アカデミア×ビジネスの研究者に聞く、データセットの魅力
～私が研究者を選んだ理由～

登壇者紹介

- 荒瀬　由紀（大阪大学）
- 加藤　誠　　（京都大学）
- 清田　陽司（株式会社LIFULL）
- 平手　勇宇（楽天株式会社）

荒瀬　由紀（大阪大学）

＜略歴＞

- 徳島県出身
- 2006年 大阪大学工部電子情報エネルギー工学科卒業
- 2007年 同大学院情報科学研究科 博士前期課程修了
- 2010年 同博士後期課程修了 博士（情報科学）
- 2010～2014年 Microsoft Research Asia, Associate Researcher
- 2014年～ 大阪大学大学院情報科学研究科准教授

大学教員の面白さ

知的好奇心の探求

誰もやっていない

上手く行く保証がない

（すぐには）役に立たない

ことにも挑戦できる

大阪大学 OSAKA UNIVERSITY

OPEN 2021

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page
{sergey, page}@cs.stanford.edu
Computer Science Department, Stanford University, Stanford, CA 94305

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date.

Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords: World Wide Web, Search Engines, Information Retrieval, PageRank, Google

1. Introduction

(Note: There are two versions of this paper -- a longer full version and a shorter printed version. The full version is available on the web and the conference CD-ROM.)
The web creates new challenges for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. People are likely to surf the web using its link graph, often starting with high quality human maintained indices such as Yahoo! or with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead automated search engines. We have built a large-scale search engine which addresses many of the problems of existing systems. It makes especially heavy use of the additional structure present in hypertext to provide much higher quality search results. We chose our system name, Google, because it is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search engines.

多様な分野の研究者

との出会いが多く、

自由に共同研究

ができる

- 異分野研究者との共同研究
  - 英語学習者支援：言語学の先生と協力
  - 認知症介護支援：医学、音楽療法の先生と協力
- CS外の研究者と知り合うチャンスが多い
  - 学内外での紹介
  - 勉強会、イベントへの参加

オープンな環境

- 研究成果の公開
  - 一生懸命作ったアノテーションデータ
  - 一生懸命作ったデモシステム

- 企業データを利用した研究もできる

企業データの面白さ

圧倒的リアリティ

- 人間が自然に書いたテキストはとても貴重
- 人工的に作るのはとても難しい
- 感情分析に使わせてもらっています
  - 楽天市場 レビューデータ
  - 不満調査データ

- 世界とつながってる感
  - 技術が本当に役に立つか検証できる
  - 実サービスへの応用可能性
- 新たな研究シーズ
  - 欠損値、不正値：ロバスト化
  - 桁違いのサイズ：高速化、省メモリ化

加藤　誠（京都大学）

＜略歴＞

- 1985年 静岡県出身
- 2008年 京都大学工学部情報学科卒業
- 2010年 Microsoft Research Asia インターンシップ（8ヶ月）
- 2011年 Microsoft Research インターンシップ（3ヶ月）
- 2012年 京都大学情報学研究科 博士後期課程修了
- 現在 京都大学 国際高等教育院
  データ科学イノベーション教育研究センター 特定講師

概要

A）なぜ大学教員を選んだのか？
→1. 漫画家と研究者のアナロジー

B）なぜ企業の研究者ではなく大学教員か？
→2. 驚くほど研究が自由

C）企業データのすごいところ
→3. データセットは研究の救世主
→4. 情報検索には実ユーザのデータが必要

1. 漫画家と研究者のアナロジー

漫画家

- いつも締切に追われている
- あり得ないことを形にする
- ストーリーも絵もいる
- 編集会議にかけられる
- 実力主義
- 友情・努力・勝利

研究者

- いつも締切に追われている
- あり得ないことを形にする
- ストーリーも手法もいる
- 編集会議にかけられる
- 実力主義
- 共同研究・努力・採択

2. 驚くほど研究が自由

修士課程

- 卒業できる研究・指導教員の暗黙的誘導

博士課程

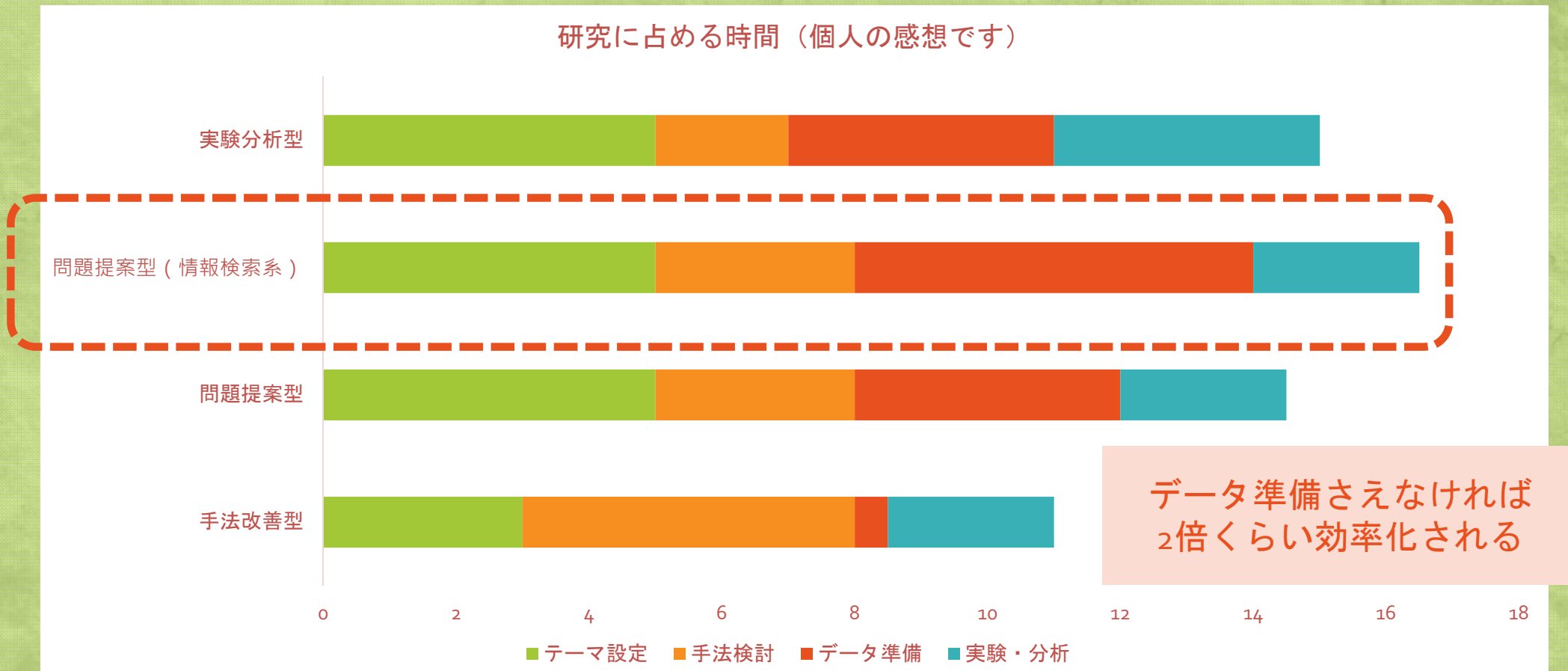
- 博士が取得できる研究・論文になりやすい研究・指導教員の暗黙的誘導

大学教員初期

- 論文になりやすい研究・共同研究の研究・科研の課題に関する研究

大学教員初期後

3. データセットは研究の救世主



私が関わってきたデータセット

- **Experimental data for Optimizing Search Result Presentation**
- **Test Collections for Conversational Relevance Feedback**
- **Cognitive Relevance Dataset**
- **Geographic Object Retrieval Dataset (GORD)**
- **Query by Analogical Example (QAE) Dataset**

- **NTCIR-9 INTENT-1（Web検索）**
- **NTCIR-10 INTENT-2 （Web検索）**
- **NTCIR-10 1CLICK-2 （モバイル検索）**
- **NTCIR-11 IMine （Web検索）**
- **NTCIR-11 MobileClick （モバイル検索）**
- **NTCIR-12 IMine-2 （Web検索）**
- **NTCIR-12 MobileClick-2 （モバイル検索**
- **NTCIR-13 OpenLiveQ （質問検索）**

情報検索は人による判定が
必要なことが多く労力が大きいため
伝統的に皆でデータセットを作る

4. 情報検索には実ユーザのデータが必要

- 特にインタラクションデータ（クリックスルーデータ）
  - マイクロソフトのクリックスルーデータを利用
    - Makoto P. Kato, Tetsuya Sakai, Katsumi Tanaka (2013) When Do People Use Query Suggestion? A Query Suggestion Log Analysis, Information Retrieval 16(6), pp. 1-22.
    - Makoto P. Kato, Tetsuya Sakai, Katsumi Tanaka (2012) Structured Query Suggestion for Specialization and Parallel Movement: Effect on Search Behaviors, WWW2012, pp. 389--398
    - Makoto P. Kato, Tetsuya Sakai, Katsumi Tanaka (2011) Query Session Data vs. Clickthrough Data as Query Suggestion Resources, ECIR 2011 Workshop
  - ヤフーの知恵袋検索におけるクリックスルーデータを利用
    - Tomohiro Manabe, Akiomi Nishida, Makoto P. Kato, Takehiro Yamamoto, Sumio Fujita (2017) A Comparative Live Evaluation of Multileaving Methods on a Commercial cQA Search, SIGIR 2017
    - Makoto P. Kato, Takehiro Yamamoto, Tomohiro Manabe, Akiomi Nishida, Sumio Fujita (2017) Overview of the NTCIR-13 OpenLiveQ Task, NTCIR-13 Conference

NTCIR-13 OpenLiveQ Task

参加チームのシステム出力をinterleaving*
して提示し，ユーザのクリックでシステム評価

これまで大学などでは利用できなかった
貴重なデータの提供

- **1,000**訓練クエリ
- 各クエリに対し上位**1,000**件の検索結果
- **ユーザ情報付き**
  クリックスルーデータ
  - 各検索結果をどれくらいの人がクリックしたか
  - クリックした人のユーザ属性
    (年齢，性別，職業, etc.)

*A/Bテストより10~100倍効率的なランキング評価手法　インターリービング（Interleaving）のまとめと実践
https://qiita.com/mpkato/items/99bd55cc17387844fd62

まとめ

A）なぜ大学教員を選んだのか？
→1. 漫画家と研究者のアナロジー
研究者は漫画家くらい夢がある職業

B）なぜ企業の研究者ではなく大学教員か？
→2. 驚くほど研究が自由
本当に驚いている

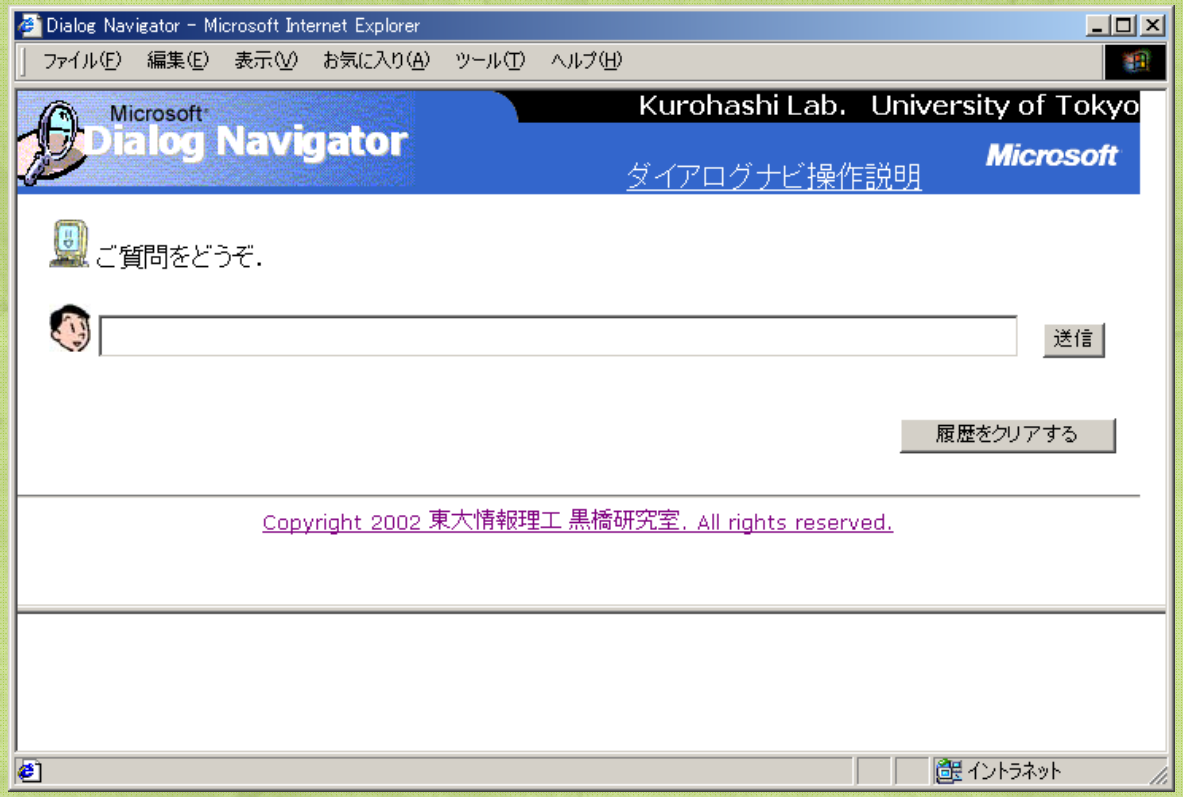
C）企業データのすごいところ
→3. データセットは研究の救世主
特に研究を始めて行う人には救世主
→4. 情報検索には実ユーザのデータが必要
特にユーザデータがすごい

清田　陽司（株式会社LIFULL）

＜略歴＞

- 1975年 福岡県出身
- 1998年 京都大学工学部電気工学第二学科（長尾研）卒業
- 2004年 京都大学大学院情報学研究科博士課程（現 黒橋研）修了
- 2004～2012年 東京大学情報基盤センター 助教→特任講師
- 2007年 株式会社リッテル設立に参画
  - 2007年～ 上席研究員（兼業）
  - 2010年～ 取締役CTO
- 2011年～ 株式会社ネクスト（現 LIFULL）上席研究員

博士課程時代の研究：ダイアログナビ（マイクロソフト日本法人との研究）



2002年～2005年にマイクロソフト社のWebサイト上で実運用

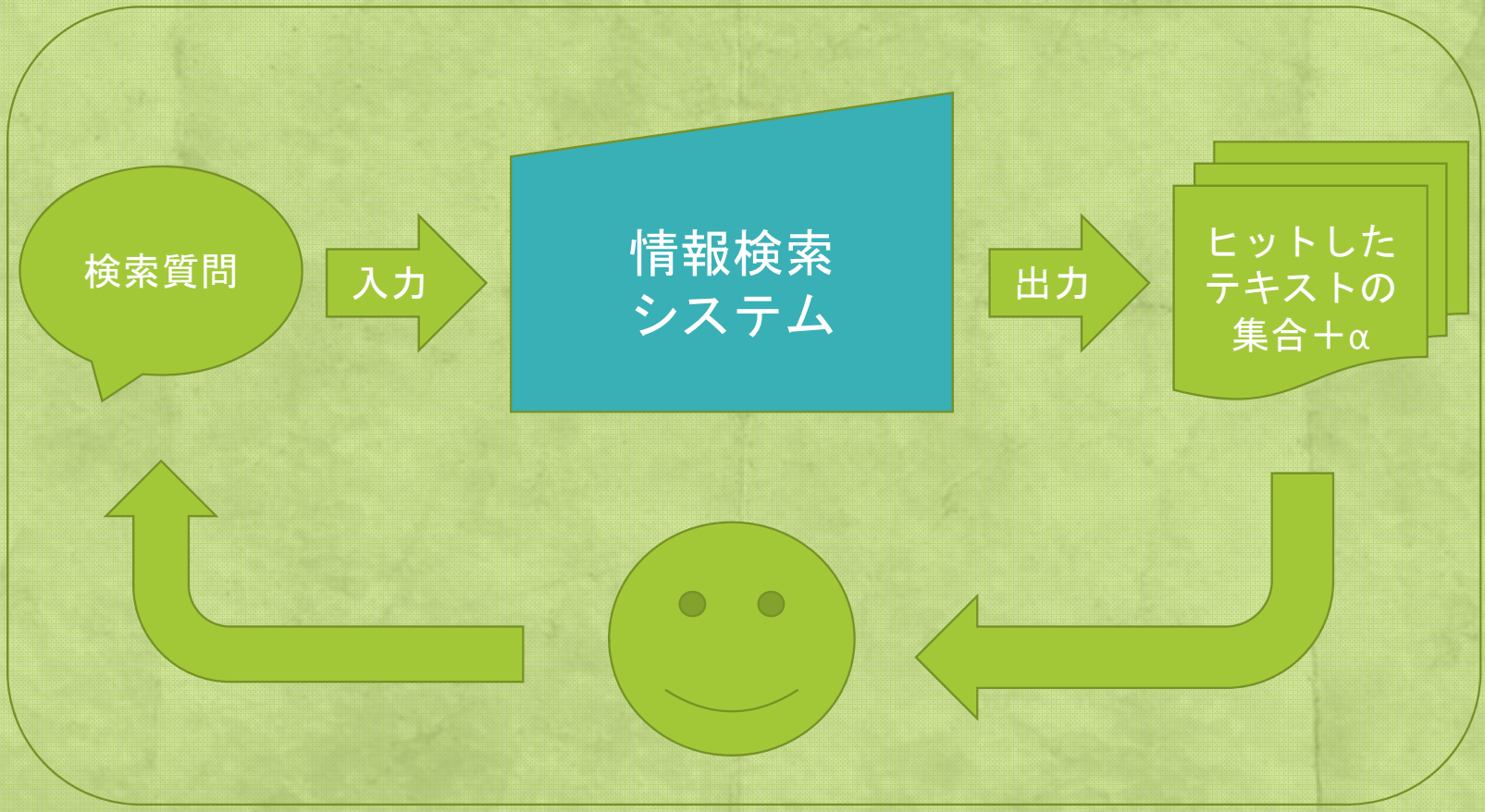
実運用
＝現実のユーザーのニーズや行動を知る

52件のテキストが見つかりました。

- Windows 98を起動したときに、…というエラーが発生する
- Windows XPでアプリケーションを起動したときに、エラーが発生する
- インターネットにダイアルアップで接続しようとしたときに、…というエラーが発生する
- 印刷中にエラーが発生して印刷できない
- ……

テストコレクションによる検索システムの評価

現実のモデル



なぜ大学教員からビジネスの世界に移ったのか？

ニーズ主導でしか発見できない本質的な研究課題がある

- ただし、本質的な研究に踏み込むまでには忍耐が必要

「生きた」データにどうやってアクセスするか？

- 「どんなデータをどうやって集めるか」自体が研究になりうる

企業で実データにアクセスして研究するということ

実データにアクセスして得られた知見はたくさんある

- 住まい探しユーザーの多様性
- ユーザーのニーズは移ろいやすい
- etc.

ただし、評価基準の確立は非常に大変

- 論文を書くのはなかなか難しい
- ビジネス上の貢献を短期で測るか、中長期で測るか？

データセットとして外部に提供することで価値を生み出す取組み

→ LIFULL HOME'Sデータセット提供開始 (2015年～)

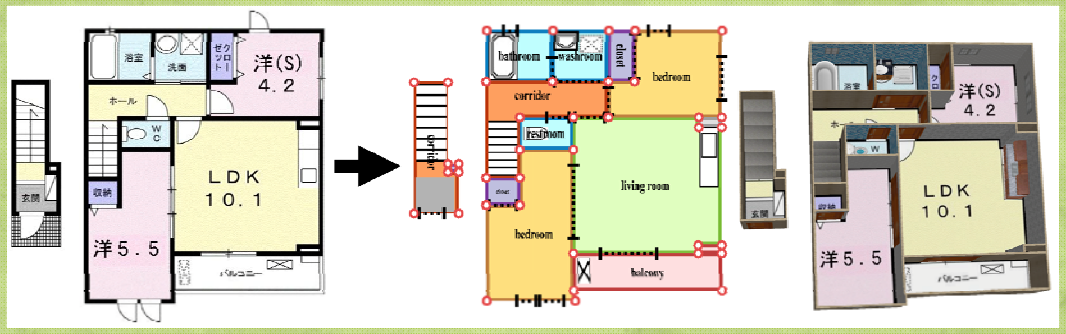
不動産物件画像・間取り図データを利用した研究によるイノベーション創出

不動産会社が入稿する画像の不整合検出

入稿画像	登録項目	整合率	判定結果
	キッチン	キッチン 70.3% 玄関 23.7% 設備 2.804%	○
	キッチン	キッチン 97.3%	◎
	キッチン	キッチン 0% 玄関 0%	×

ユーザーに提供する
不動産情報品質の向上

間取り図からの3Dモデル生成

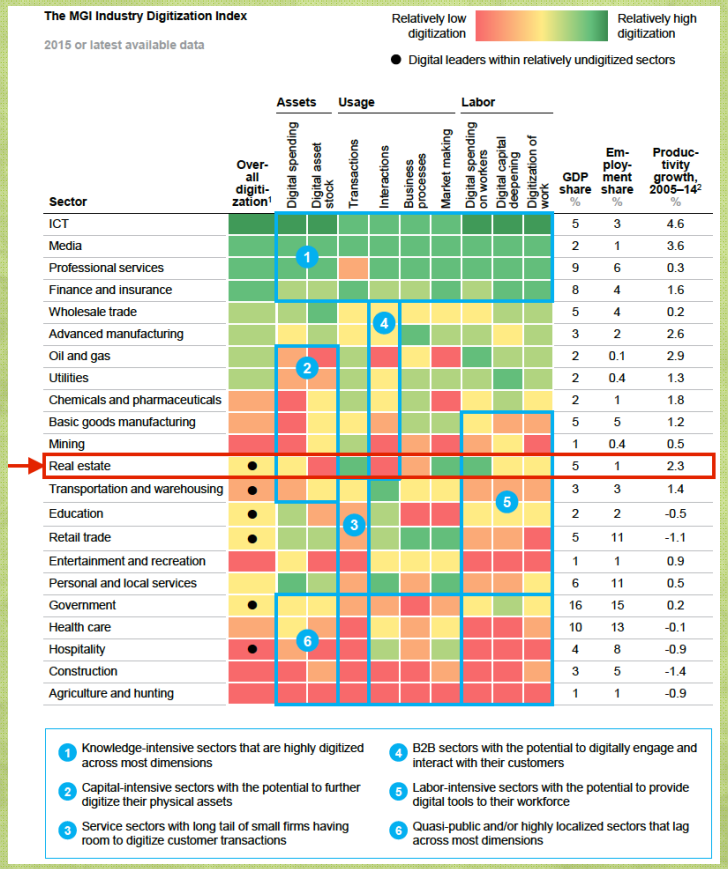


古川康隆准教授（Simon Fraser Univ.）による
LIFULL HOME'Sデータセット利用研究
ICCV 2017に採択

VRコンテンツなどの
新たなUXの提供

データサイエンスアワード2017
ファイナリストに選出

不動産データの面白さ



- データのマルチモダリティ
  - 画像、価格、属性データ、位置情報などがセットになっている
- 不動産はこれからデジタル化の浸透が期待される分野
- 学際的な研究テーマがたくさん転がっている
  - 不動産学、経済学、建築学、都市学、…

出典: Manyika, James et al. Digital America: A tale of the haves and have-mores. McKinsey Global Institute. 2015
http://www.mckinsey.com/industries/high-tech/our-insights/digital-america-a-tale-of-the-haves-and-have-mores

平手　勇宇（楽天株式会社）

楽天技術研究所 インテリジェンスドメイングループ　マネージャー

機械学習，データマイニング，NLP関連の研究プロジェクトを統括しています．

＜略歴＞

- 1980年 富山県出身
- 2008年03月 早稲田大学大学院理工学研究科情報・ネットワーク専攻修了（博士・工学）
- 2006年04月～2009年03月 早稲田大学メディアネットワークセンター助手
- 2009年04月～　楽天技術研究所

＜学会活動＞

- IPSJ DBS研　運営委員・幹事，IEICE　DE研　専門委員など
- IPSJ　TOD　編集委員，IEICE　データ工学特集号編集委員など

楽天技術研究所

楽天の戦略的な研究開発部門　5拠点で100名以上.

Singapore

Tokyo

NY

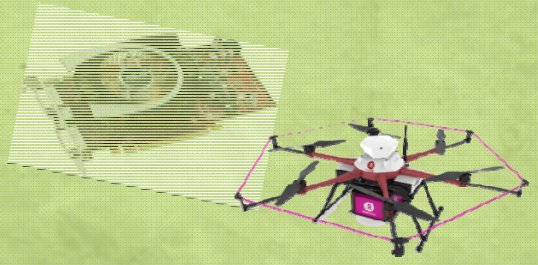
Paris

Boston

楽天技術研究所の研究分野

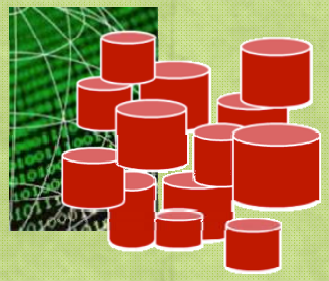
Power

- 分散コンピューティング
- 大規模並列処理
- IoT



Intelligence

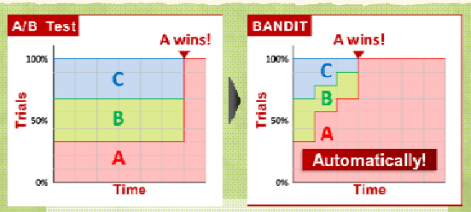
- 機械学習・深層学習
- 自然言語処理
- データマイニング



Reality

- ユーザインターフェイス
- ユビキタスコンピューティング
- 画像処理

インテリジェンスドメイングループの研究プロジェクト例

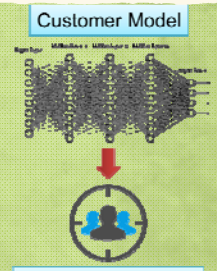


ABテストの最適化

レビューデータ解析

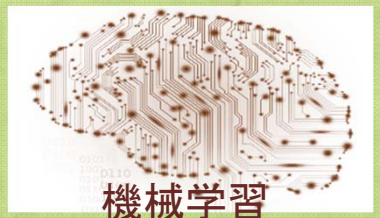
User Behavior Data
In Rakuten Ecosystem
(Huge and Complex)

Existing Service
Customers

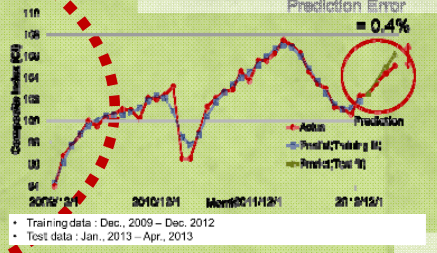


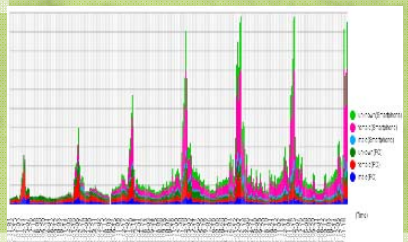
潜在顧客抽出

クーポン配信の最適化



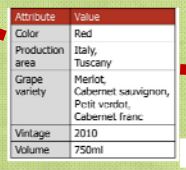
機械学習





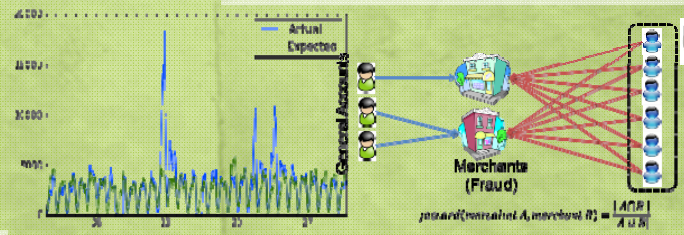
経済予測／需要予測

商品推薦システム



商品データの構造化

機械翻訳/音声認識



異常検知／不正検知

M1の11月まで，博士課程への進学という
選択肢は，考えてもいませんでした.

博士課程に進学を決めたきっかけ

ICDM2004 workshopでの研究発表 （M1の11月）

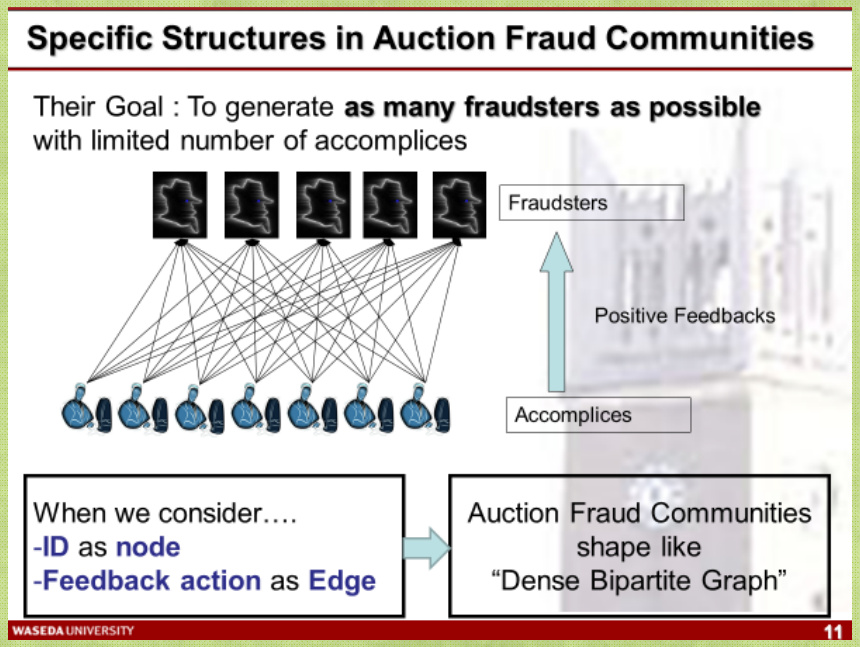
http://icdm04.cs.uni-dortmund.de/

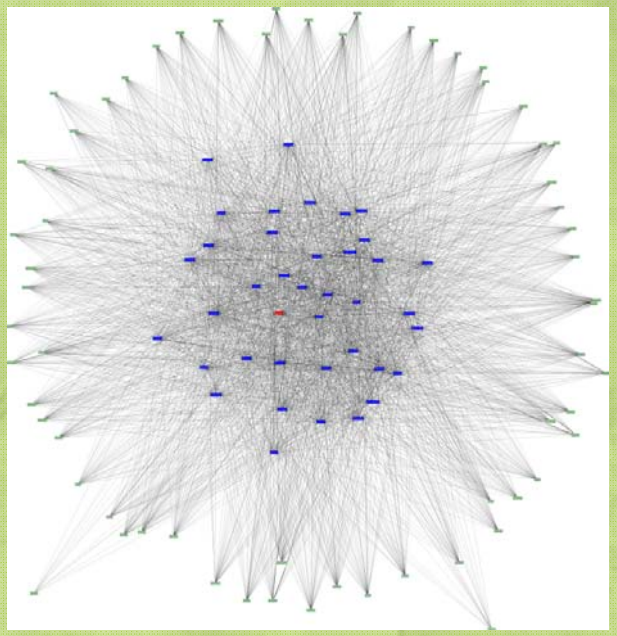
指導教員からの一言 （M1の12月）

※http://www.yama.info.waseda.ac.jp/~yamana/yamana_eng2015.htmより転載

企業研究所を目指すことになったきっかけ（１）

- Yahoo! Japan様との共同研究（オークションからの不正アカウント抽出）

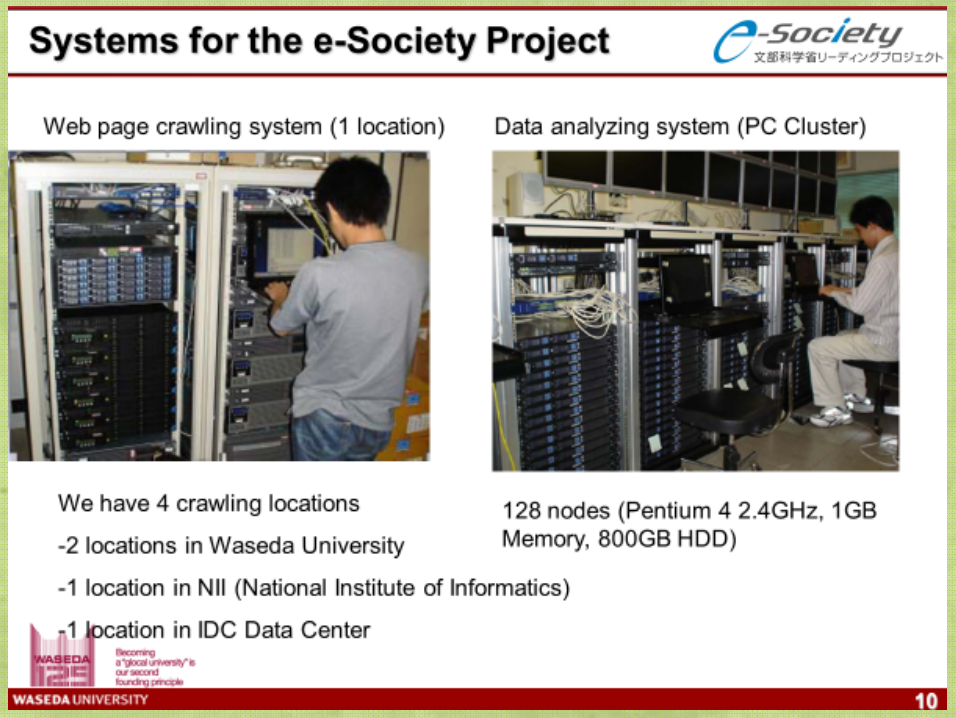




2007年度 IPSJ CS領域奨励賞受賞

企業研究所を目指すことになったきっかけ（２）

- 全世界のWebをクローリングして，解析して実施



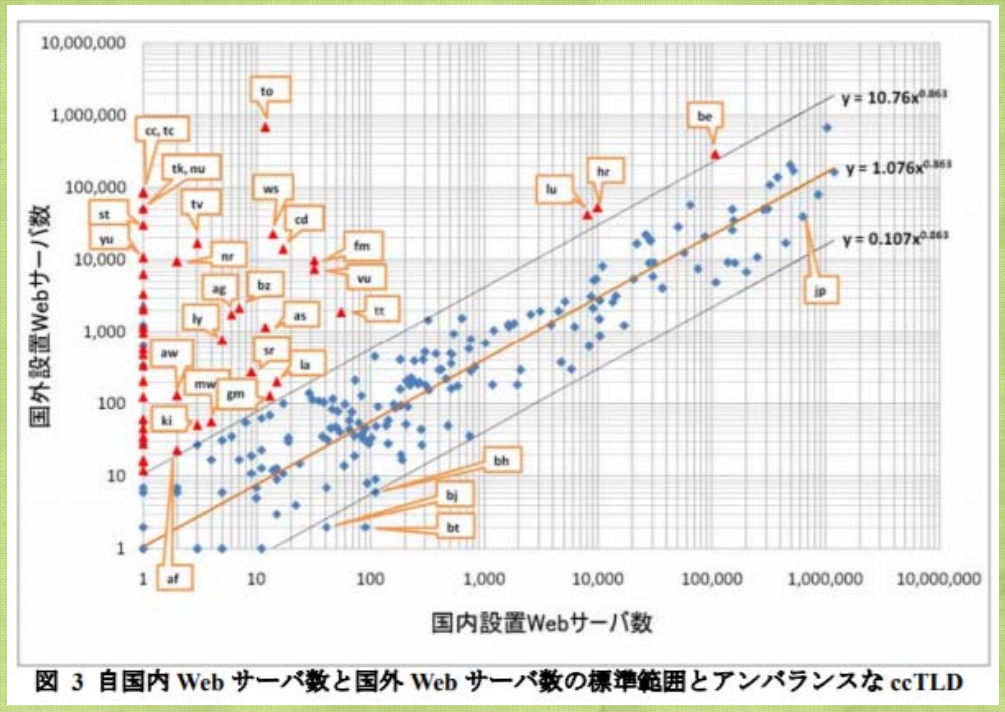
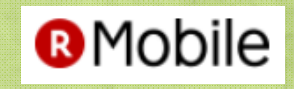


図 3 自国内 Web サーバ数と国外 Web サーバ数の標準範囲とアンバランスな ccTLD

第9回 DBSJ論文賞受賞

企業データの面白さ

- Query Log Data, Click Through Data, Purchase History Data等は，どの企業に入ってもすぐにアクセスできる環境が整えられています．
- 入社前には想像もつかなかったデータを対象としたプロジェクトに従事することも．



- （何より大事なのは）考えたアルゴリズムを適用する環境がすぐ近くにあること．

質疑応答