人工知能学会
インタラクティブ情報アクセスと可視化マイニング研究会（第10回）

トレンドと検索頻度とWikipediaページビューの三角関係

吉田光男,
角田孝昭, 山本幹雄, 荒瀬由紀, 津川翔
yoshida [at] cs.tut.ac.jp / @ceekz

WEB SCIENCE 2015
Association for Computing Machinery
#WebSci15
REGISTRATION

Ceek.jp News
Robot type news sear
検索クエリの分類
ウェブ
ニュース
最新ニュースを手軽に。ロボット型ニュース検索サイトです。5分で反映。
AND
検索
学術文献のソーシャルメディアでのインパクトを計測する Ceek.jp Altmetrics をリリースしました
いいね！ 367
Hot Keywords
発表 43.54 pt - 26
ロシア 31.92 pt - 13
東京 27.93 pt - 21
バラクーダネットワークス 24.25 pt - 1
制裁 23.96 pt - 9
協議 22.28 pt - 3
北海道神宮 21.32 pt - 1
家族 21.01 pt - 2
北朝鮮 20.21 pt - 2
高原 18.54 pt - 1
イチロー 18.43 pt - 1
甲子園 18.01 pt - 6
開幕 17.84 pt - 15
田中 16.86 pt - 1
出場 15.74 pt - 15
優勝 15.14 pt ↑ 8
政府 14.83 pt ↓ 9
現在 14.14 pt ↓ 12
次元 13.49 pt ↓ 1
優秀 13.49 pt ↑ 1
Search Keywords
北朝鮮 1.544 % ↑ 2
最新情報
(121749件) 新着順に表示 RSS
薬物脱却 女性支える 武蔵野ダルク 日野に専用入寮施設
入寮者が生活を送る部屋。落ち着いた環境で薬物依存からの脱却を目指す＝日野市内で 多摩地区で薬物依存患者のリハビリや社会復帰支援に取り組む民間団体「武蔵野ダルク」が、日野市内に多摩地区初の女性専用入寮施設を開いた。女性向け施設は全国的に不足しており、受け皿の準備を進めていた。(小松... 本文 >> (東京新聞 07:15)
薬物脱却 女性支える 武蔵野ダルク 日野に専... (地域・地方) 07:15 東京新聞
同性 でトーク (地域・地方) 07:15 東京新聞
現 会場常設 (地域・地方) 07:15 東京新聞
兵 対策強化 (地域・地方) 07:10 神戸新聞
日 中国... (中国・朝鮮) 07:10 レコードチャイナ
藍は 07:05 デイリースポーツ
鶴竜、綱とりへ第一関門・日馬を突破！ (スポーツ) 07:05 デイリースポーツ
全国高校柔道で90キロ級県勢初V 工大福井の白川剛章 (地域・地方) 07:05 福井新聞
レッズのチャブマンが打球受け顔面骨折 (スポーツ) 07:05 デイリースポーツ
ヘビー級へ…石田順裕の大増量作戦！ (スポーツ) 07:05 デイリースポーツ
本文抽出
社会
(10237件) 新着順に表示 RSS
「血糖値下げる漢方薬」で重体 注意を 06:41 NHK
EU、対ロ制裁強化決定へ 資産凍結など拡大 06:40 スポニチ
重機がJR高架に引っかかる 現場は一時通行止め 東京・千代田区 06:37 フジテレビ
ていたと発表 06:36 日本テレビ
よそ20万ビットコインが残存 06:35 フジテレビ
の一部区間で通行止め続く 06:32 フジテレビ
管接続ミス トイレ用の水を飲用に 06:30 毎日新聞
これまでの研究概要

最近の興味

- ソーシャルメディアの活用・分析
  - 投稿者の属性（活動地域）推定
    - 投稿テキストから
    - 人間関係から
  - 情報の伝搬モデル
- トレンドの要因分析
  - ウェブでの流行は何日で終わる？
  - トレンドを誘引する情報源は？

本日の話のモチベーション

本日の話の概要

- Wikipedia ページビューと検索頻度
  - Wikipedia ページビューで検索頻度を推定
  - 予備調査結果

- 自動車購買台数と検索行動量
  - 検索行動量の有用性を検証
  - 検索行動量
    - Google Trends
    - Wikipedia ページビュー

WIKIPEDIA ページビューと検索頻度との関係

話題の盛り上がりを知る（1/2）

- 話題「インフルエンザ」のピークは？
  - 2015年1月18日

- 話題「アン・ハサウェイ」のピークは？
  - 2014年12月12日

検索頻度はこれらの答えを知っている

Google トレンド

www.google.co.jp/trends/

Google

Search Google Trends

ログイン

トレンド　日本　2015年2月13日

2014 年、Google では膨大な数の検索が行われました。
検索キーワードで一年を振り返りましょう。

「Year in Search: 検索で振り返る 2014」を見る

最新トレンド

キーワード　日本

倉科カナ
検索件数: 50,000+
画像ソース - モデルプレス

テラフォーマーズ
検索件数: 20,000+
画像ソース - ファミ通.com

安田由紀奈
検索件数: 10,000+
画像ソース - livedoor

2014 のチャート

急上昇ワード　日本

ドラマ

1 昼顔
2 明日ママがいない
3 花子とアン
4 ごちそうさん
5 失恋ショコラティエ

http://www.google.co.jp/trends/

YouTube　日本

Introducing Spot
▲ 地域での今日の再生回数 50,000 回以上

総合ランキング　日本

総合

Google トレンド - ウェ

www.google.co.jp/trends/explore#q=%E3%82%A4%E3%83%B3%E3%83%95%E3%83%AB%E3%82%A8%E3%

Google

Search Google Trends

ログイン

調べる　すべての国　1月 2012 - 1月 2015　すべてのカテゴリ　ウェブ検索

比較　キーワード

インフルエンザ
検索キーワード

+キーワードを追加

人気度の動向

ニュースのヘッドライン　予測

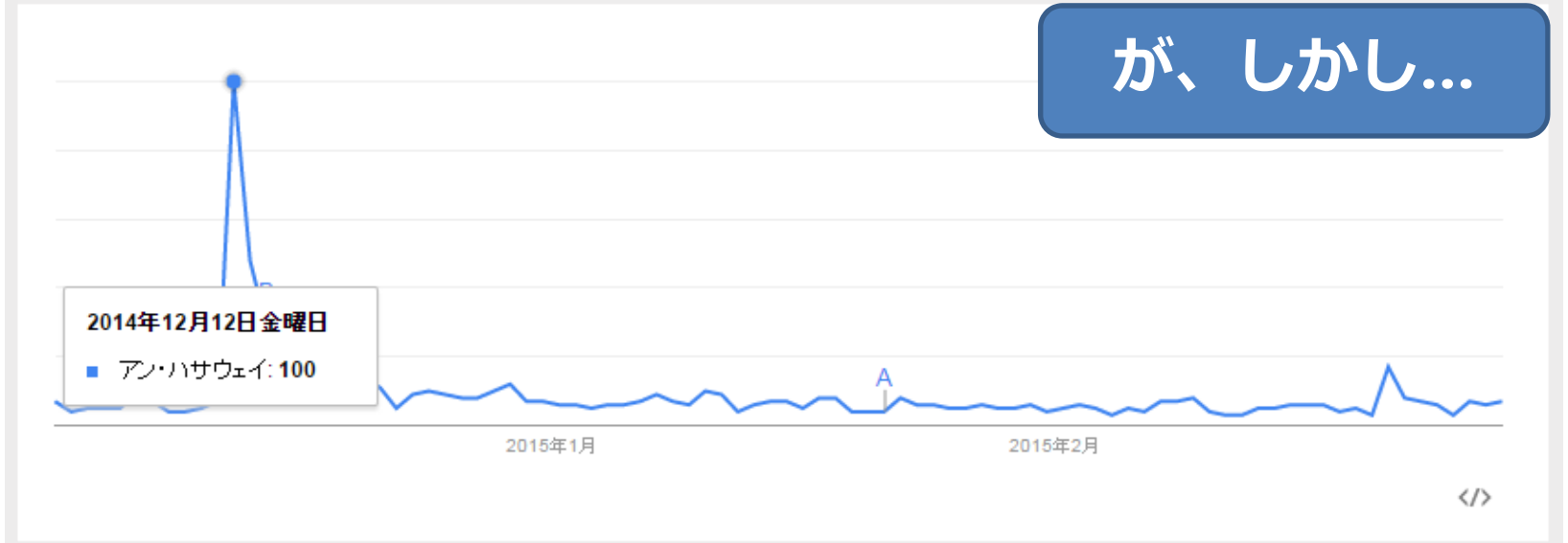
Google Trends : インフルエンザ

地域別人気度

国/地域 | 都市

Google Trends で答えがわかった！

が、しかし…



話題の盛り上がりを知る（2/2）

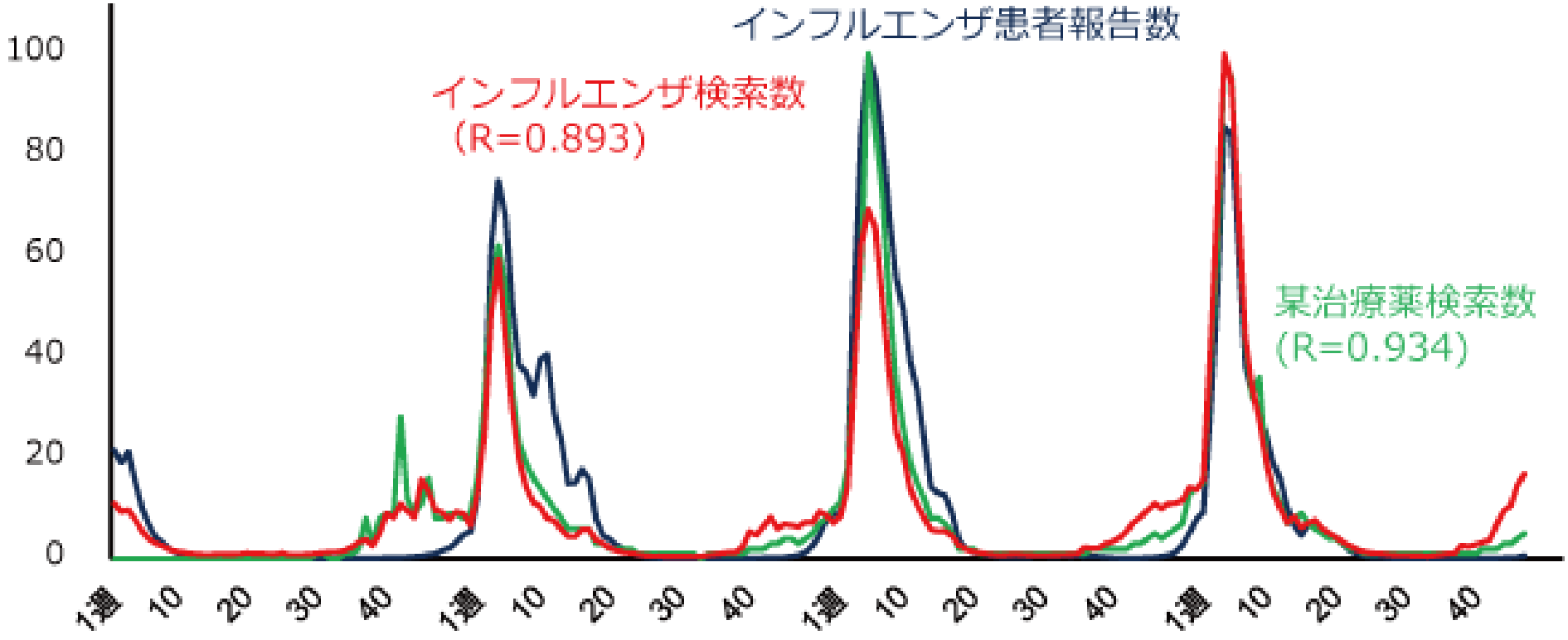
- 2014年12月12日に盛り上がった話題は？
  - アン・ハサウェイ
  - ダークナイト ライジング
  - 長澤奈央
  - まるか食品

有用なデータ

この調査には時系列データが必要
Google Trends に大規模リクエスト？

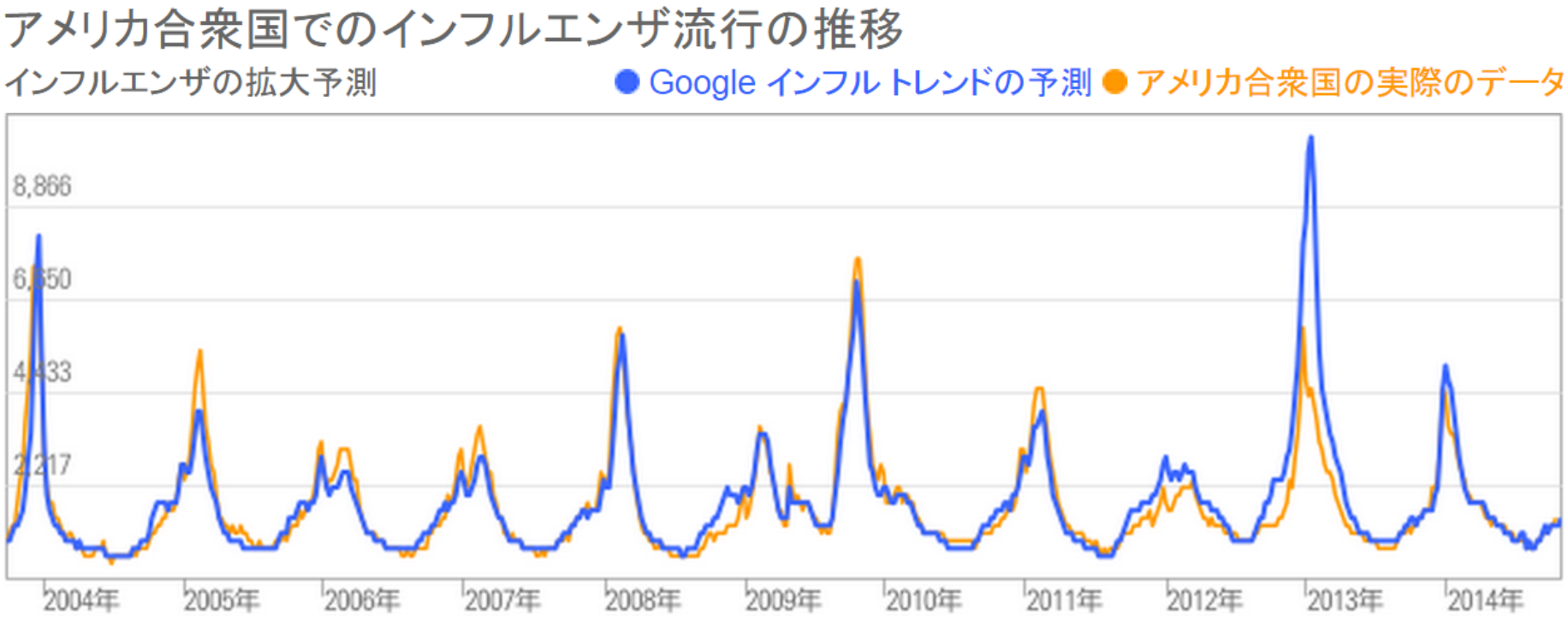
ビッグデータレポート -
docs.yahoo.co.jp/info/bigdata/
ビッグデータレポート
YAHOO! JAPAN
トップ
サービス一覧
会社概要一覧
プレスルーム
IR情報
(Investor Relations)
採用情報
企業の社会的責任
Corporate Home(English)
Yahoo! JAPANビッグデータレポート
「Yahoo! JAPANビッグデータレポート」は、日本最大級のポータルサイトYahoo! JAPANに匿名化され、蓄積された検索・広告・ショッピング・地域情報・ソーシャル上のトレンド情報など あらゆるカテゴリーの膨大なデータを分析・活用し、世の中の課題解決を行っていくことを目的に展開していくもので、定期的に情報を更新してまいります。
ビッグデータ分析でみるインフルエンザ感染状況：2014－2015
2015年2月9日 更新
第47回衆院選の議席数予測を振り返る
2015年1月9日
ビッグデータが導き出した第47回衆院選の議席数予測
2014年12月12日 更新
大相撲の人気がジャニーズ
http://docs.yahoo.co.jp/info/bigdata/
検索ログは有用な情報源である

インフルエンザの流行予測（1/2）



http://docs.yahoo.co.jp/info/bigdata/influenza/2013/

インフルエンザの流行予測（2/2）



https://www.google.org/flutrends/intl/ja/about/how.html

映画・ゲームなどの売上数予測

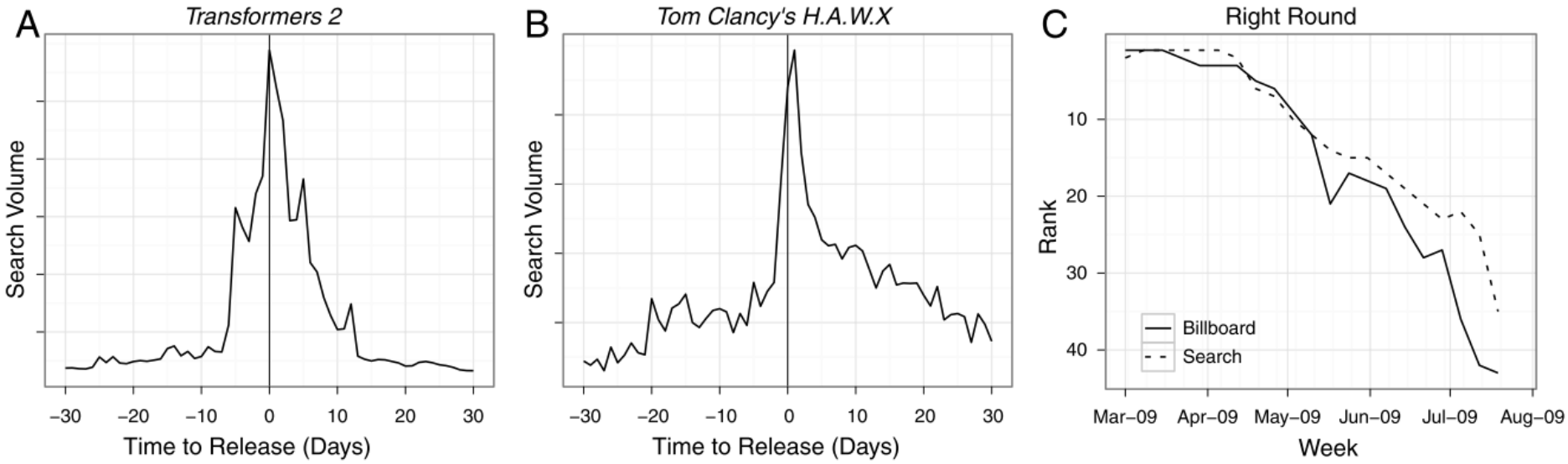


Fig. 1. Search volume for the movie *Transformers 2* (*A*) and the video game *Tom Clancy's H.A.W.X.* (*B*) prior to and after their release, and search and Billboard rank for the song "Right Round" by Flo Rida (*C*).

[Goel 10] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J.: Predicting consumer behavior with Web search, Proceedings of the National Academy of Sciences, Vol. 107, No. 41, pp. 17486–17490 (2010)

検索頻度データ入手の困難性

- 検索ログは一般的に公開されない
  - 研究用のデータセットはあるが…
- Google Trends
  - キーワードに対するトレンドが取得可能
  - 任意の日のトレンドワードの探索が困難
- **公開データをもとに推定したい**

アン・ハサウェイ - Goog
https://www.google.co.jp/search?q=アン・ハサウェイ&oq=アン・ハサウェイ
Google
アン・ハサウェイ
ログイン
ウェブ
画像
動画
ニュース
ショッピング
もっと見る
検索ツール
約 743,000 件 (0.13 秒)
Wikipedia
アン・ハサウェイ - Wikipedia
ja.wikipedia.org/wiki/アン・ハサウェイ
父親のジェラルド・ハサウェイは弁護士、母親のケイト・マッコーレーは舞台女優。名前の由来は劇作家ウィリアム・シェイクスピアの妻からである。マイケルという兄とトーマスという弟がいる。先祖はアイルランド人とフランス人、さらにたどるとドイツ人とアメリカ・ ...
ワン・デイ 23年のラブストーリー - プラダを着た悪魔 - ブリティ・プリンセス
その他の画像
検索上位ページの訪問率が高いという調査結果あり
美少女 ⇒ 大女優 アン・ハサウェ
matome.naver.jp/odai/21336127119894
2014/12/30 - 【随時更新中】1999年～現在の
ます。デビュー当時は17歳。衝撃的な美少女ぶりで話題になりました。
女優
アン・ジャクリーン・ハサウェイは、アメリカ合衆国の女優。身長173cm。
ウィキペディア
生年月日: 1982年11月12日 (32歳)
アン・ハサウェイ - NAVER まとめ
matome.naver.jp/topic/1LwS2
Page View で検索頻度を推定できそう
予備調査としての類似性調査
画『ブルックリンの恋人たち』は、同作のスタッフとアン・ハサウェイが ...
Google: アン・ハサウェイ
アン・ハサウェイ - 映画.com
eiga.com › 俳優・監督
アン・ハサウェイの新作映画、写真、画像、動画、関連ニュースの情報。ニューヨーク・ブルックリン出身。弁護士の父と舞台女優の母の間に生まれる。99年、TVシリーズ「ゲット・リアル」の主人公役に抜擢され、女優としてのキャリアをスタート。01年「プリティ・ ...
インターステラー
2014年
ダークナイトライジング
2012年
レ・ミゼラブル
2012年
プラダを着た悪魔
2006年
ブリティ・プリンセス
2001年
他の人はこちらも検索
他 15 件以上を表示

Page view statistics for

dumps.wikimedia.org/other/pagecounts-raw/

Page view statistics for Wikimedia projects

(For up-to-date information (outages, ...) about this dataset, please consult the dataset's wiki page.)

Pagecount files per year

- 2007
- 2008
- 2009
- 2010
- 2011
- 2012
- 2013
- 2014
- 2015

What are the page view statistics files and what do they contain?

Each request of a page, whether for editing or reading, whether a "special page" such as a log of actions generated on the fly, or an article from Wikipedia or one of the other projects, reaches one of our squid caching hosts and the request is sent via udp to a filter which tosses requests from our internal hosts, as well as requests for wikis that aren't among our general projects. This filter writes out the project name, the size of the page requested, and the title of the page requested.

Here are a few sample lines from one file:

fr.b Special:Recherche/Achille_Baraguey_d%5C%27Hilliers 1 624

Wikipedia 記事ページの PV は入手可能

（言語、ファイルパス、1時間単位のアクセス数）

In

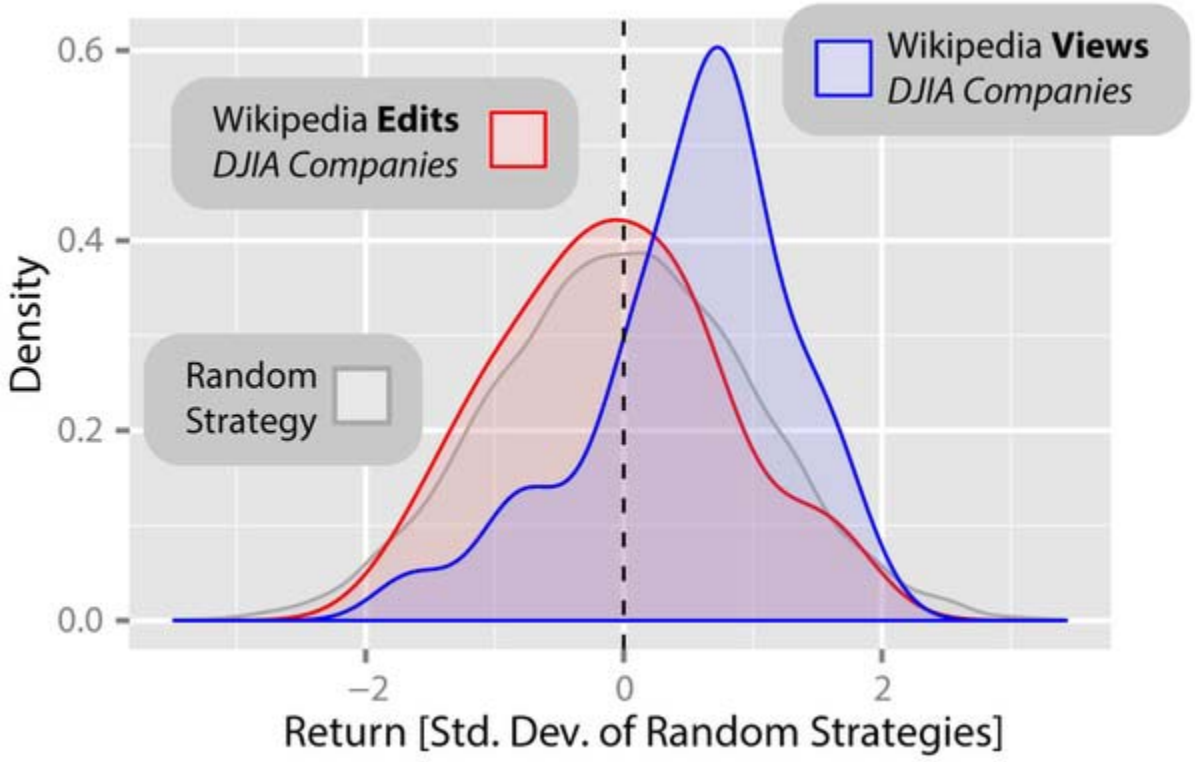
- wikimedia: ".m"
- wikipedia mobile: ".mw"
- wikinews: ".n"
- wikiquote: ".q"
- wikisource: ".s"
- wikiversity: ".v"
- mediawiki: ".w"

http://dumps.wikimedia.org/other/pagecounts-raw/

Projects without a period and a following character are wikipedia projects.

The second column is the title of the page retrieved, the third column is the number of requests, and the fourth column is the size of the content returned.

Wikipedia PV で株価予測



※ Wikipedia Page Views を活用する研究はいくつかある

[Moat 13] Moat, H. S., et.al.: Quantifying Wikipedia Usage Patterns Before Stock Market Moves, Scientific Reports 3, Vol. 3, No. 1801 (2013)

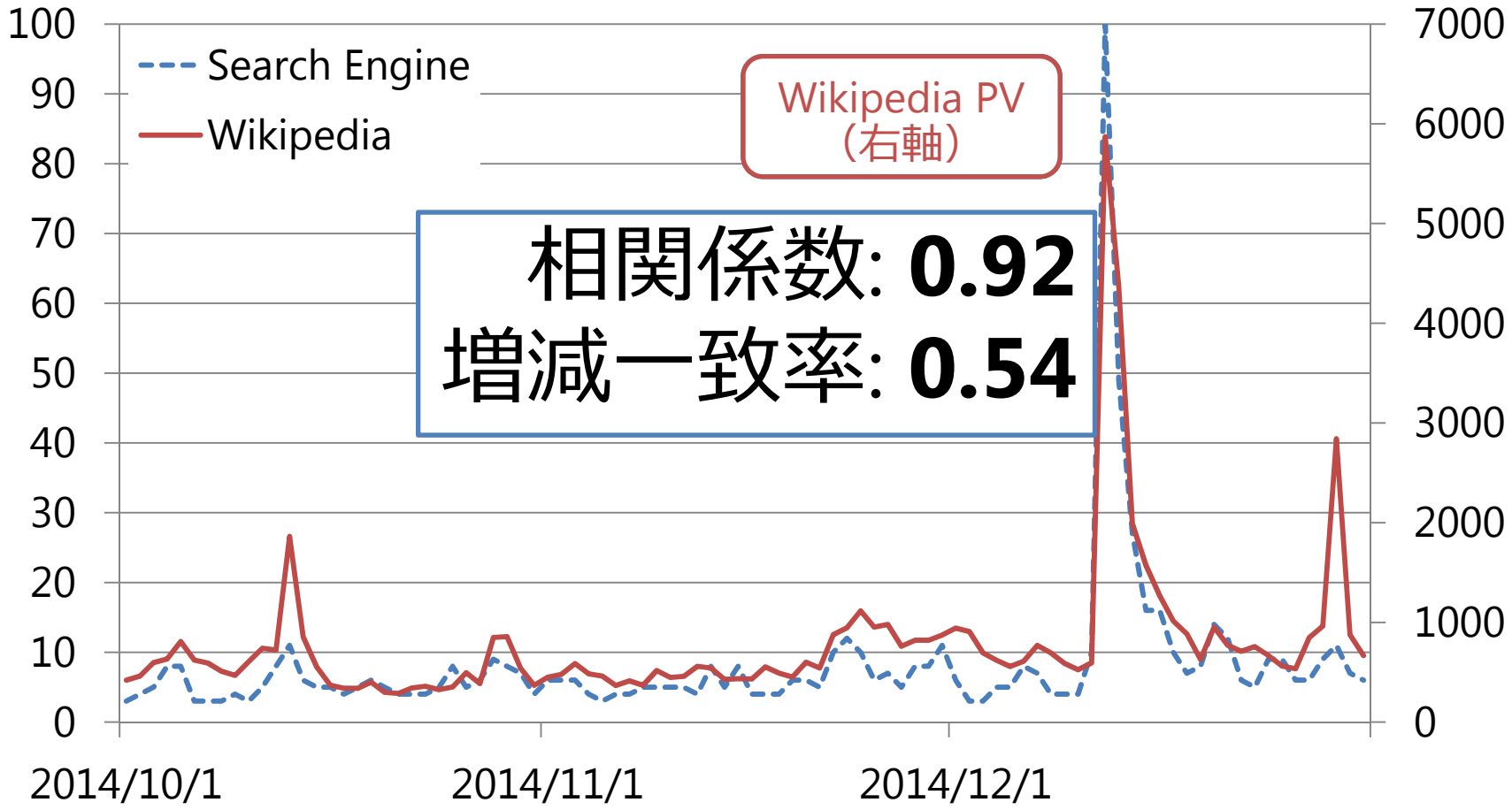
類似性評価の指標・区間

- ピアソンの積率相関係数
  - トレンドの各値を集合と見なした類似性
- 増減一致率（←おすすめ）
  - トレンドの増減がどの程度一致するか
    （ランダムの場合は 0.33 となる）

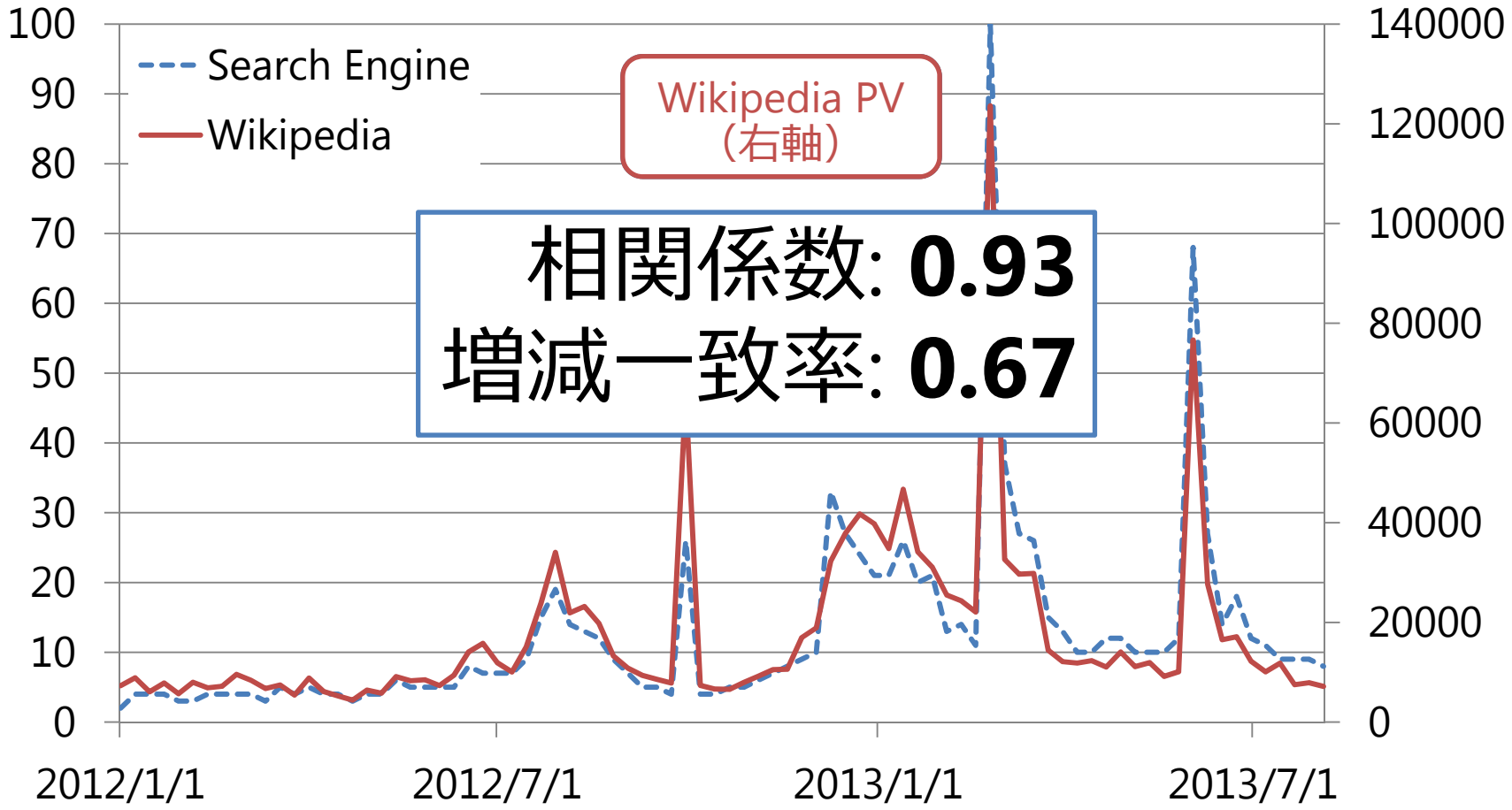
- 評価区間
  - 日別 : 2014年10月～12月
  - 週別 : 2012年～2014年
  - 月別 : 2008年～2014年

Google Trends の制約
（データ取得期間が限定される）

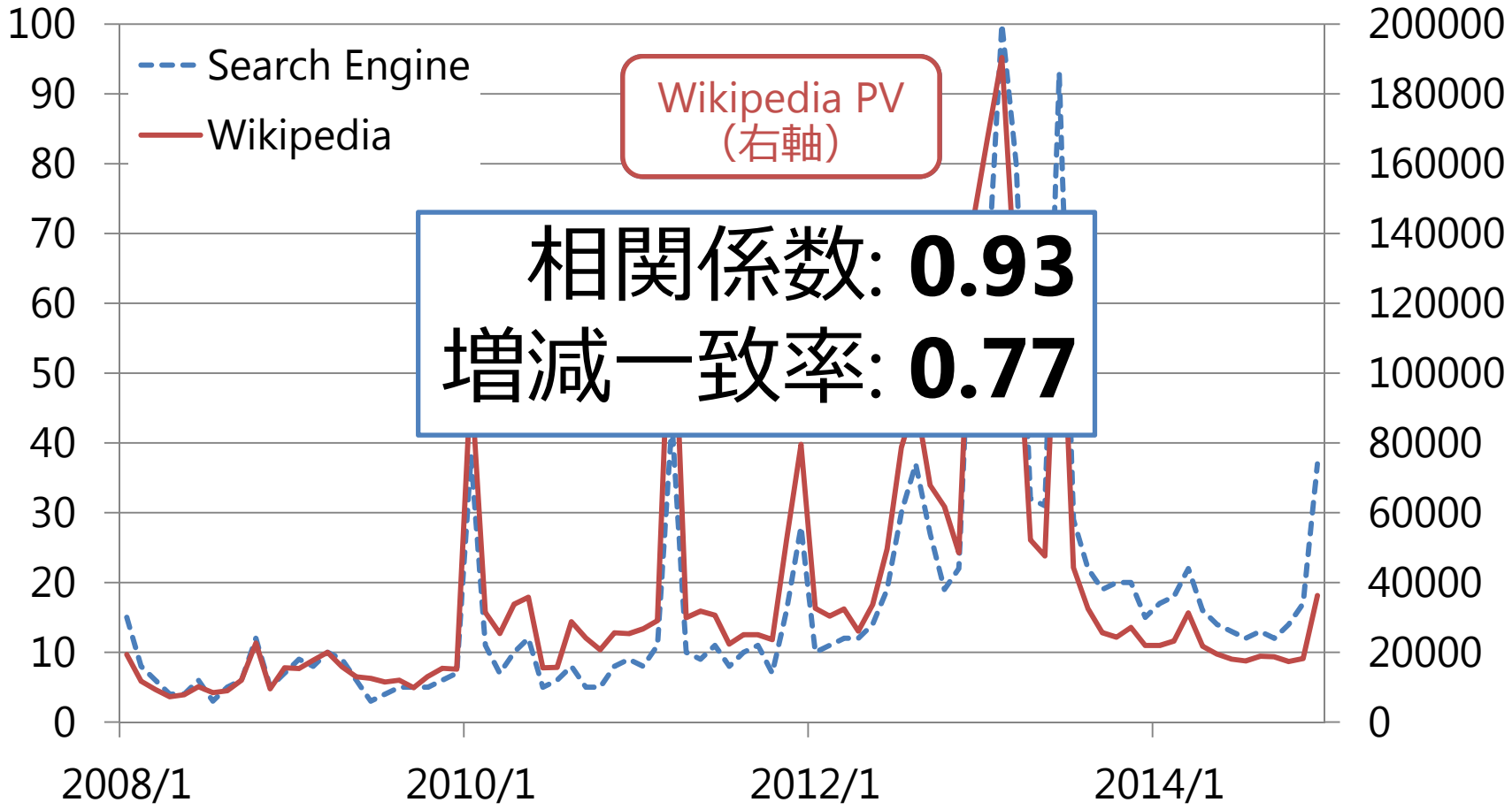
日別トレンドの例（アン・ハサウェイ）



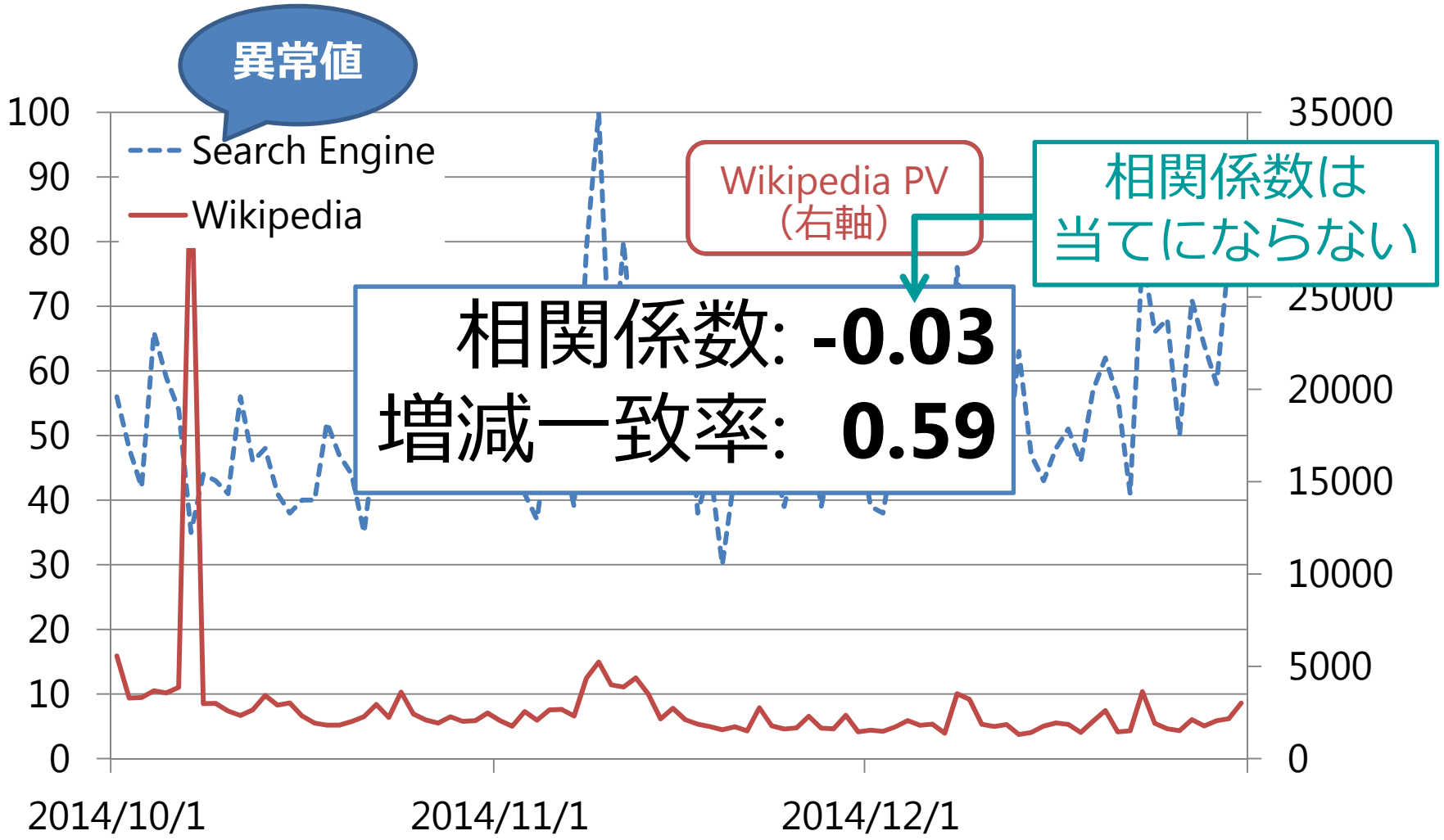
週別トレンドの例（アン・ハサウェイ）



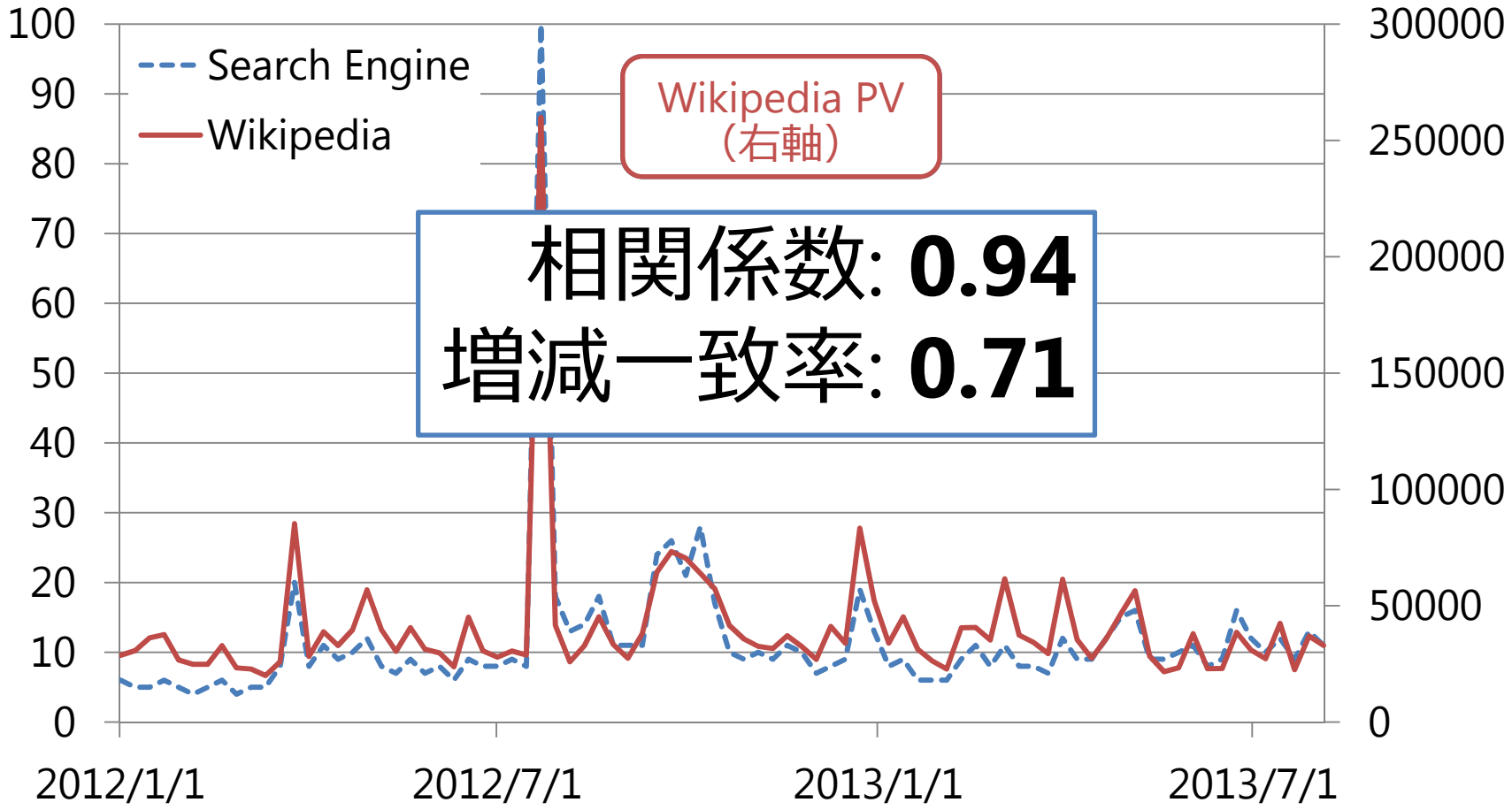
月別トレンドの例（アン・ハサウェイ）



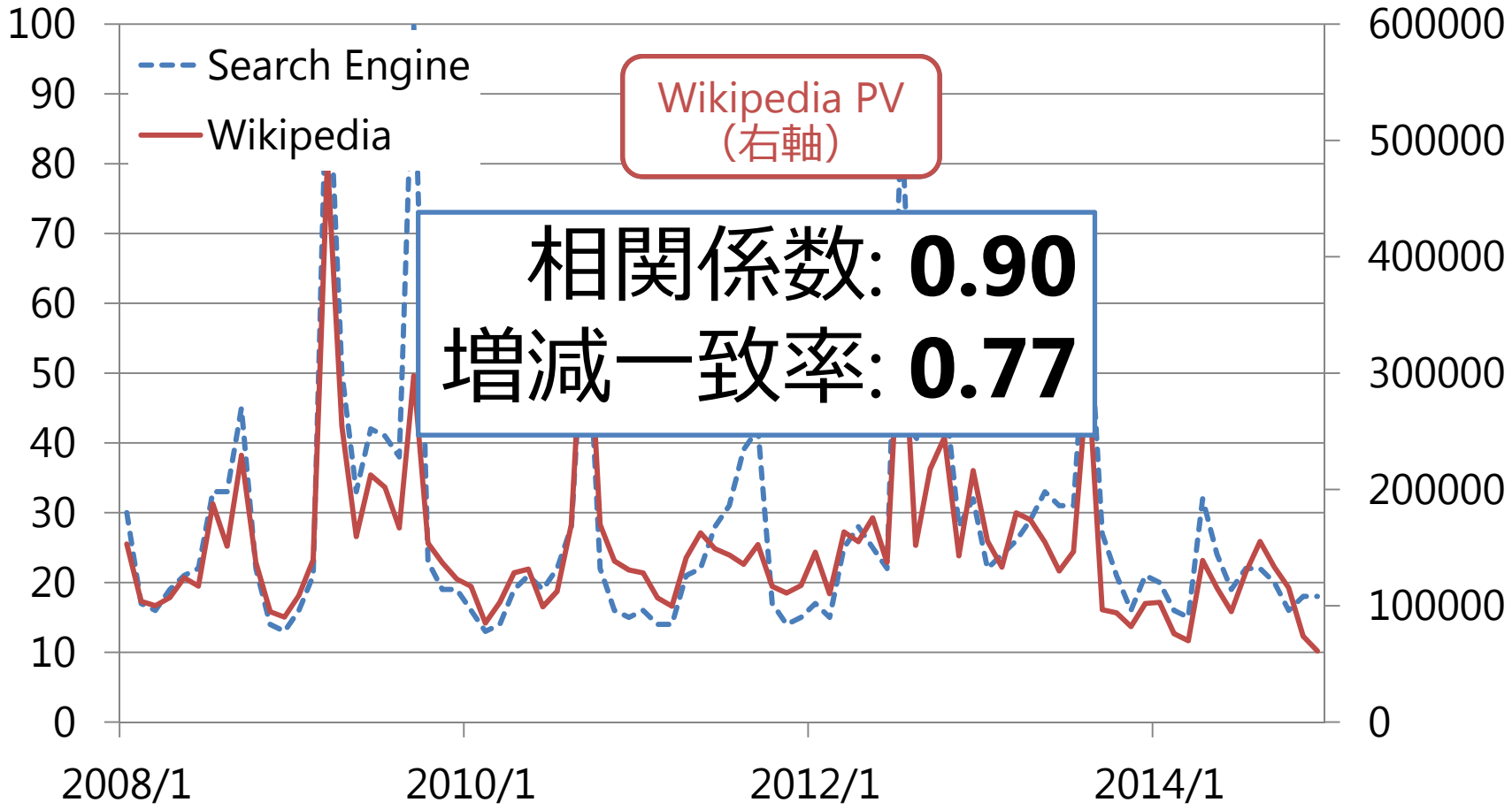
日別トレンドの例（イチロー）



週別トレンドの例（イチロー）



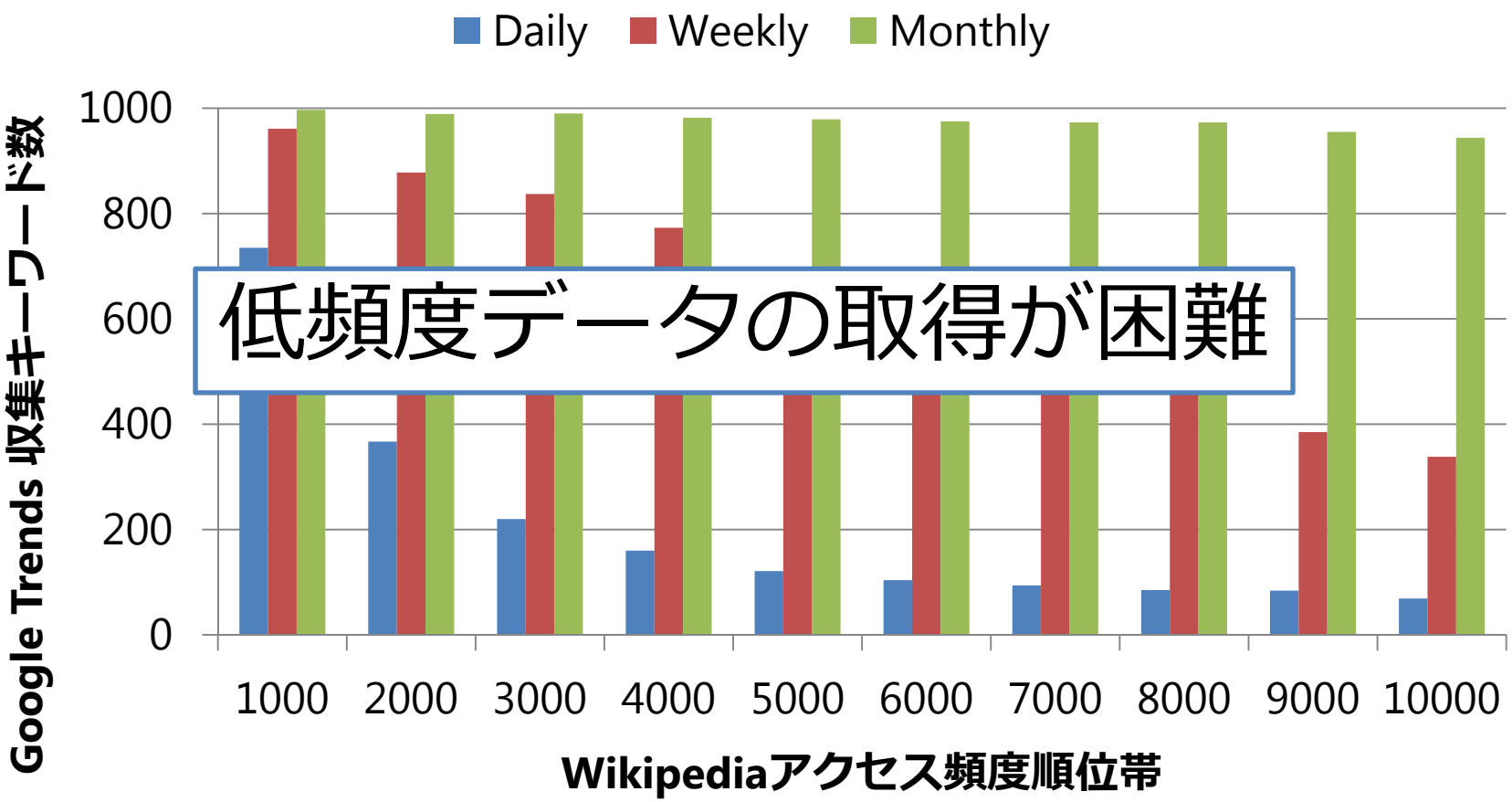
月別トレンドの例（イチロー）



評価対象キーワード

- Wikipedia のアクセス上位の見出し語
  - 1,000 PV/day（2008年～2014年）
- 対象キーワード数
  - アニメ: 163
  - 漫画　: 182
  - 映画　: 48
  - 人名　: 837
    - アクセス上位 10,000 キーワードも追加

Google Trends における制限状況



日本語版 Wikipedia より人名10,000件

カテゴリ別の収集数

		実数			割合		
	対象数	日別	週別	月別	日別	週別	月別
アニメ	163	122	153	159	0.75	0.95	0.98
漫画	182	134	170	178	0.74	0.93	0.98
映画	48	35	45	46	0.73	0.94	0.96
人名	837	645	810	835	0.74	0.98	1.00

日本語版 Wikipedia より 1,000pv/day の見出し語を抽出

カテゴリ別の類似度

		相関係数			増減一致率		
	対象数	日別	週別	月別	日別	週別	月別
アニメ	163	0.43	0.71	0.78	0.53	0.57	0.65
漫画	182	0.37	0.70	0.75	0.53	0.56	0.63
映画	48	0.49	0.74	0.76	0.51	0.57	0.63
人名	837	0.57	0.73	0.72	0.52	0.66	0.70

上昇

日本語版 Wikipedia より 1,000pv/day の見出し語を抽出

カテゴリ別の類似度

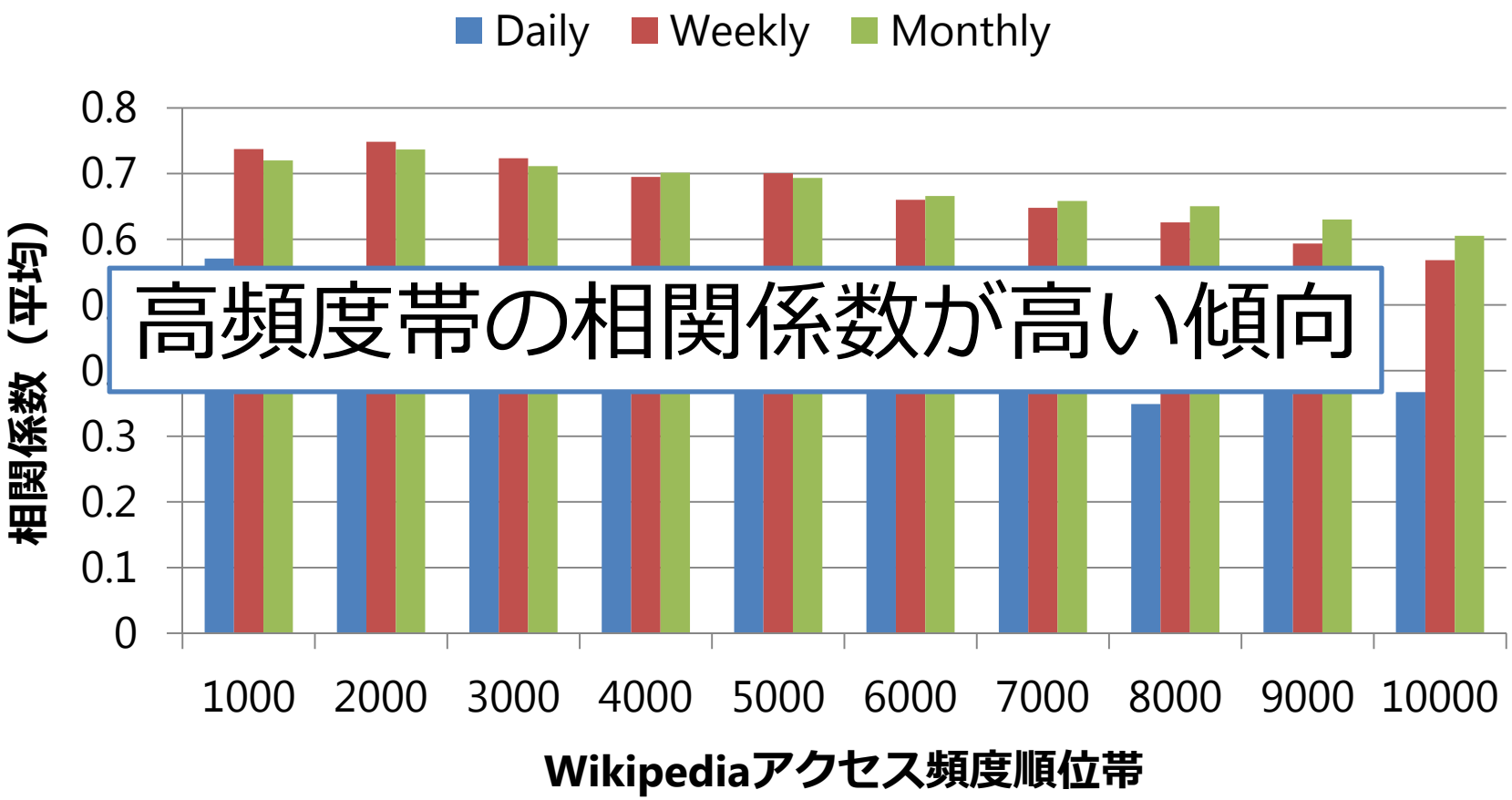
		相関係数			増減一致率		
	対象数	日別	週別	月別	日別	週別	月別
アニメ	163	0.43	0.71	0.78	0.53	0.57	0.65
漫画	182	0.37	0.70	0.75	0.53	0.56	0.63
映画	48	0.49	0.74	0.76	0.51	0.57	0.63
人名	837	0.57	0.73	0.72	0.52	0.66	0.70

詳細にみる

上昇

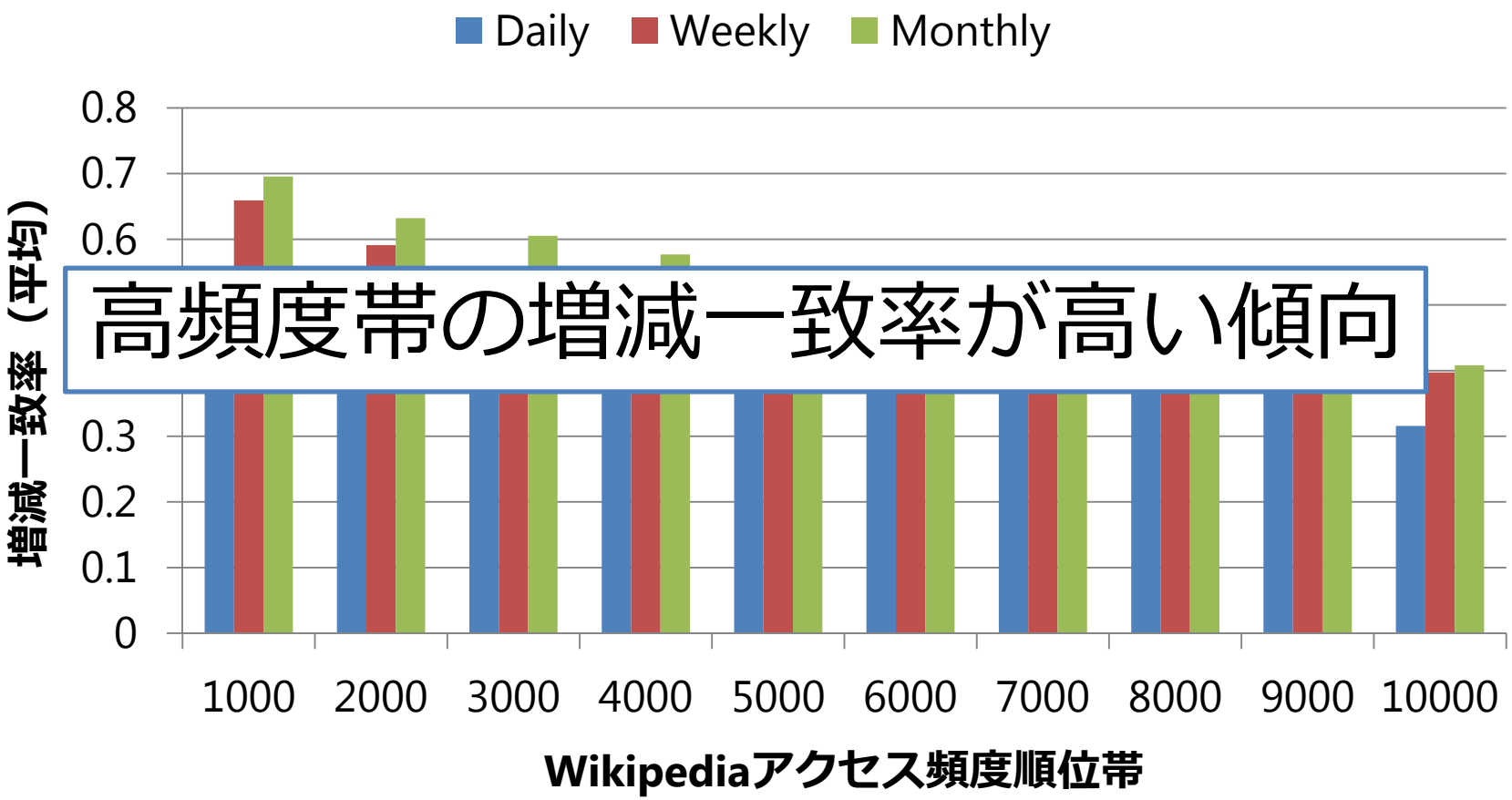
日本語版 Wikipedia より 1,000pv/day の見出し語を抽出

頻度帯別の相関係数



日本語版 Wikipedia より人名10,000件

頻度帯別の増減一致率



日本語版 Wikipedia より人名10,000件

要因の考察

- 日別の類似度が低くなる要因
  - ボットなどの一時的アクセス増大の影響（PV はアクセスユーザ数ではない）
  - Google Trends のタイムゾーンが不明
  - Wikipedia のデータに遅延あり（？）
    - WebSci ’15 で仕入れた情報
- アニメ等の類似度が低くなる要因
  - 検索キーワードと見出し語との異なり

学園黙示録 - Google 検索

https://www.google.co.jp/search?q=学園黙示録

Google 学園黙示録 ログイン

ウェブ 画像 動画 ショッピング ニュース もっと見る 検索ツール

約 585,000 件 (0.14 秒)

他のキーワード: 学園黙示録2期 学園黙示録 動画 学園黙示録 ss
学園黙示録 4話 学園黙示録 夢小説

学園黙示録 HIGHSCHOOL OF THE DEAD 公式サイト

www.nbcuni.co.jp/rondorobe/anime/hotd/

学園黙示録 HIGHSCHOOL OF THE DEAD TVアニメ化決定！！

学園黙示録 HIGHSCHOOL OF THE DEAD - Wikipedia

ja.wikipedia.org/wiki/学園黙示録_HIGHSCHOOL_OF_THE_DEAD

学園黙示録 HIGHSCHOOL OF THE DEAD』（がくえんもくしろく ハイスクール・オブ・ザ・デッド）は、原作：佐藤大輔、作画：佐藤ショウジによる日本の漫画作品。略称は、「H.O.T.D.」「HOTD」「スクデッド」など。2010年に単行本第4巻までの内容がテレビアニメ化 ...

概要 - 連載と休載 - 物語 - 登場人物

学園黙示録 の画像検索結果

画像を報告

Google: 学園黙示録

比較 キーワード

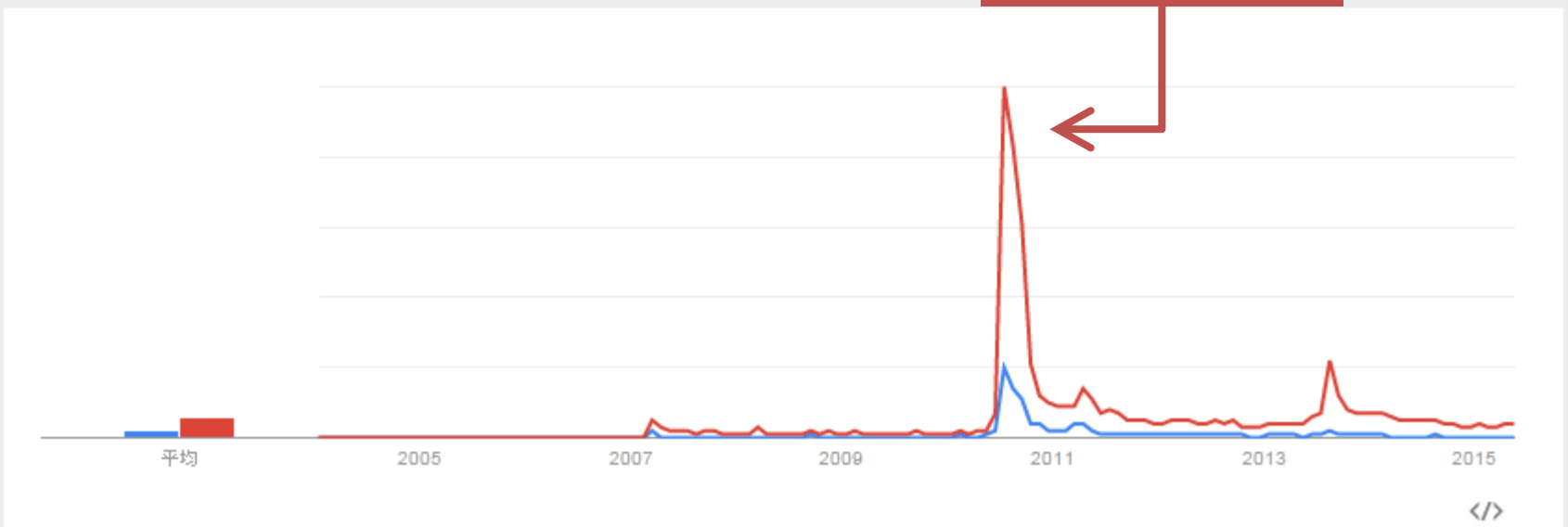
学園黙示録HIGHSCHOOL ...
検索キーワード

学園黙示録
検索キーワード

+キーワードを追加

人気度の動向

検索頻度が4倍

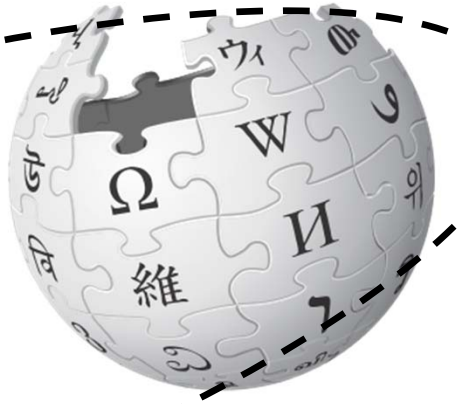


「学園黙示録」

検索頻度が4倍

今回の評価対象

「学園黙示録 HIGHSCHOOL OF THE DEAD」

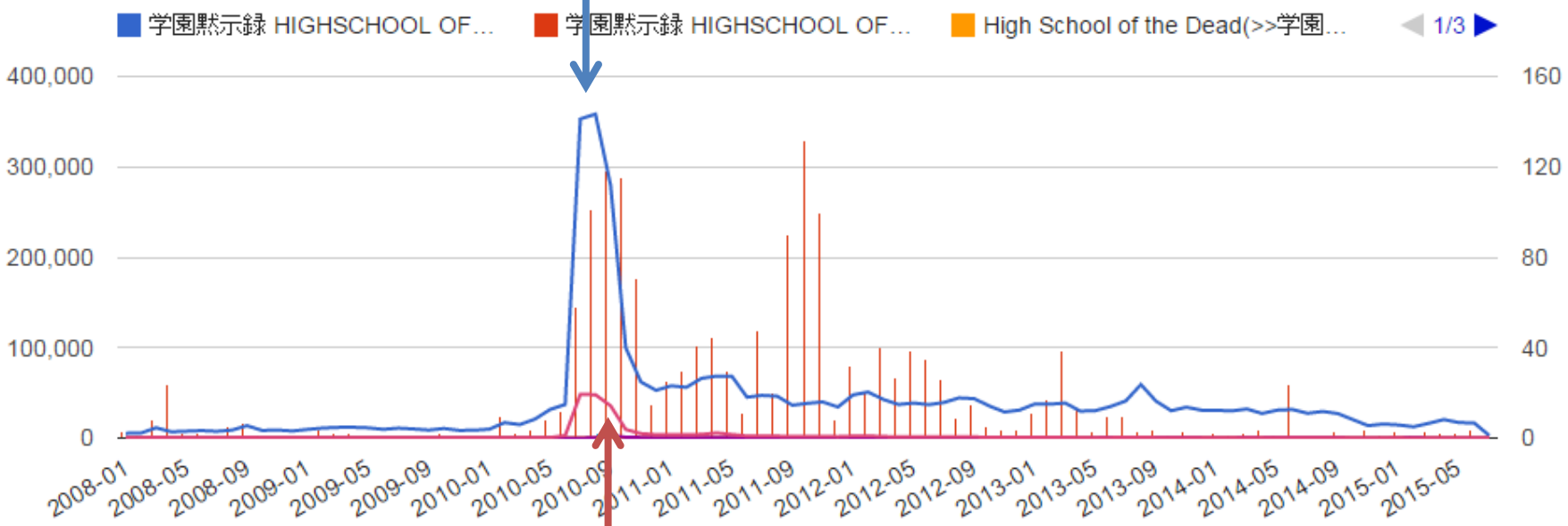


WIKIPEDIA
The Free Encyclopedia

学園黙示録
HIGHSCHOOL OF THE DEAD

Wikipedia

学園黙示録
HIGHSCHOOL OF THE DEAD



学園黙示録

まとめ・今後の方針（1/2）

- 検索頻度推定のための予備調査
  - Wikipedia Page View との類似性を調査
  - 高い類似性が明らかに
    - 特に高頻度帯、週別・月別の粒度

- 検索頻度推定に向けて
  - Wikipedia 見出し語に対する部分一致文字列
    - 調査対象を任意の文字列にする
      → 高類似度となるモデルを探る

まとめ・今後の方針（2/2）

- 語の重要度計算
  - 出現頻度（TF）
  - 文書頻度（DF）

  } **TF-IDF**

  ↪ 検索エンジンでのヒット件数で代用

- Wikipedia ページビューを活用
  - ページビューを語の重要度に使えないか

詳細資料

- 検索頻度推定のためのWikipediaページビューデータの分析
  - 吉田光男, 荒瀬由紀, 角田孝昭, 山本幹雄
  - 第29回人工知能学会全国大会
- Wikipedia Page View Reflects Web Search Trend
  - Mitsuo Yoshida, Yuki Arase, Takaaki Tsunoda, Mikio Yamamoto
  - The 2015 ACM Web Science conference

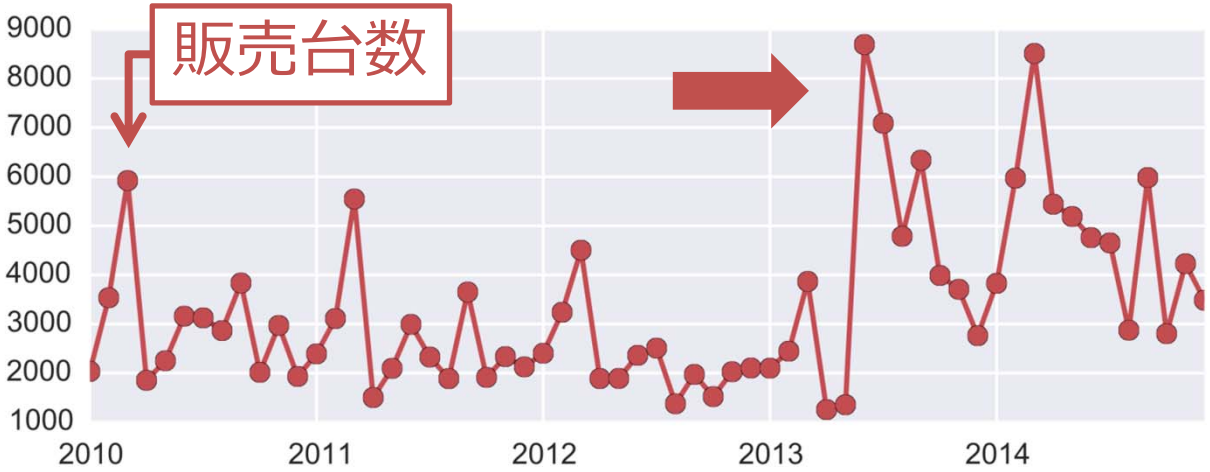
商品購買予測における三角関係

本研究の目的

- 商品購買数の予測を行いたい
  - 自社製品の生産計画
  - 他社製品の一早い動向把握
- よりリアルタイムに予測したい
  - 公式指標の場合、リアルタイム性に欠ける
    （発表が月初のみ、遅れて発表…など）
  - 消費者に近い情報を用いて素早く予測したい

従来研究: 過去のデータから予測

- 従来研究
  - 周期的要因やトレンド要因に基づいて自己回帰的に予測
  - ピークの突出やトレンドの転換に弱い

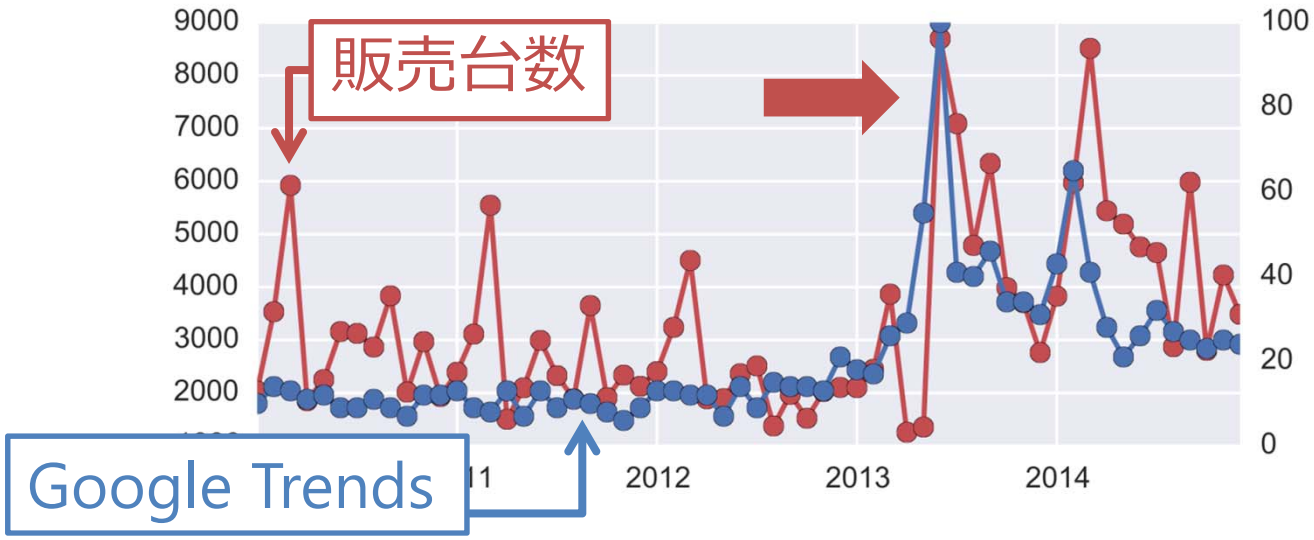


三菱 eK の
販売台数

➡ の期間は
予測が難しい

本研究: 検索行動量も利用

- 本研究
  - Google での検索数や Wikipedia の記事閲覧数 (検索行動量) を予測モデルに組み込む
  - ピークやトレンドのより正確な予測が目標

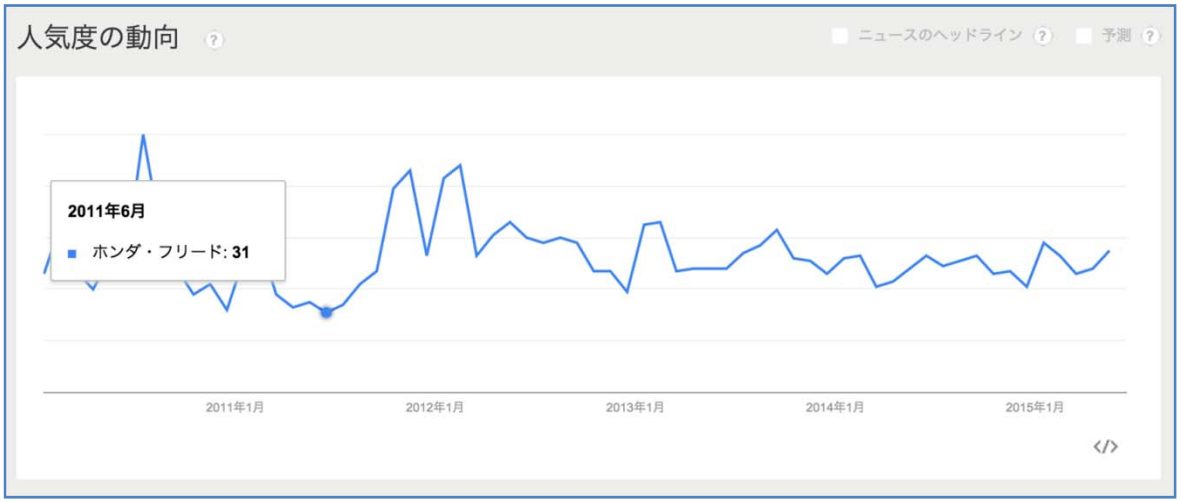


三菱 eK の
販売台数

➡ の期間で
検索量上昇！

注目データ①: 検索トレンド指数

- Google Trends★
  - 指定キーワードの検索頻度を取得可能
    （期間内での相対値 0-100; 日別又は月別）



★ Google トレンド: https://www.google.co.jp/trends/

注目データ②: Wikipedia 閲覧数

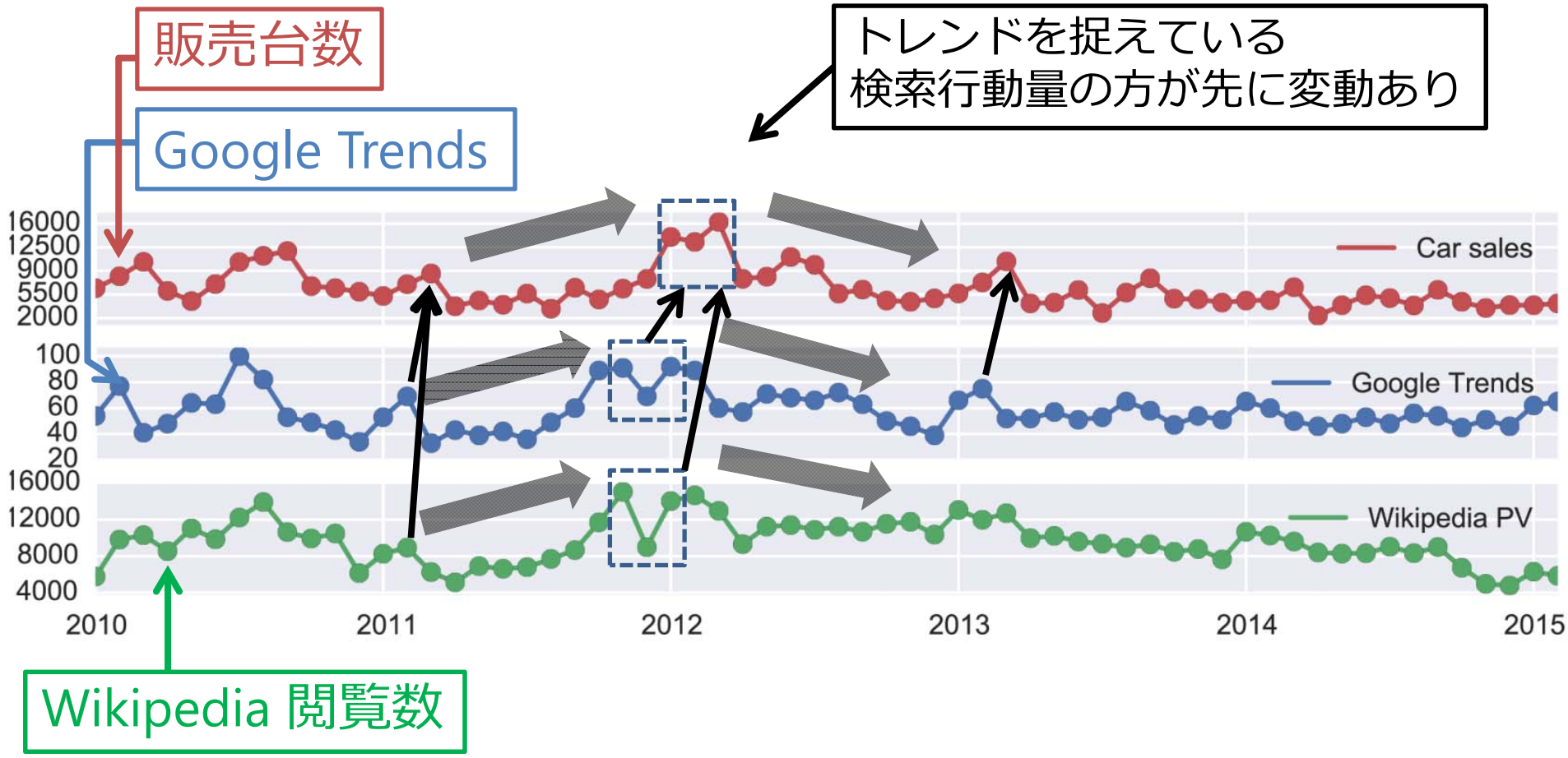
- Wikipedia
  - 検索結果の上位
  - 購買の参考になる情報量

  ➡ 検索数 ∝ 記事閲覧数

  - 公開データ！★

★ Page view statistics for Wikimedia projects:
http://dumps.wikimedia.org/other/pagecounts-raw/

[例] ホンダ・フリードの場合



データ検証における仮説・期待

- 仮説1: Google Trends が優位
  - 検索頻度データはトレンド把握の王様データ
  - Wikipedia PV は検索頻度の推定に過ぎない
- 仮説2: Wikipedia PV が優位
  - 検索は情報探索の入口
  - 情報に興味があれば Wikipedia を閲覧する

データの収集 (1/2)

- データの概要
  - 車種: 22車種
    (販売台数が多く、現時点でも発売されている車種)
  - 期間: 2010年1月 ~ 2015年2月

プリウス, カローラ, パッソ, ノア, ヴィッツ, ヴォクシー, ヴェルファイア, ウィッシュ, フィット, フリード, ステップワゴン, セレナ, キューブ, ノート, モコ, デミオ, スイフト, ワゴンR, ムーヴ, ミラ, タント, eK

データの収集 (2/2)

- 自動車販売台数
  - 各機関が公開する新車登録台数データを利用
    - 日本自動車販売協会連合会 (普通・小型乗用車)
    - 全国軽自動車協会連合会 (軽自動車)
- Google Trends
  - 「車名+メーカー名」を検索キーワードと考える
- Wikipedia 閲覧数
  - 対応する記事の閲覧数を閲覧数統計データから集計（リダイレクトページの閲覧数を含む）

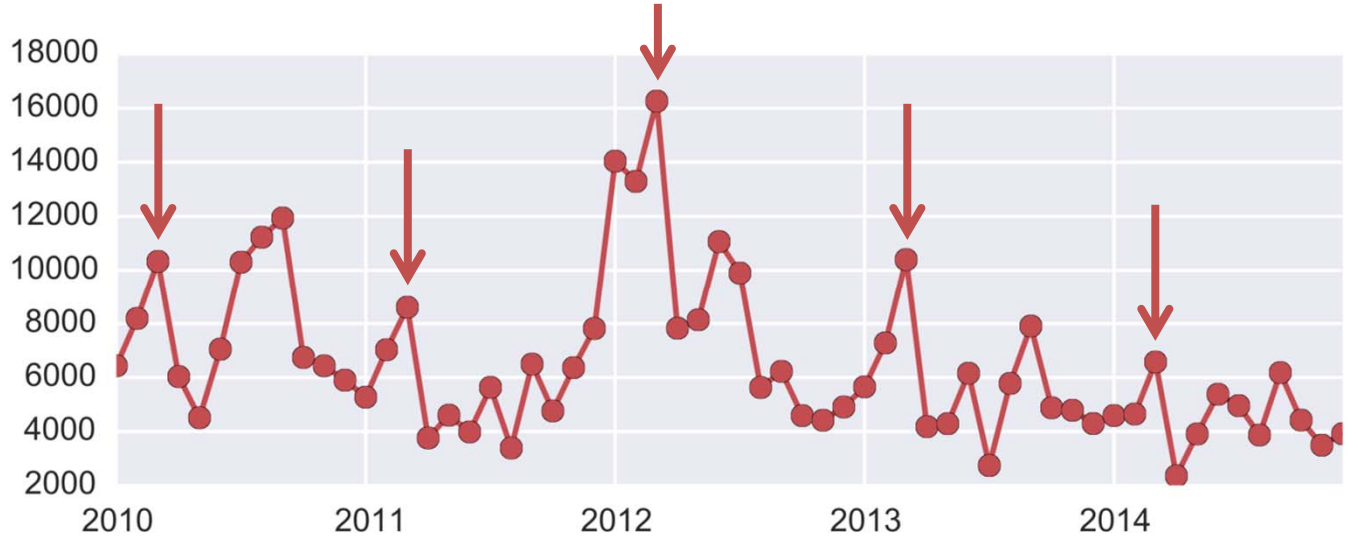
販売台数と検索行動量間の相関

- 相関係数を計算
  - 検索行動量の系列を1ヶ月・2ヶ月先行させた場合についても算出
- 一部の車種については相関有り
  - 相関係数が 0.4 以上であった車種数

検索行動量	同時期	-1ヶ月	-2ヶ月
Google Trends	7	10	4
Wikipedia 閲覧数	6	6	2

季節・トレンド成分を分解する (1/2)

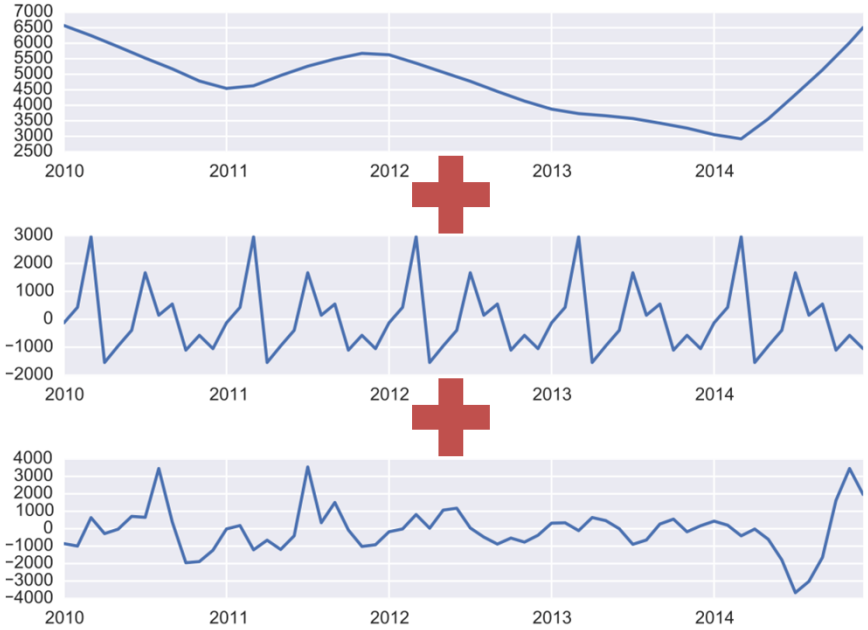
- 自動車販売台数は周期性有り



その他の要因を検索行動量から予測させたい

季節・トレンド成分を分解する (2/2)

- 各系列を STL[Cleveland 90] により分解



トレンド成分

＋

季節成分

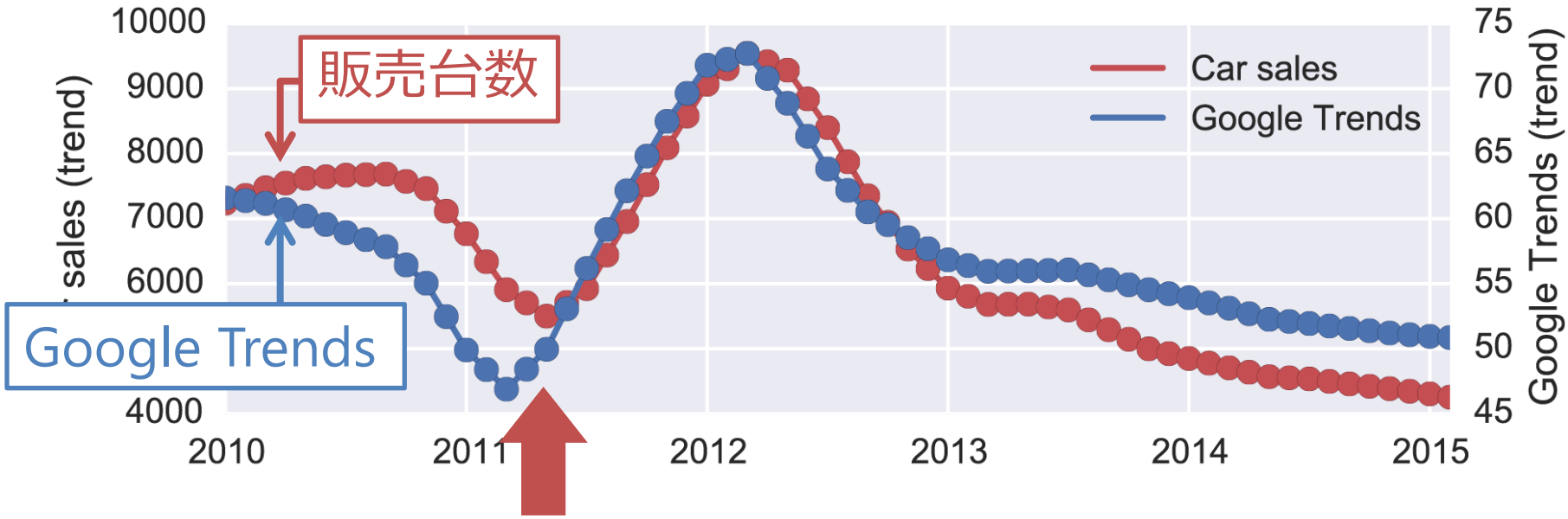
＋

残差

[Cleveland 90] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: A seasonal-trend decomposition procedure based on loess, Journal of Official Statistics, Vol. 6, No. 1, pp. 3–73 (1990)

トレンド成分における相関 (1/2)

- ホンダ・フリードのトレンド成分比較



  - フリードの場合、Google Trends が1~2ヶ月程先行してトレンドの転換を捉えている

トレンド成分における相関 (2/2)

- 特に相関が高い車種 (トレンド成分間)

検索行動量の先行期間

車種	Google Trends			Wikipedia 閲覧数		
	同時期	-1ケ月	-2ケ月	同時期	-1ケ月	-2ケ月
デミオ	0.65	0.67	0.67	-0.12	-0.1	-0.1
eK	**0.84**	0.77	0.69	0.62	0.58	0.54
フリード	**0.81**	**0.85**	**0.83**	0.5	0.44	0.35
ミラ	0.64	0.62	0.57	0.79	**0.84**	**0.89**
ムーブ	**0.88**	**0.84**	**0.8**	0.67	0.66	0.64
ノート	**0.97**	**0.94**	**0.9**	**0.9**	**0.92**	**0.92**
パッソ	**0.9**	**0.82**	0.74	0.27	0.22	0.16
ワゴンR	0.72	0.71	0.69	0.19	0.28	0.36

太字は相関係数が 0.8 以上、赤字下線は相関係数が 0.9 以上

予測手法の概要

- ベースライン
  - トレンド・季節要因に分解し、過去の時系列のみから予測
- 提案手法
  - トレンド・季節要因に加え、検索行動量を利用して予測

検索行動量活用の有効性を検証

実験条件 (1/2)

デミオ, eK, フリード, ミラ,
ムーブ, ノート, パッソ, ワゴンR

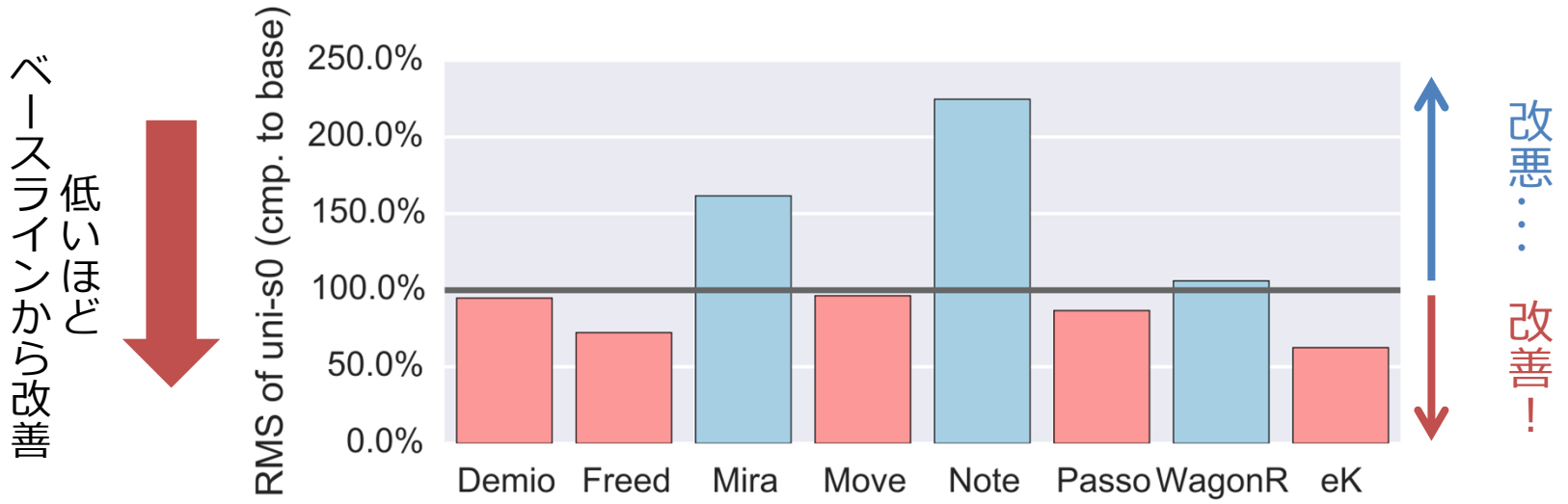
- 実験データ
  - 車種　　: 検索行動量と相関が高い 8 車種
  - 学習期間　: 2010年1月 ~ 2013年8月
  - テスト期間 : 2013年8月 ~ 2015年1月 (1.5年)
- 評価指標
  - 実販売台数に対する予測販売台数の比率の RMS (誤差二乗和の平均)
    → 低い値ほど性能が良いことを示す

実験条件 (2/2)

- 各月の予測 (一期先予測)
  1. 予測月の1ヶ月前までのデータに基づき
     学習期間で構築したモデルの内部状態系列を求める
     (カルマンフィルタ)
  2. 求めた内部状態系列を利用し、予測月における
     販売台数を求める

実験結果：シフト期間を全車種固定

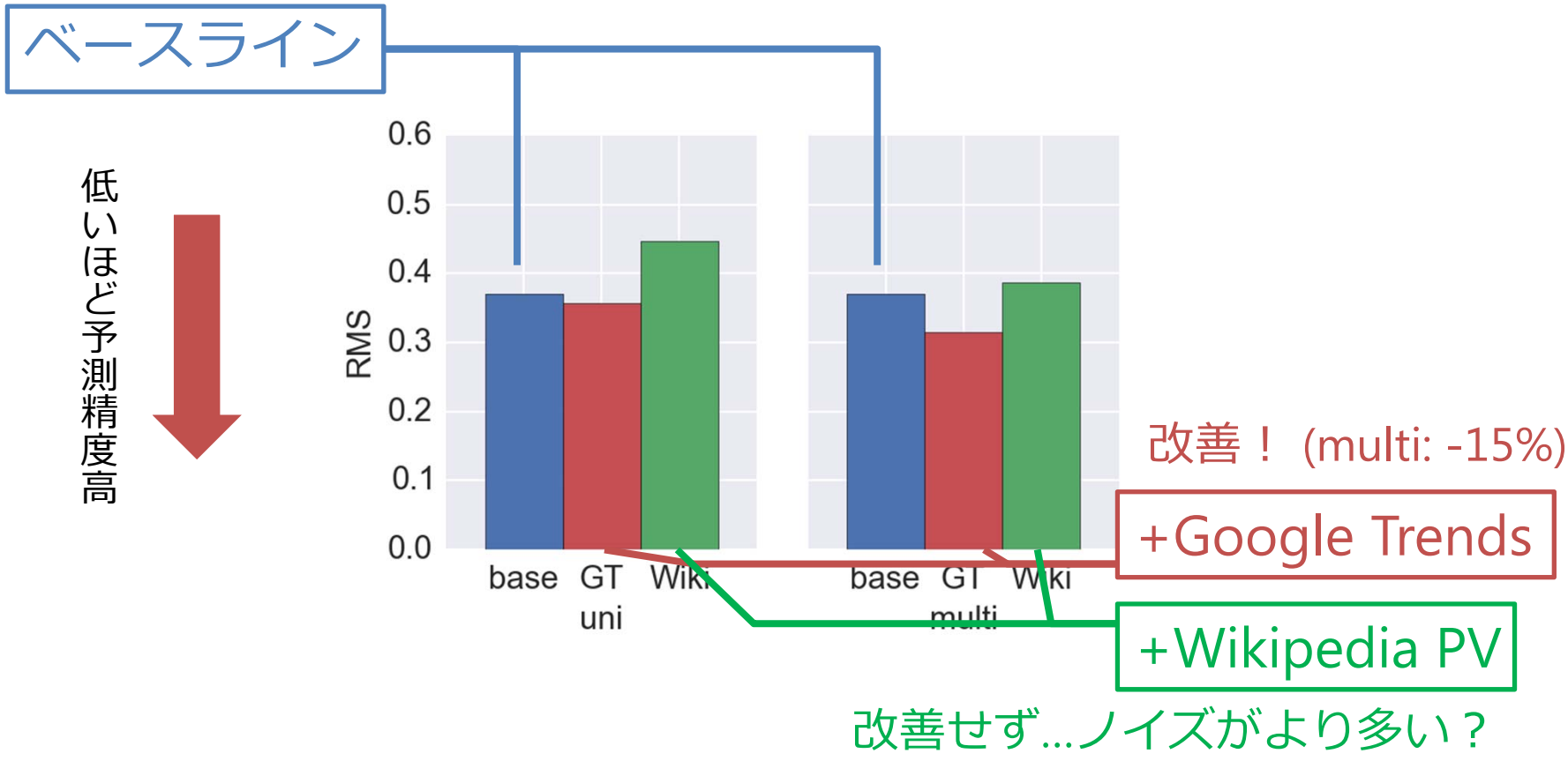
- ベースラインと uni モデルの比較
  (全ての車種で検索行動量系列を -0ヶ月 シフト)



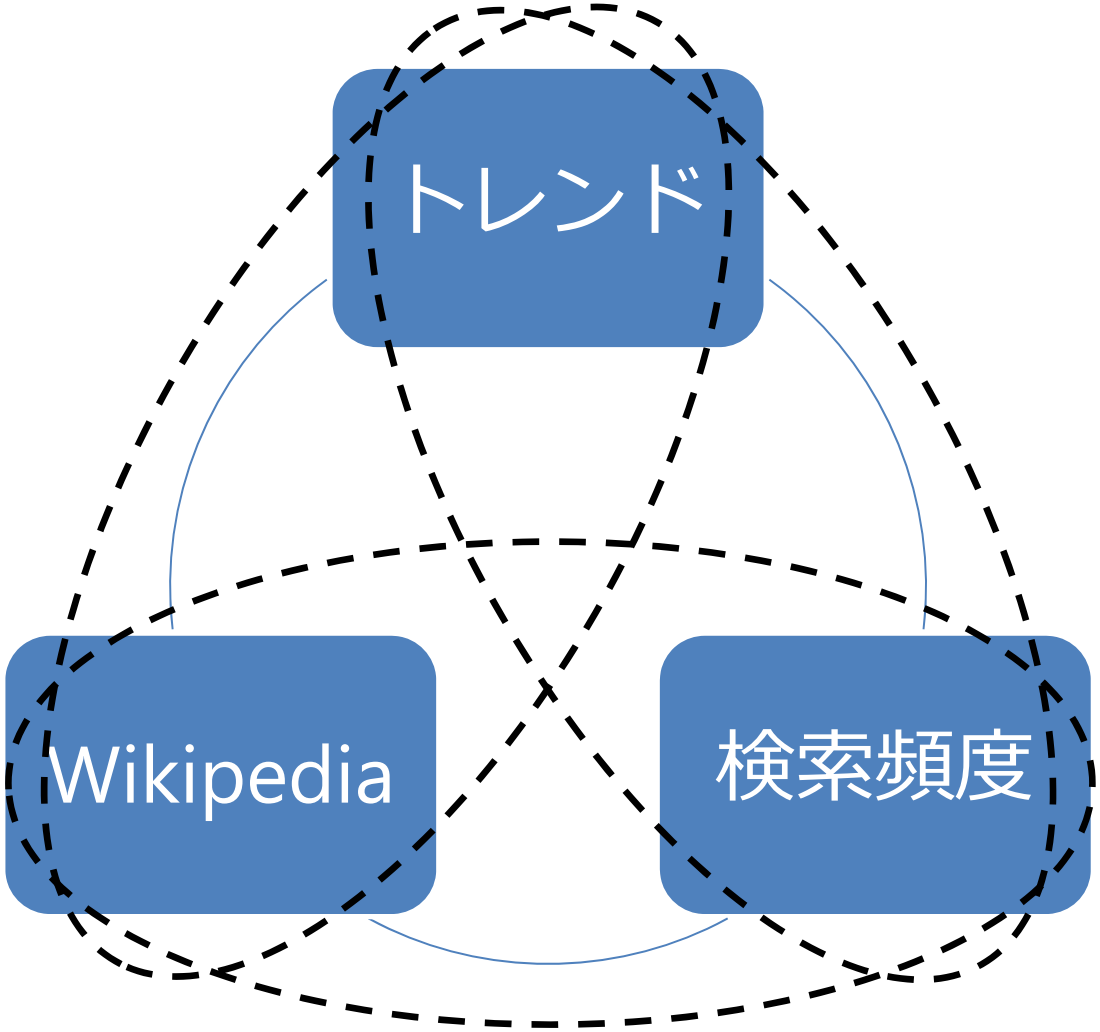
  - 改善する車種とそうでない車種に分かれる
    (最適なシフト期間は車種によって異なるため)

実験結果：シフト期間を車種毎可変

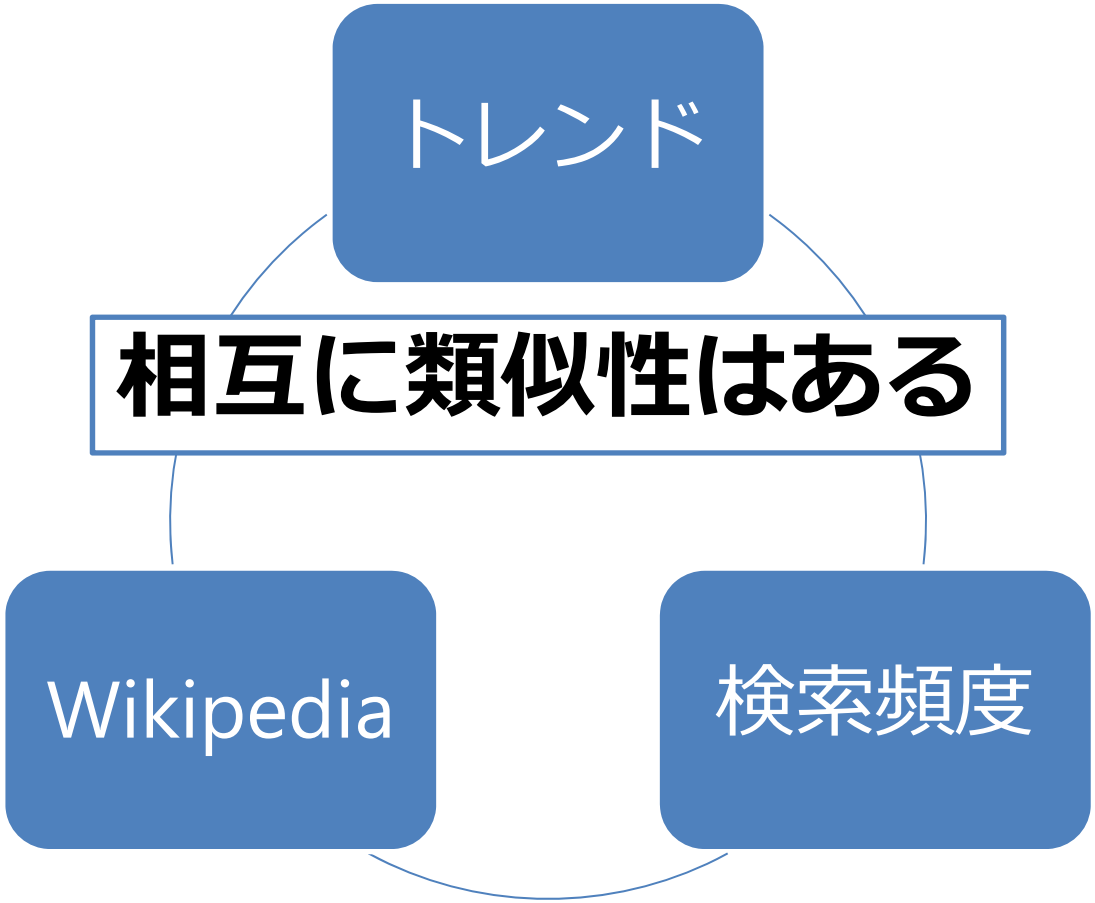
- 車種毎に最適シフト期間を学習データから推定



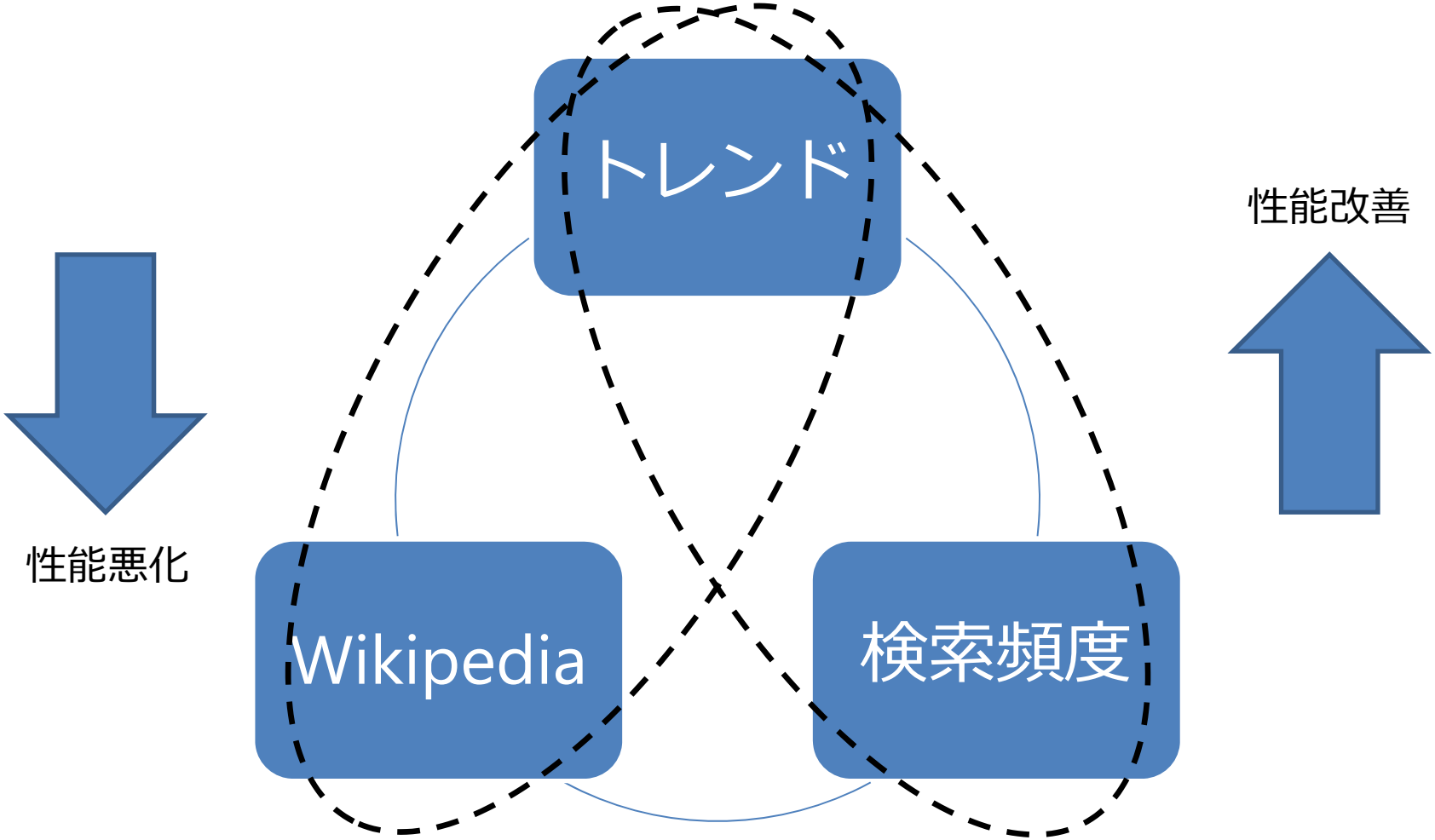
三角関係



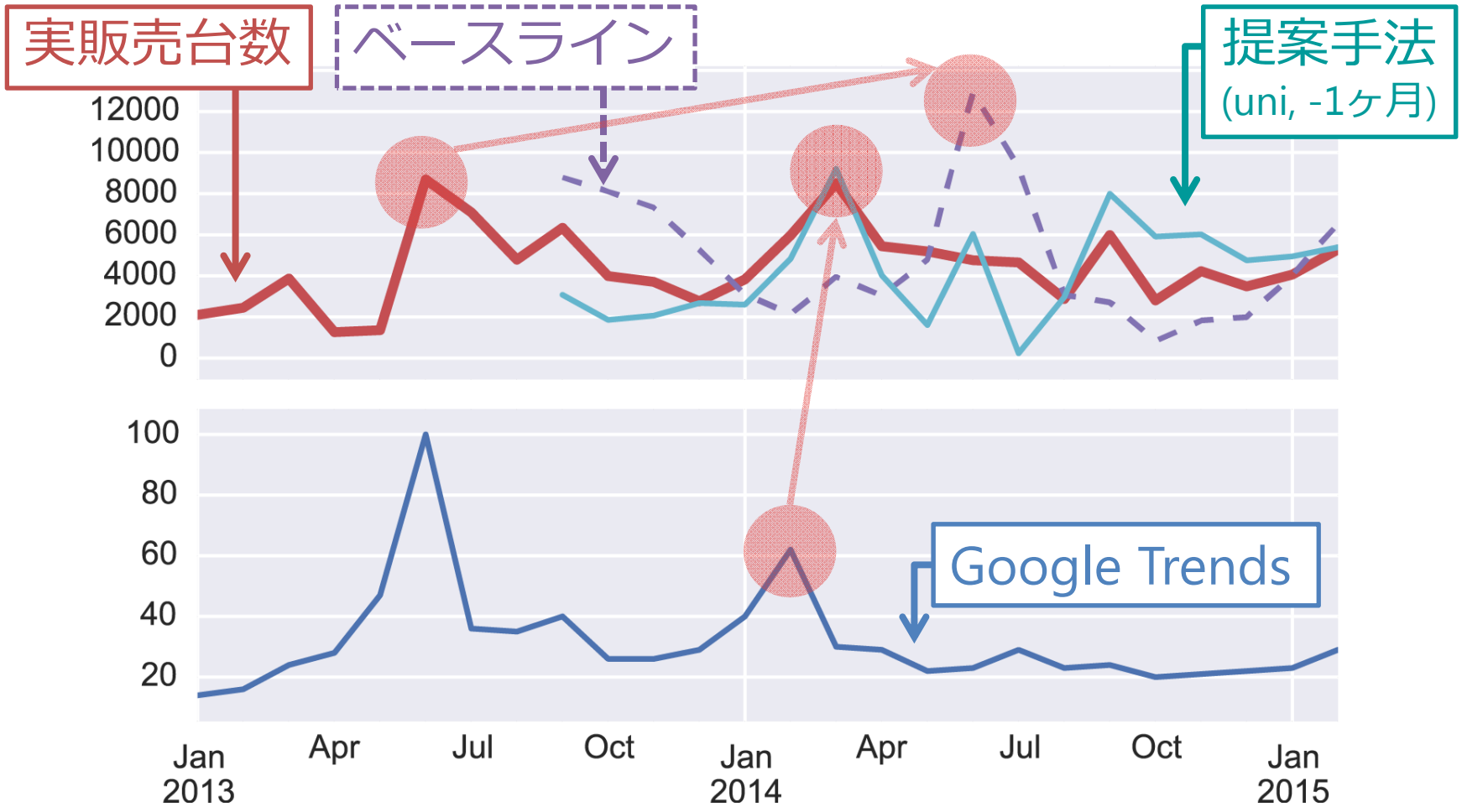
三角関係



三角関係



実験結果 – 三菱 eK の予測例

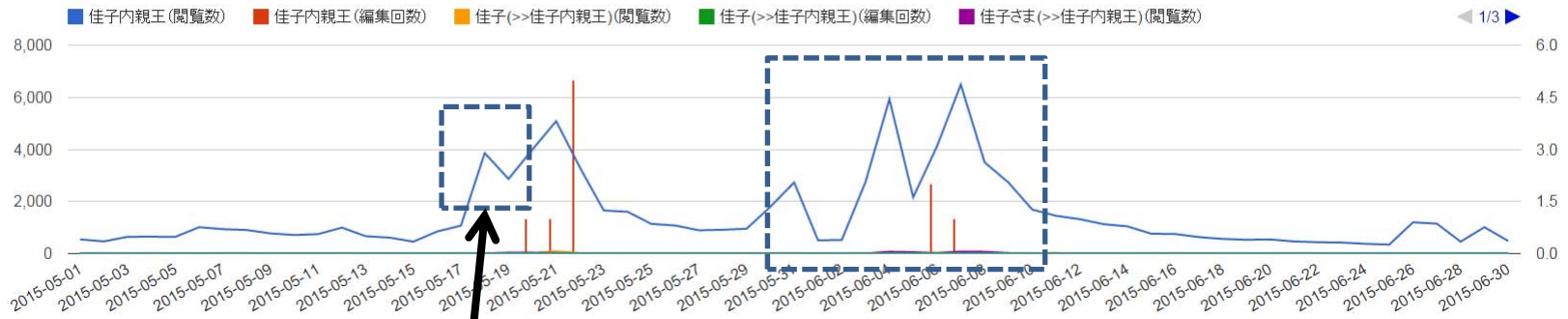


まとめと今後の課題

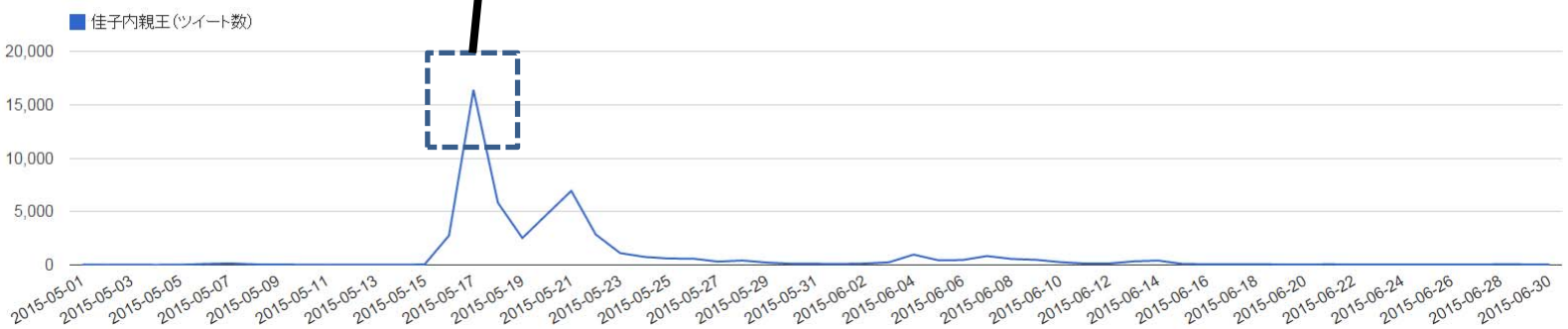
- まとめ
  - 一部の車種において自動車販売台数と検索行動量の間に相関があった
  - 一部の車種において Google Trends を利用した提案モデルにより予測性能が改善
- 今後の課題
  - Wikipedia 閲覧数を用いたモデルの分析
  - ソーシャルメディア投稿の活用

ソーシャルメディア投稿（1/2）

Wikipedia



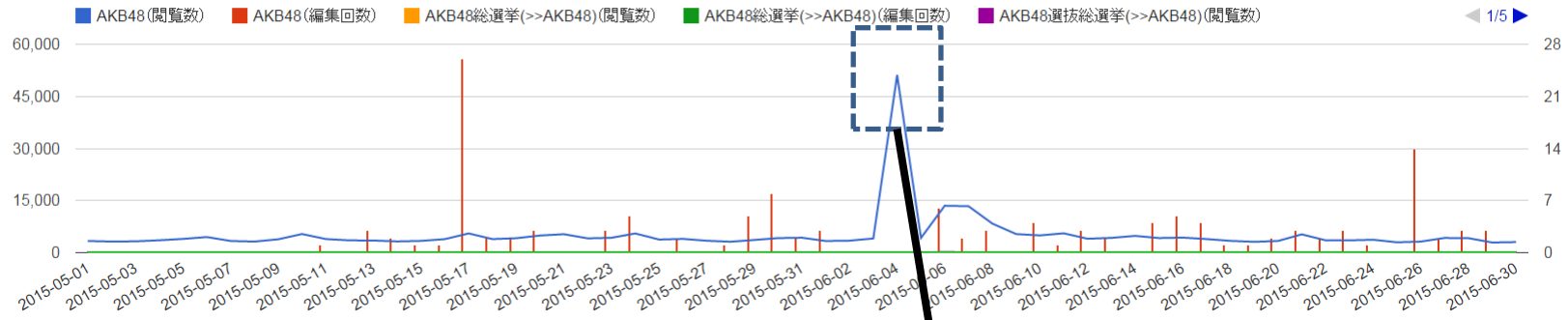
Twitter



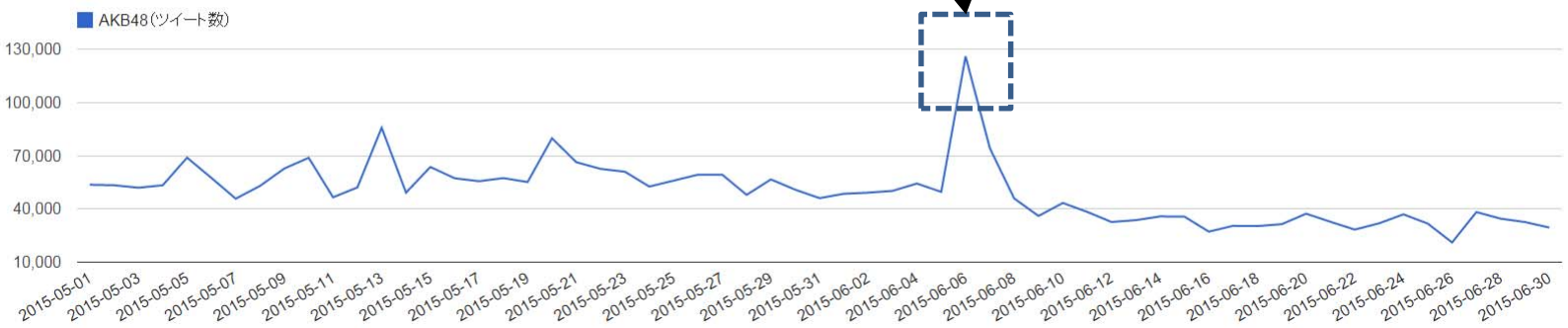
佳子内親王（Twitter先行型）

ソーシャルメディア投稿（2/2）

Wikipedia



Twitter



AKB48（Wikipedia先行型）

詳細資料

- 状態空間モデルを用いた検索トレンドとページビューからの自動車販売台数の予測
  - 角田孝昭, 吉田光男, 津川翔, 山本幹雄
  - 第29回人工知能学会全国大会

WIKIPEDIA データ集計ツール

近日中に公開予定

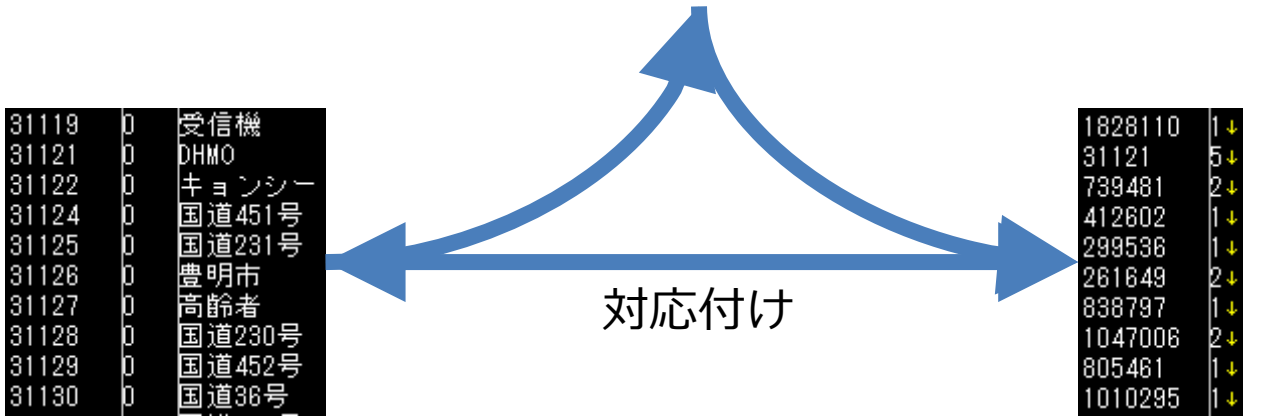
- GitHub で公開予定
  - https://github.com/ceekz

- ツール概要
  - 見出し語リスト生成
    - 見出し語、作成日、カテゴリ、リダイレクト、...
  - ページビューデータ生成
    - 時間別、日別、月別

ページビューデータ生成

```
1283 ja %E3%81%AA%E3%81%AE%E3%81%AF 2 2↓
1284 ja %E3%81%AA%E3%81%AF_%28%E5%88%97%E8%BB%8A%29 1 1↓
1285 ja %E3%81%AA%E3%81%BE%E3%81%9A 3 3↓
1286 ja %E3%81%AA%E3%81%BE%E3%81%AF%E3%81%92 3 3↓
1287 ja %E3%81%AA%E3%82%81%E3%82%8D%E3%81%86 1 1↓
1288 ja %E3%81%AA%E3%82%89%E3%81%AF%E3%81%97%E3%81%BF%E3%81%8D 1 1↓
1289 ja %E3%81%AA%E3%82%8B%E3%81%9F%E3%82%8B 3 3↓
1290 ja %E3%81%AA%E3%82%8B%E3%81%BF 1 1↓
1291 ja %E3%81%AA%E3%82%8B%E3%83%88%E3%83%A2! 1 1↓
1292 ja %E3%81%AA%E3%82%8B%E3%83%88%E3%83%A2%21 1 1↓
1293 ja %E3%81%AA%E3%82%8F%E3%81%A8%E3%81%B3 1 1↓
1294 ja %E3%81%AA%E3%82%93%E3%81%A7%E3%82%82Q 1 1↓
1295 ja %E3%81%AA%E3%82%93%E3%81%A8%E3%81%8B%E5%8B%95%E7%94%BB 4 4↓
```

配布データ（raw data）
（言語、ファイルパス、頻度）



ID ⇔ 見出し語

ID ⇔ 頻度

まとめ

まとめ

- 検索頻度推定のための予備調査
  - Wikipedia Page View との類似性を調査
  - 高い類似性が明らかに
    - 特に高頻度帯、週別・月別の粒度
- 自動車購買台数予測と検索行動量
  - 一部の車種において自動車販売台数と検索行動量の間に相関があった
  - 一部の車種において Google Trends を利用した提案モデルにより予測性能が改善
    → Wikipeda Page Views は NG ☹