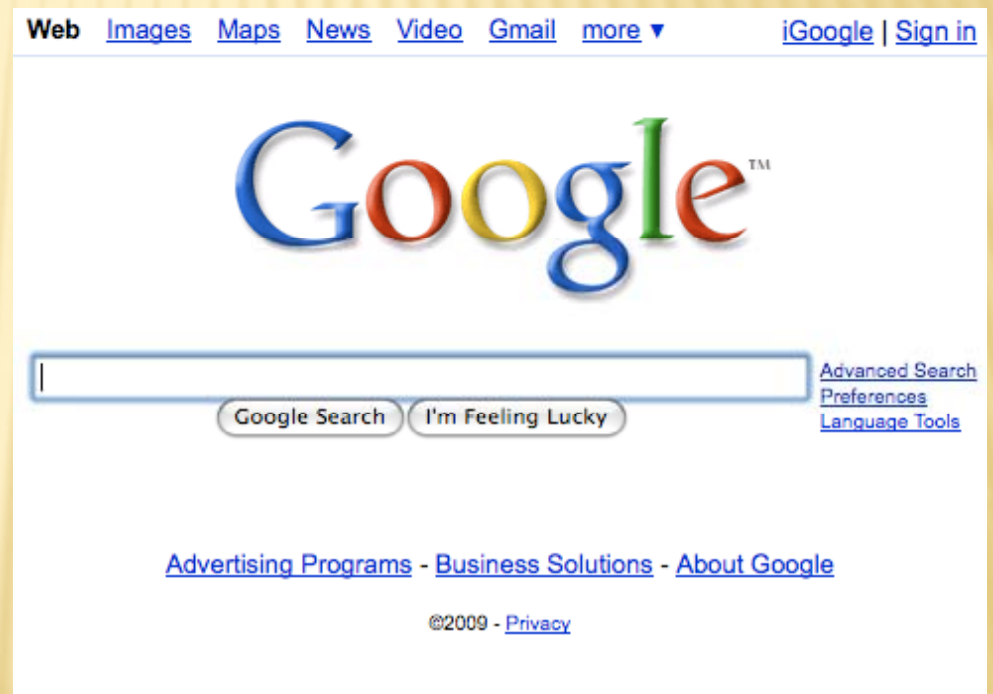


Barack Obama

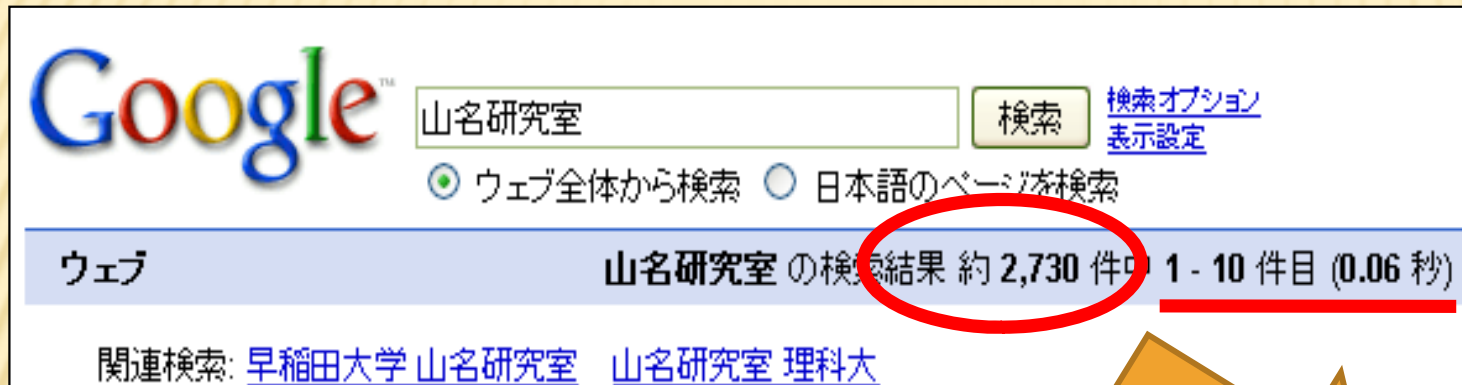
Results 1 - 10 of about 104,000,000 for Barack Obama. (0.10 seconds)

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...

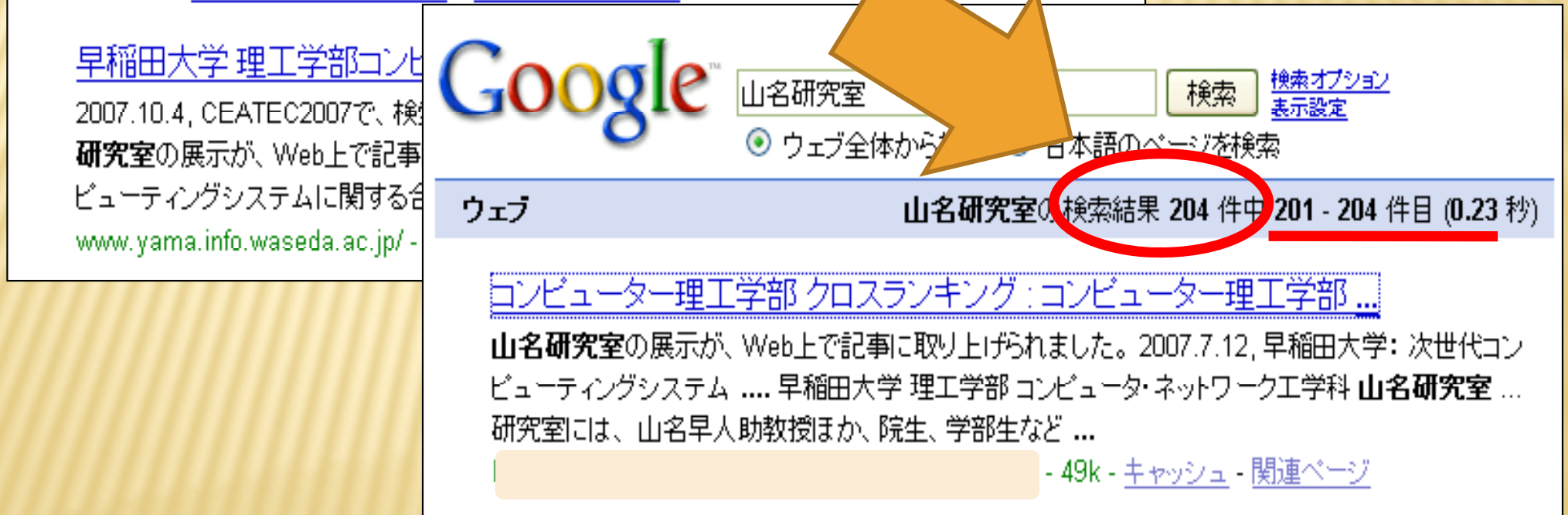


Google 山名研究室 検索 検索オプション 表示設定

ウェブ全体から検索 日本語のページを検索

ウェブ 山名研究室 の検索結果 約 2,730 件中 1 - 10 件目 (0.06 秒)

関連検索: [早稲田大学 山名研究室](#) [山名研究室 理科大](#)



Google 山名研究室 検索 検索オプション 表示設定

ウェブ全体から検索 日本語のページを検索

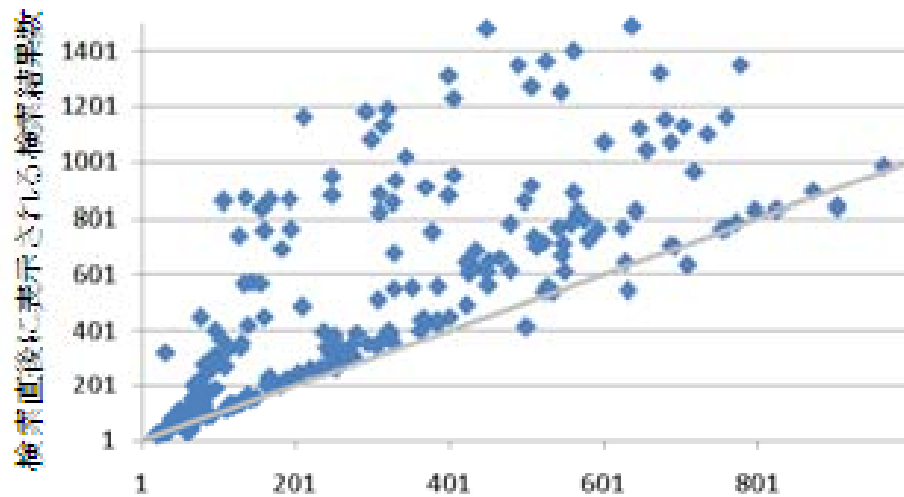
ウェブ 山名研究室 の検索結果 204 件中 201 - 204 件目 (0.23 秒)

[コンピュータ工学部 クロスランキング: コンピュータ工学部 ...](#)

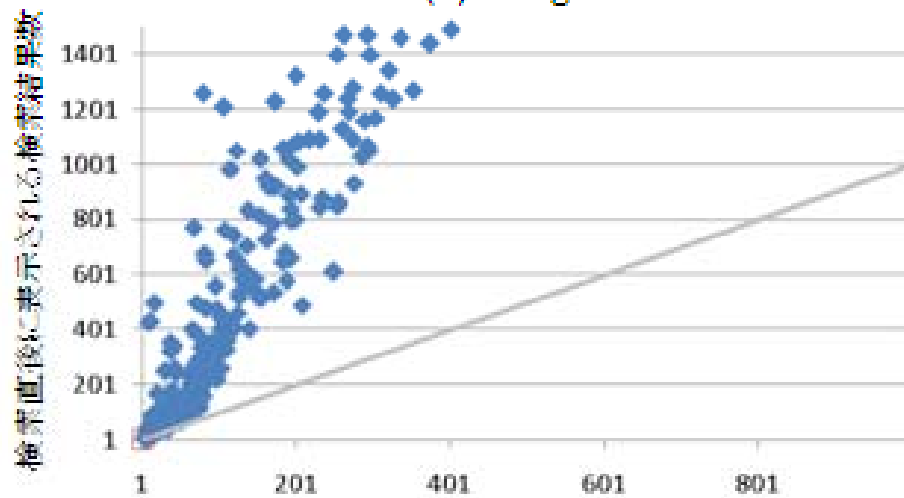
山名研究室の展示が、Web上で記事に取り上げられました。2007.7.12, 早稲田大学: 次世代コンピューティングシステム ... 早稲田大学 理工学部 コンピュータ・ネットワーク工学科 山名研究室 ... 研究室には、山名早人助教授ほか、院生、学部生など ...

- 49k - [キャッシュ](#) - [関連ページ](#)

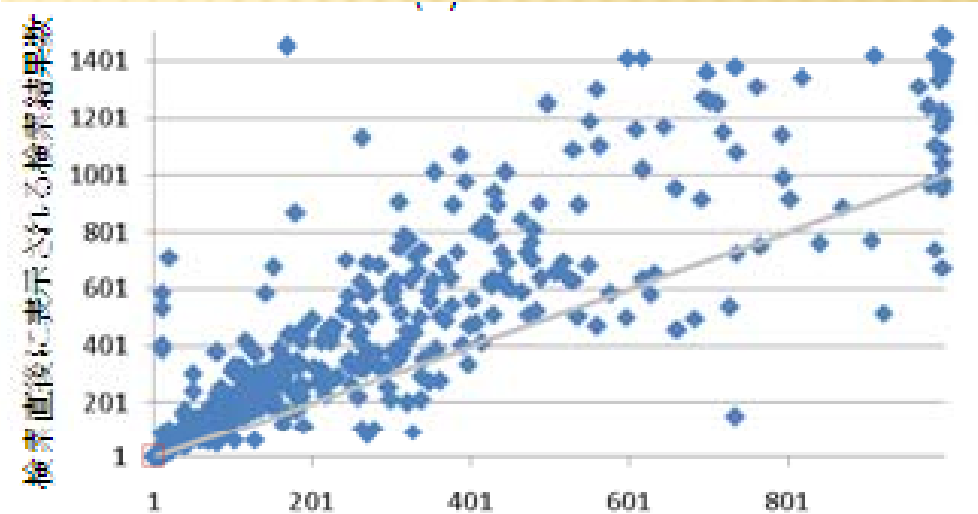
検索結果ヒット数と実際に得られる数



実際に取得できた件数
(a) Google



実際に取得できた件数
(b) Yahoo! JAPAN



実際に取得できた件数
(c) MSN

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

ヒット数が増える3つのケース

Case1: 「検索」ボタンを何度も押した場合

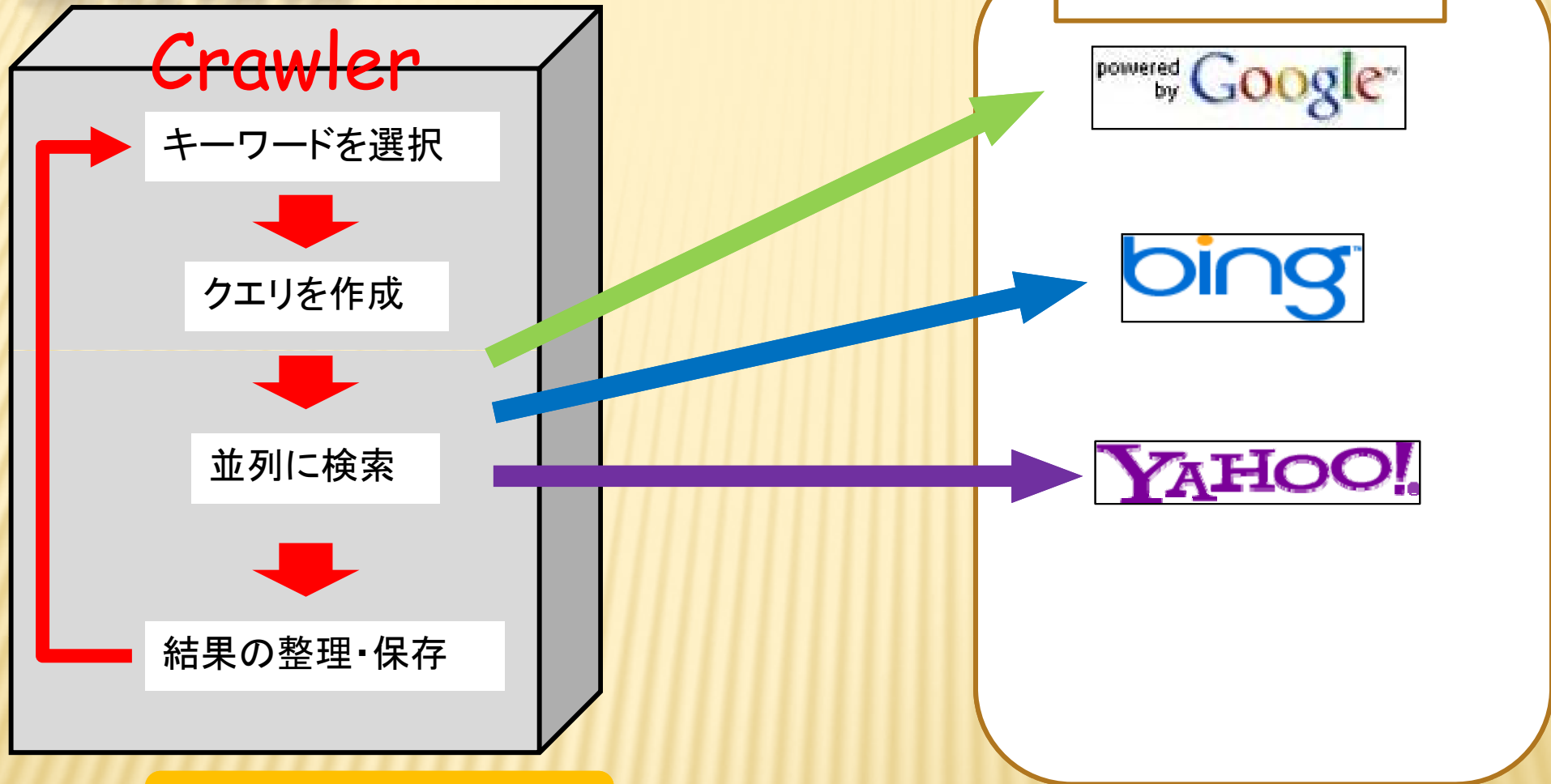
Case2: 「次へ」ボタンを何度も押した場合

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

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

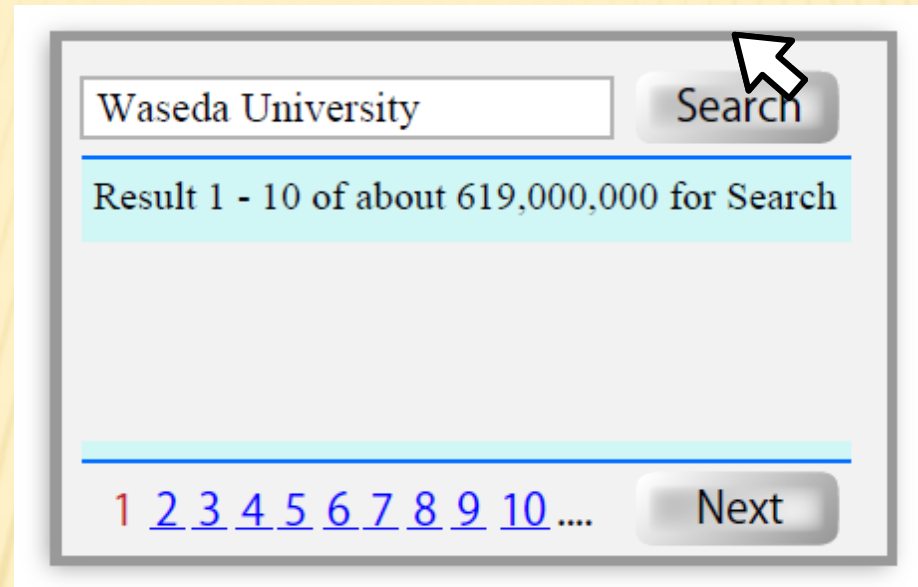
...

A : 991-1000

検証事項

- ×  ,  ,  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

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

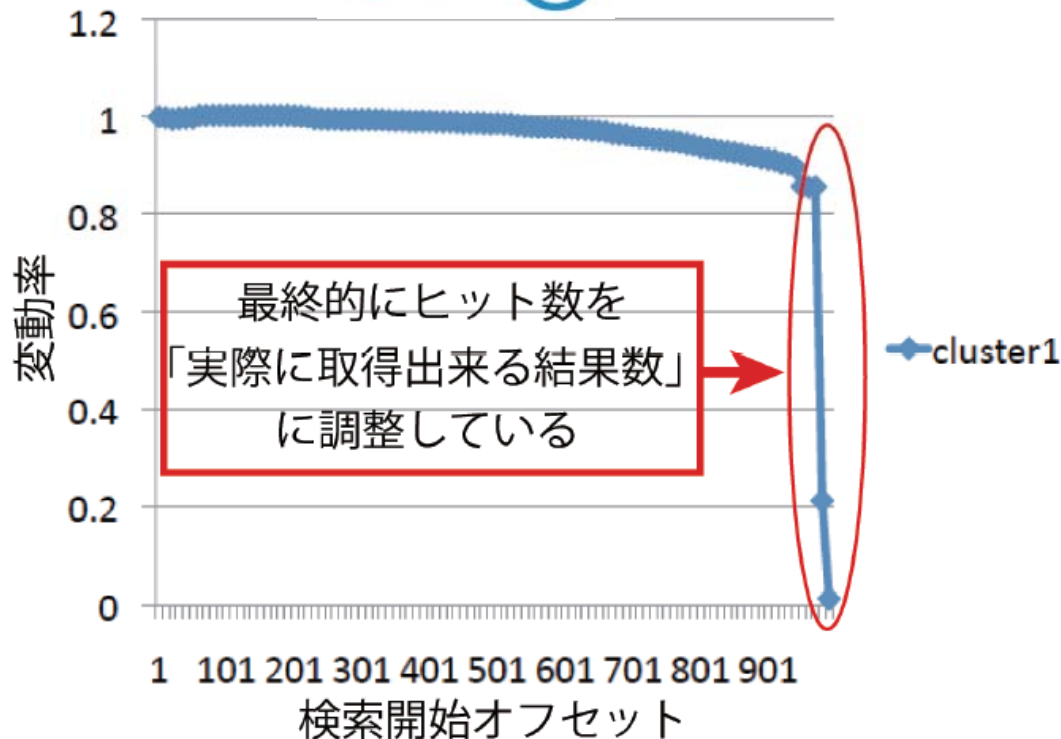
$$DV = \left\{ \frac{HitCount(1,10)}{HitCount(1,10)}, \frac{HitCount(11,20)}{HitCount(1,10)}, \dots, \frac{HitCount(991/1000)}{HitCount(1/10)} \right\}$$

- × $HitCount(k, k+9)$: 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

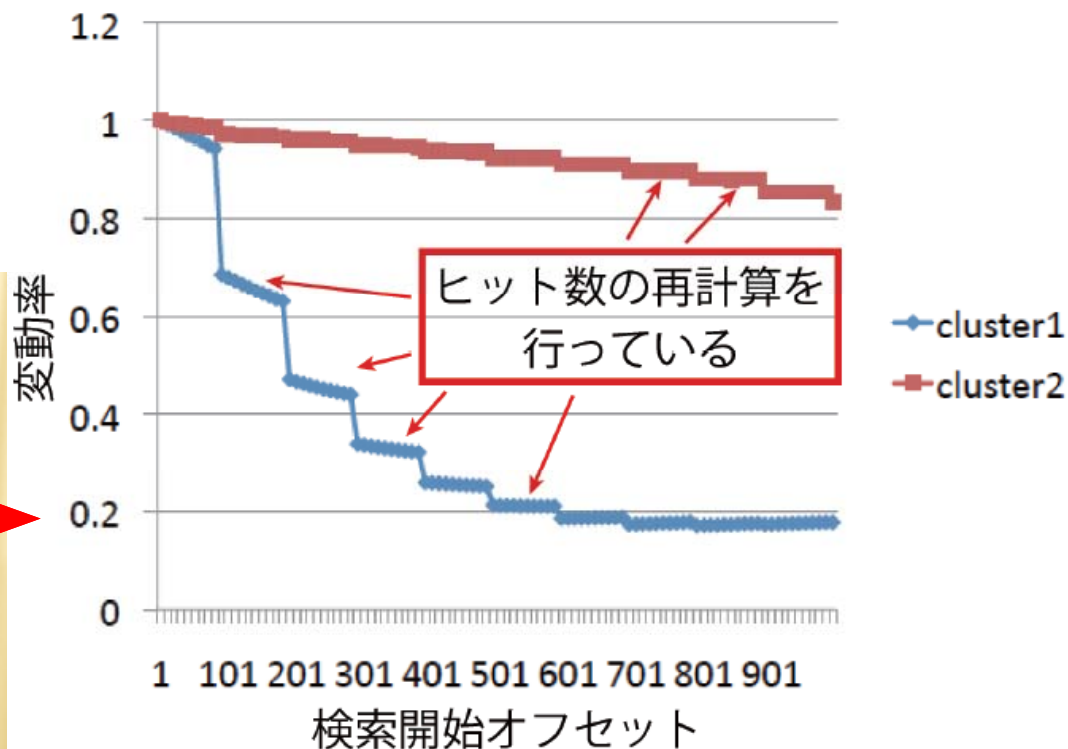
+ DVを全てのクエリに対して求め, クラスタリング

CASE2: 検証結果

ークラスタリング



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

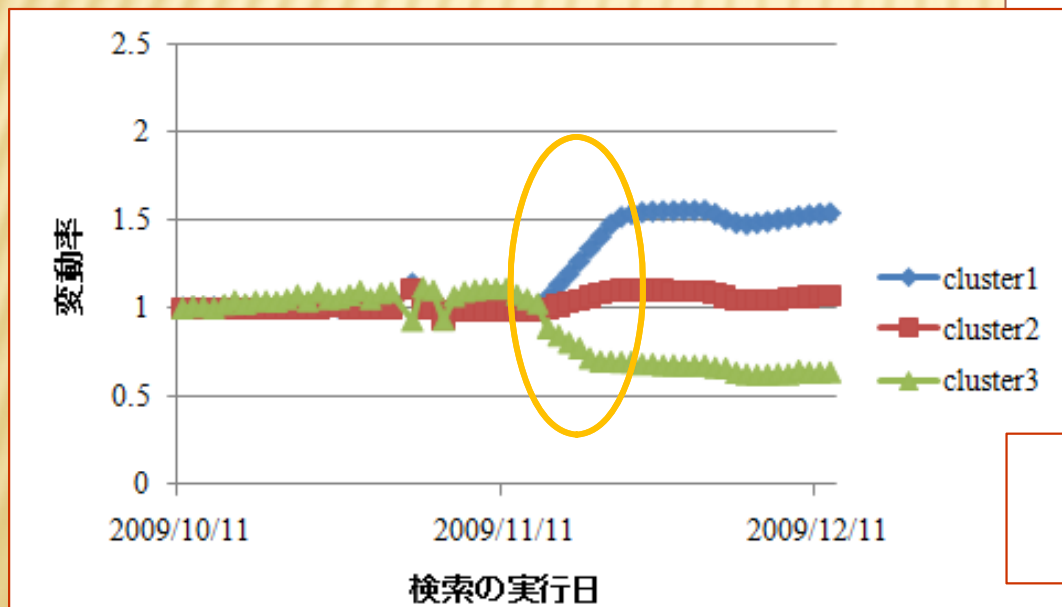
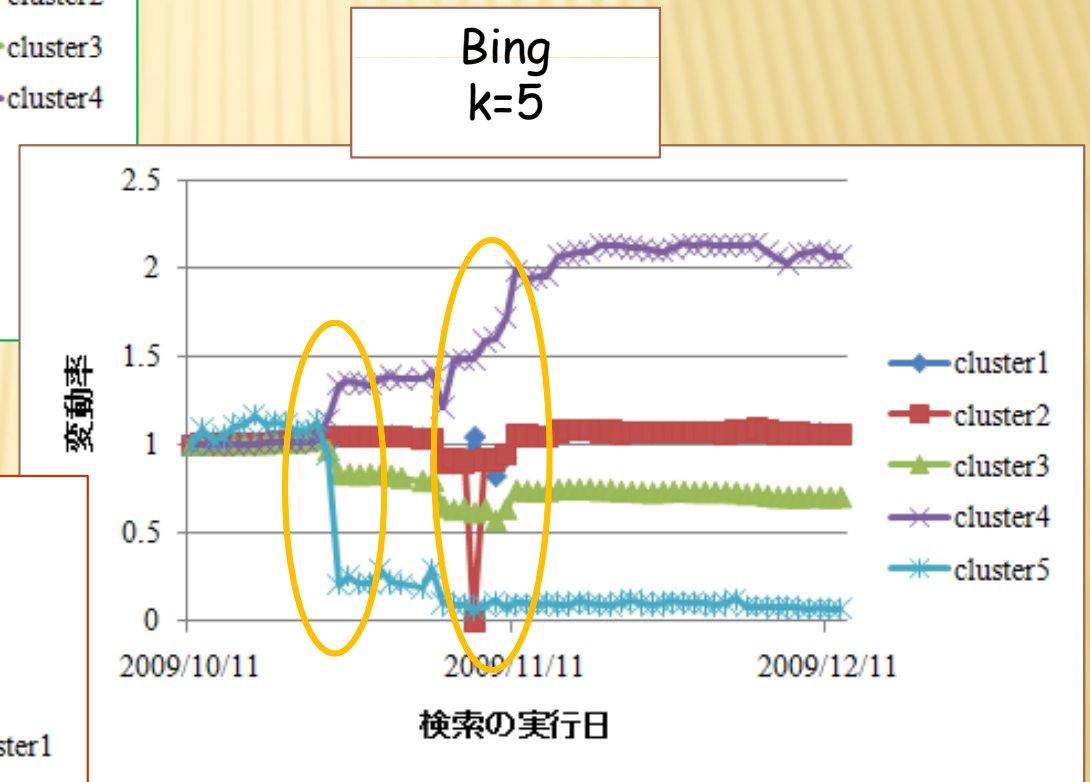
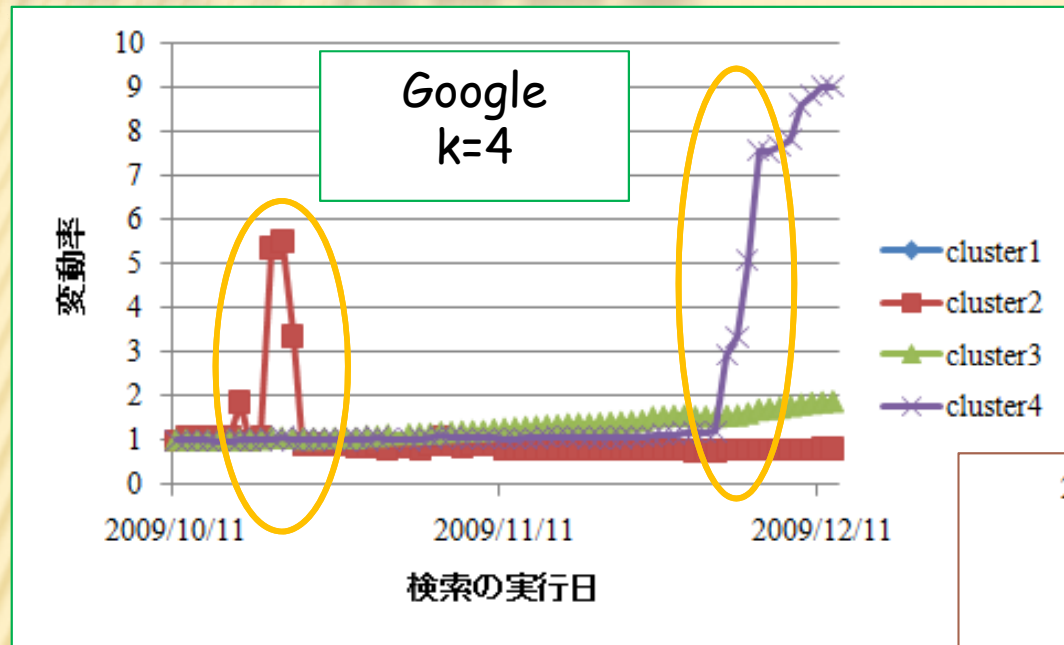


ヒット数は、検索開始オフセットが100変化するごとに大きく変動

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
 $\{HitCount(10/11), HitCount(10/12), \dots, HitCount(12/12)\}$
- 10/11のヒット数に対する変動率ベクトルを算出
$$\left\{ \frac{HitCount(10/11)}{HitCount(10/11)}, \frac{HitCount(10/12)}{HitCount(10/11)}, \dots, \frac{HitCount(12/12)}{HitCount(10/11)} \right\}$$
- 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



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なので、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

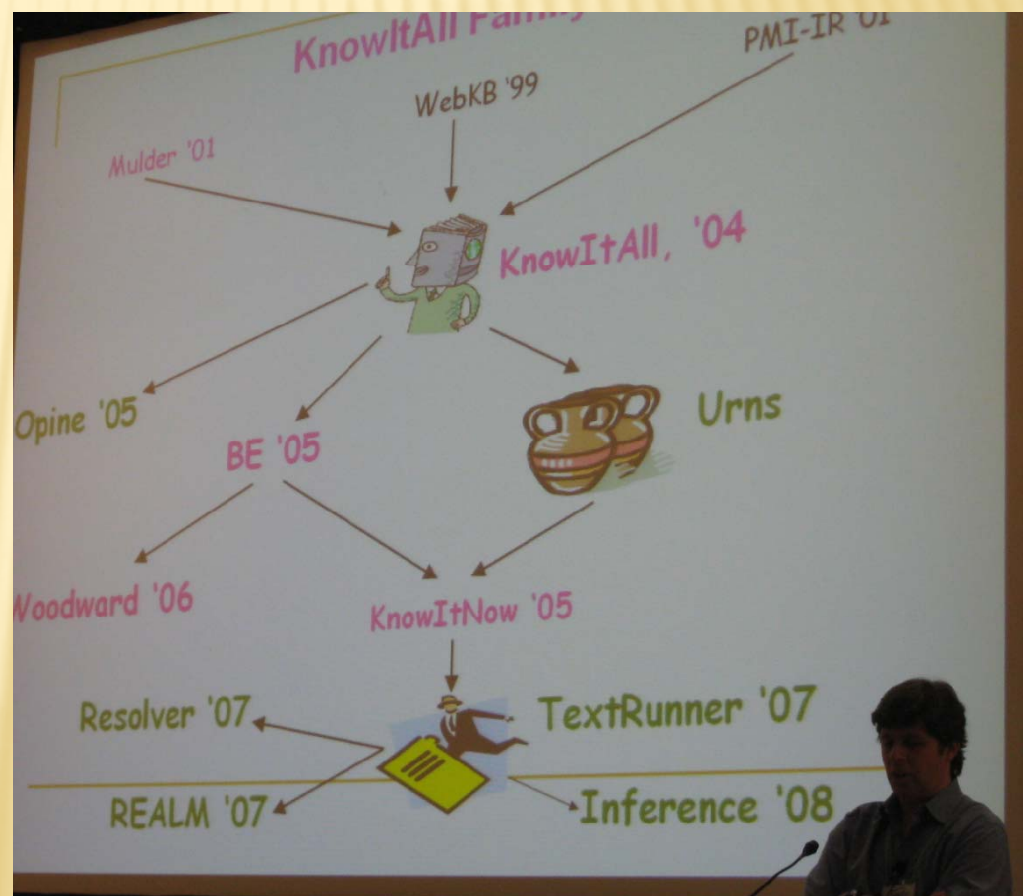
これを区別できますか？

- × Book by children
- × Book for children
- × Book of

① TEXT 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
<ul style="list-style-type: none">■ High precision■ Broad scope■ Sentence-by-sentence■ High comprehension■ Background Knowledge.■ Single language■ Slow	<ul style="list-style-type: none">■ Noisy■ Limited scope■ Corpus-wide statistics■ Minimal reasoning■ Bottom up■ General■ 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.

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

(C) 2002-2011 All Rights Reserved Hayato YAMANA@Waseda Univ.

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

② POWERSETの出力結果

Wikipedia Articles

What did Thomas Edison invent?

Factz from Wikipedia: we found that **Thomas Edison** invented the following

Thomas Edison **invented** : **phonograph**, bulb, device, technology, term, machine, microphone, cylinder, means, [more](#)

Results for Thomas Edison invented phonograph

[Backmasking](#) In 1877 **Thomas Edison** **invented** the **phonograph**, a device that allowed sound to be recorded and reproduced on a rotating cylinder with a stylus (or "needle") attached to a diaphragm mounted at the narrow end of a horn.

[Information science](#) Alexander Graham Bell and **Thomas Edison** **invented** the **phonograph** and telephone in 1876 and 1877 respectively, and the American Library Association was founded in Philadelphia.

[1877 in music](#) **Thomas Edison** **invents** the **phonograph**.

[show all results for "Thomas Edison invented phonograph"](#)

Wikipedia Articles: results 1 - 10 of 705

- [Thomas Edison](#) **Edison** did not **invent** the first electric light **bulb**, but instead **invented** the **first** commercially practical incandescent light. ... **Thomas Edison**
- [Edisonian approach](#) **Edison** did not just **invent** a light **bulb**, he invented an economically viable system of lighting including its generators, cables, metering and so on. ... **Thomas Edison**
- [Edison Records](#) **Thomas A. Edison** **invented** the **phonograph**, the first **device** for recording and **playing** back sound, in 1877.

POWERSET

Poweraset

minimize

Website: powerset.com

Location: San Francisco, California, United States

Founded: October, 2006

Funding: \$12.5M



Poweraset is a search engine that focuses on natural language processing. In other words, Poweraset 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

Wikipedia Articles

How many people in the world?

World: Population source: [freebase™](#) (view topic) ?

6,602,224,175 (July 1, 2007)

Wikipedia Articles: results 1 - 10 of 216007 ?

- [ILR scale](#) S-3 is what is usually used to **measure how many people in the world** know a given language.
- [People's Weekly World](#) The **People's Weekly World** (PWW) is the newspaper associated with the Communist Party USA.
- [World](#) The **world** population is over 6.60 billion **people**. ... By extension, a **'world'** may refer to any planet or heavenly body, especially when it is thought of as inhabited.

検索例2

Wikipedia Articles

What is the smallest country in the world?

Wikipedia Articles: results 1 - 10 of 307534

- ▼ [List of countries and outlying territories by total area](#) | 234 | Vatican City | 0.44 | <0.01% | **Smallest country** in the **world**. | ... Lists of **countries** with rankings Geography
- ▼ [Tuvalu](#) In terms of physical land size, **Tuvalu** is the fourth **smallest country** in the **world**, larger only than the Vatican City—0.44 km²; Monaco—1.95 km² and Nauru—21 km². ... The **smallest** island, Niulakita, was uninhabited until it was resettled by people from Niutao in 1949.
- ▼ [Vatican City](#) The **Vatican City** is the **smallest** sovereign **state** in the **world** at 0.44 square kilometres (44 ha/110 acres). ... As a **country** that is 1.05 kilometres (0.6 mi) long and .85 kilometres (0.5 mi) wide, it has a **small** transportation system with no airports or highways.
- ▼ [Microstates and the European Union](#) The **Vatican City** is the **smallest state** in the **world**. ... As the **smallest country** minting them, the coins have become extremely rare and the hardest and most expensive to collect of any of the microstate coins.

検索例3

Wikipedia Articles

please find books written for children

Wikipedia Articles: results 1 - 10 of 144904

- ▼ **Children's literature** Rene Villanueva (born 1954): award-winning Filipino writer, who has **written books** and **plays** for **children**. ... Those who don't believe in magic will never **find** it.": Roald Dahl
- ▼ **Religious debates over Harry Potter** "It's very disappointing that the author would have to make one of the characters gay", said Roberta Combs, president of the Christian Coalition of America, "It's not a good example for our **children**, who really like the **books** and the movies. ... ↑ **Please**, Mrs. Rowling, May We Have Some More?"
- ▼ **Harry Potter** She has revealed she is currently **penning** two **books**, **one** for **children** and **one** not for **children**. ... US National Endowment for the Arts Chairman Dana Gioia said in 2007 "God bless Harry Potter, and **please** send us many more."

検索例4

The screenshot shows a search interface with a search bar containing the text "please find books written by children" and a "search" button. Below the search bar, it displays "Wikipedia Articles: results 1 - 10 of 144904" and an "advanced" button. Three search results are listed, each with a dropdown arrow icon:


- Children's literature** Another type of **children's literature** is work **written by children**, such as *The Young Visitors* **by** Daisy Ashford (aged 9) or the juvenilia of Jane Austen or Lewis Carroll, **written** to amuse brothers and sisters. ... Those who don't believe in magic will never **find** it." : Roald Dahl
- Religious debates over Harry Potter** I have met thousands of **children** and not even one time has a **child** come up to me and said, "Ms Rowling, I'm so glad I've read these **books** because now I want to be a witch." ... ↑ **Please**, Mrs. Rowling, May We Have Some More?
- Picture book** Picture **books** are most often aimed at young **children**, and while some may have very basic language especially designed to help **children** develop **their** reading skills, most are **written** with **vocabulary** a **child** can understand but not necessarily read. ... As rotary presses proliferated around the country, industry pressure was increased to **find** economically viable **publishing** content.

検索例 5

Wikipedia Articles

junichiro koizumi

Junichiro Koizumi source: [freebase](#) (view topic) ?



is a Japan politician who served as Prime Minister of Japan from 2001 to 2006. Widely seen as a maverick leader of the Liberal Democratic Party (LDP), he became known as an economic reformer, focusing on Japan's government debt and the privatization of its postal service. In 2005, Koizumi led the LDP... [Read complete Wikipedia article](#)

Date of Birth: 1942
Place of Birth: [Yokosuka](#)
Religion: [Shinto](#), [Buddhism](#)

Factz from Wikipedia: we found the following about Junichiro Koizumi ?

Factz fro ?

Junichiro Koizumi **visited :** North Korea, **claim** and shrine.

Results for Junichiro Koizumi visited claim

[Junichiro Koizumi](#) Several journals and news reports in Japan, such as one published by Kyodo News Agency on August 15, 2006, questioned the validity of the **claim** that **Koizumi** was **visiting** as a private citizen, as he recorded his name on the shrine's guestbook as prime minister, and visited the shrine yearly as part of his campaign pledge, which was political in nature.

検索例 5

Wikipedia Articles

junichiro koizumi

Junichiro Koizumi



Factz from Wikipedia: we found the following about Junichiro Koizumi

Junichiro Koizumi visited : North Korea, **claim** and shrine.

Results for Junichiro Koizumi visited claim

[Junichiro Koizumi](#) Several journals and news reports in Japan, such as one published by Kyodo News Agency on August 15, 2006, questioned the validity of the **claim** that **Koizumi** was **visiting** as a private citizen, as he recorded his name on the shrine's guestbook as prime minister, and visited the shrine yearly as part of his campaign pledge, which was political in nature.

[LDP ... Read complete Wikipedia article](#)

Factz from Wikipedia: we found the following about Junichiro Koizumi

Junichiro Koizumi visited : North Korea, claim and shrine.

appointed : commentator, Heizo Takenaka and Hiroshi Oki.


defeated : stalwarts, former and Ryutaro Hashimoto.

showing 3 of 32

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

③ TRUE KNOWLEDGE



How can I help? [GO](#)

[Home](#) | [Forum](#) | [Blog](#) | [Wiki](#) | [Recent Activity](#) | [League Table](#) | [Add Knowledge](#)

Examples of what you can do:
[How many legs does a butterfly have?](#)
[List the US states](#)
[\(More examples...\)](#)

Here is the answer that I found:

- [the integer 6](#)

I used the following facts to provide this answer:

- [butterfly is a subclass of lepidopteran](#) ([endorse](#)) ([contradict](#))
- [lepidopteran is a subclass of insect](#) ([endorse](#)) ([contradict](#))
- [6 is the count for the meronym holonym pair leg and insect](#) ([endorse](#)) ([contradict](#))

(I understood your question to mean: **How many legs (an appendage that supports weight) does a butterfly (diurnal insect typically having a slender body with knobbed antennae and broad colorful wings) have?**)

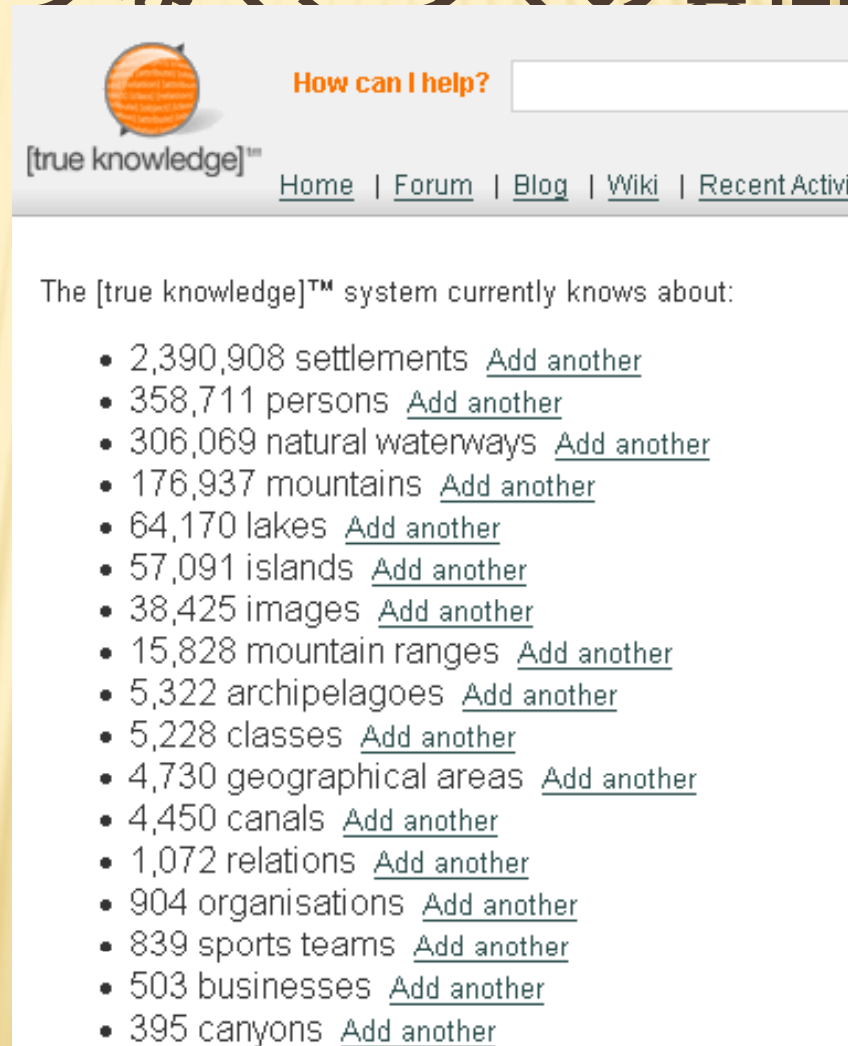
There were other interpretations of your request based on rarely used names. If the interpretation above is not what you were expecting, you can [try again with the rare cases included](#).

[Click here for a more detailed explanation of the answer](#)

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



[true knowledge]™ Home | Forum | Blog | Wiki | Recent Activi

The [true knowledge]™ system currently knows about:

- 2,390,908 settlements [Add another](#)
- 358,711 persons [Add another](#)
- 306,069 natural waterways [Add another](#)
- 176,937 mountains [Add another](#)
- 64,170 lakes [Add another](#)
- 57,091 islands [Add another](#)
- 38,425 images [Add another](#)
- 15,828 mountain ranges [Add another](#)
- 5,322 archipelagoes [Add another](#)
- 5,228 classes [Add another](#)
- 4,730 geographical areas [Add another](#)
- 4,450 canals [Add another](#)
- 1,072 relations [Add another](#)
- 904 organisations [Add another](#)
- 839 sports teams [Add another](#)
- 503 businesses [Add another](#)
- 395 canyons [Add another](#)



true knowledge™

How can I help?

how many people in the world?

GO

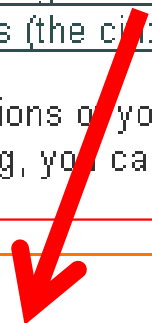
Examples of what
Is a lemur a primate
new subclass of r
(More examples..

[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)

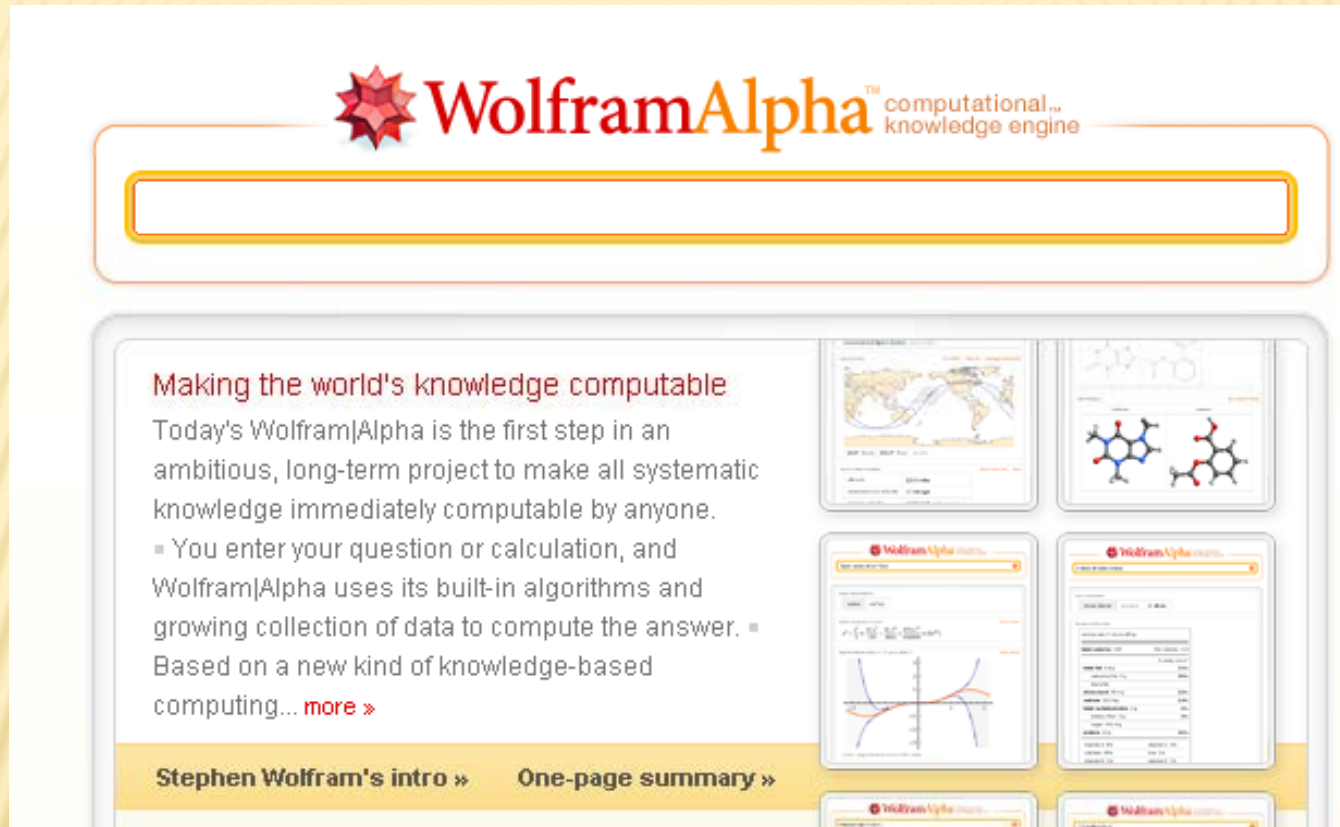
- × best solution for doc shared with 10 people is different than for doc shared with the world
- × sharing patterns of a document might change over time

+ 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>

⑤ WOLFRAMALPHA(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

- × 38000 CPU (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をリーダーとした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/>

⑥ BING

アクセス元を
認識している

The image shows two screenshots of the Bing search engine interface. The top screenshot shows a search for 'apple' with a sidebar on the left containing categories like 'Memory', 'Downloads', 'Reviews', 'Jobs', and 'Monitor'. The bottom screenshot shows a search for 'weather' with a sidebar on the left containing 'RELATED SEARCHES' such as 'Apple iPod', 'Apple Laptop Computer', 'iPhone', 'Apple iTunes', 'Apple MP3 Players', 'Apple Inc', 'Apple Schools', and 'Apple Farming'. The main content area of the bottom screenshot displays weather information for Shinjuku-Ku, Tokyo, including a 5-day forecast and links to '10 Day Forecast', 'Hourly Forecast', and 'Weather Maps'. A blue arrow points from the text box in the top right to the search bar in the bottom screenshot.

APPLE ALL RESULTS 1-20 of 192,000,000 results - [Advanced](#)

Memory Best match

Downloads

Reviews

Jobs

Monitor

RELATED SEARCHES

- Apple iPod
- Apple Laptop Computer
- iPhone
- Apple iTunes
- Apple MP3 Players
- Apple Inc
- Apple Schools
- Apple Farming

weather ALL RESULTS 1-10 of 231,000,000 results - [Advanced](#)

Weather in Shinjuku-Ku, Tokyo [Change location](#)

Today	Thu	Fri	Sat	Sun
73°F · (°C) Wind: 14 mph ENE Humidity: 73%	85° / 74°	86° / 76°	89° / 78°	85° / 68°

80° / 69°

[10 Day Forecast](#) · [Hourly Forecast](#) · [Weather Maps](#) · [Data provided by Foreca](#)

[National and Local Weather Forecast, Hurricane, Radar and Report](#)

The **Weather** Channel and **weather.com** provide a national and local **weather** forecast for cities, as well as **weather** radar, report and hurricane coverage.

[www.weather.com](#) - [Cached page](#)

Weather for Top 100 Cities	On TV
Weather - Search by State	About
National Forecast	Contact
Flight Status	Careers

Show more results from [www.weather.com](#)

⑦NAVER

検索結果数が
表示されない！

統合検索 | ウェブ | 画像 | 動画 | ブログ | クチコミ | テーマ | まとめ 会員登録(無料) ログイン

NAVER

「山名研究室」について良いページをまとめてみませんか？ [+ リンク集を作る](#) 「山名研究室」について話題を投げかけてみませんか？ [+ フリートークをはじめ](#)

ウェブ

 **京大炉・山名研究室**
京都大学 原子炉実験所 原子力基礎工学研究部門 量子リサイクル 工学 研究分野 京都大学大学院 工学研究科 原子核工学専攻 協力講座 **山名 元 研究室** 研究室の概要 メンバー紹介 研究紹介 実験設備 研究業績 おしらせ **研究室紹介(あとみん)** English 〒590-0494 大阪府 泉南郡 熊取町...
hlweb.rri.kyoto-u.ac.jp/hpc-lab/ - キャッシュ | +

www.isc.senshu-u.ac.jp/~thc0640/mda.html
...クレジットデータ社会人部門中間発表だ2回」発表チーム: 矢島**研究室**.175Rs,GYOEN,ウイングバード,金鉦掘蔵,吹けよ嵐、呼べよ嵐 第3回研究部会... 発表チーム: AIZAWAINOGA,Kurosu Road,立教大学岡太**研究室**,やればできる子たち,櫻井**研究室** 伴在健太郎,AKI,矢上256,ESPRESSO ...
www.isc.senshu-u.ac.jp/~thc0640/mda.html - キャッシュ | +

IBM 早稲田大学 理工学術院 山名研究室 - Japan
早稲田大学**山名**早人**研究室**が研究を進めている、インターネット上のWebページのデータ解析に、IBMのインターネット経由でハイパフォーマンス・コンピューティング(HPC)環境を提供するソリューション「IBM Deep Computing Capacity On Demand(DCCoD)」を活用したことを発表しました。
www-06.ibm.com/jp/solutions/casestudies/20080606waseda.html - キャッシュ | +

画像



[もっと](#)

トレンドランキング

- 1 地震速報 誤報
- 2 亀田興毅
- 3 田母神俊雄
- 4 衆議院選挙
- 5 くめ納豆

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

END
