

# 開発運営を続ける楽しさ

ニュース検索サイト CEEK.JP NEWS の開発を通じて

吉田光男

筑波大学大学院システム情報工学研究科

有限会社てっくてっく / CEEK.JP

静岡大学 大学院GP

2010/7/5

- 自己紹介
- Web検索エンジンの基礎・課題
- 事例
  - CEEK.JP
  - CEEK.JP NEWS
  - 有限会社てっくてっく
- キャリアを考える
- おわりに

- 吉田光男（よしだみつお）
- 所属
  - システム情報工学研究科 CS専攻
    - 自然言語処理 on the Web 研究室
  - 産学リエゾン共同研究センター 客員研究員
  - 有限会社てっくてっく
- 1984年生（25歳）
- 和歌山県出身

- 2003年 情報学類 入学
  - AC入試 (AO入試)
    - 図書館の情報をインターネットに解放するぞ！
- 2005年 留年 (2年生2回目)
  - 年度末に起業
- 2006年 留年 (2年生3回目)
- 2009年 システム情報工学研究科 入学
  - 研究漬け
- 2011年 進学できたらしいな…

- システム情報工学研究科 CS専攻  
- 自然言語処理 on the Web 研究室



ACCC Photo Archives, Univ. of Tsukuba  
<http://photo.cc.tsukuba.ac.jp/>

- 産学リエゾン共同研究センター  
– 客員研究員



ACCC Photo Archives, Univ. of Tsukuba  
<http://photo.cc.tsukuba.ac.jp/>



株式会社しずくらボ



小野永貴  
図書館情報メディア研究科

- 有限会社てっくてっく  
- 代表取締役



- サービス運営
- 受託開発
- コンサルティング

# Web検索エンジンの基礎

- 大量のウェブ
  - 2,600万 unique URLs (1998)
  - 1兆 unique URLs (2008.07)
- 拡張するウェブ
  - 企業サイト
  - eコマース
  - SNS (ブログ)
  - 画像, 動画

We knew the web was big...  
(Official Google Blog, 2008)



- 必要な情報を得るために
  - ウェブ検索エンジン
- ウェブ検索エンジンの利用率
  - 93.7% (日本 2008)

インターネット検索エンジンの現状と市場規模等に関する調査研究  
(総務省情報通信政策研究所, 2009)

- 主要なウェブ検索エンジン
  - Yahoo!
  - Google
  - Bing (マイクロソフト)



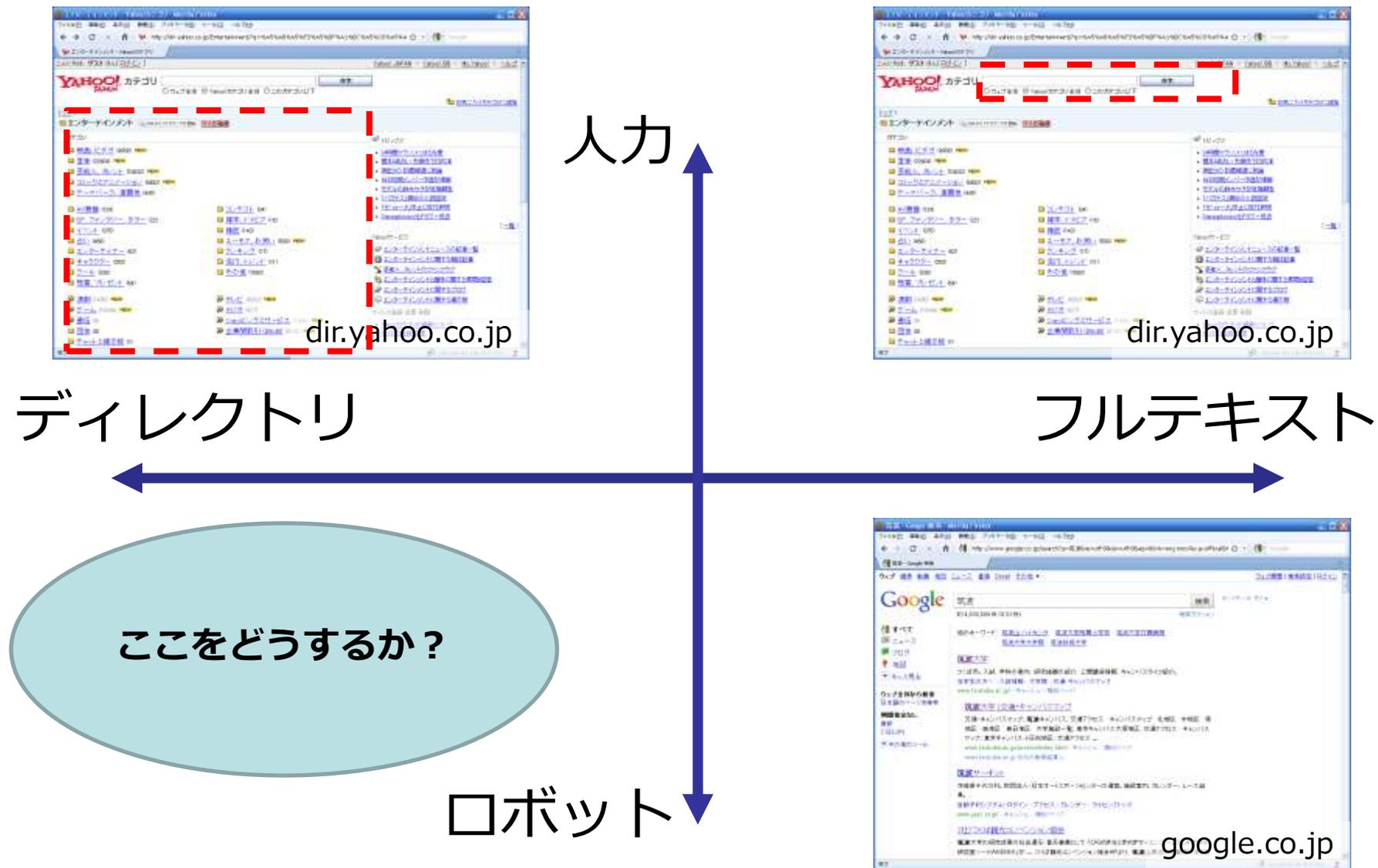
- ディレクトリ型 (Yahoo!)
  - サーファードが収集 (人力)
  - 信頼性の高いウェブサイト
  - 登録サイト数に限界



フルテキスト型

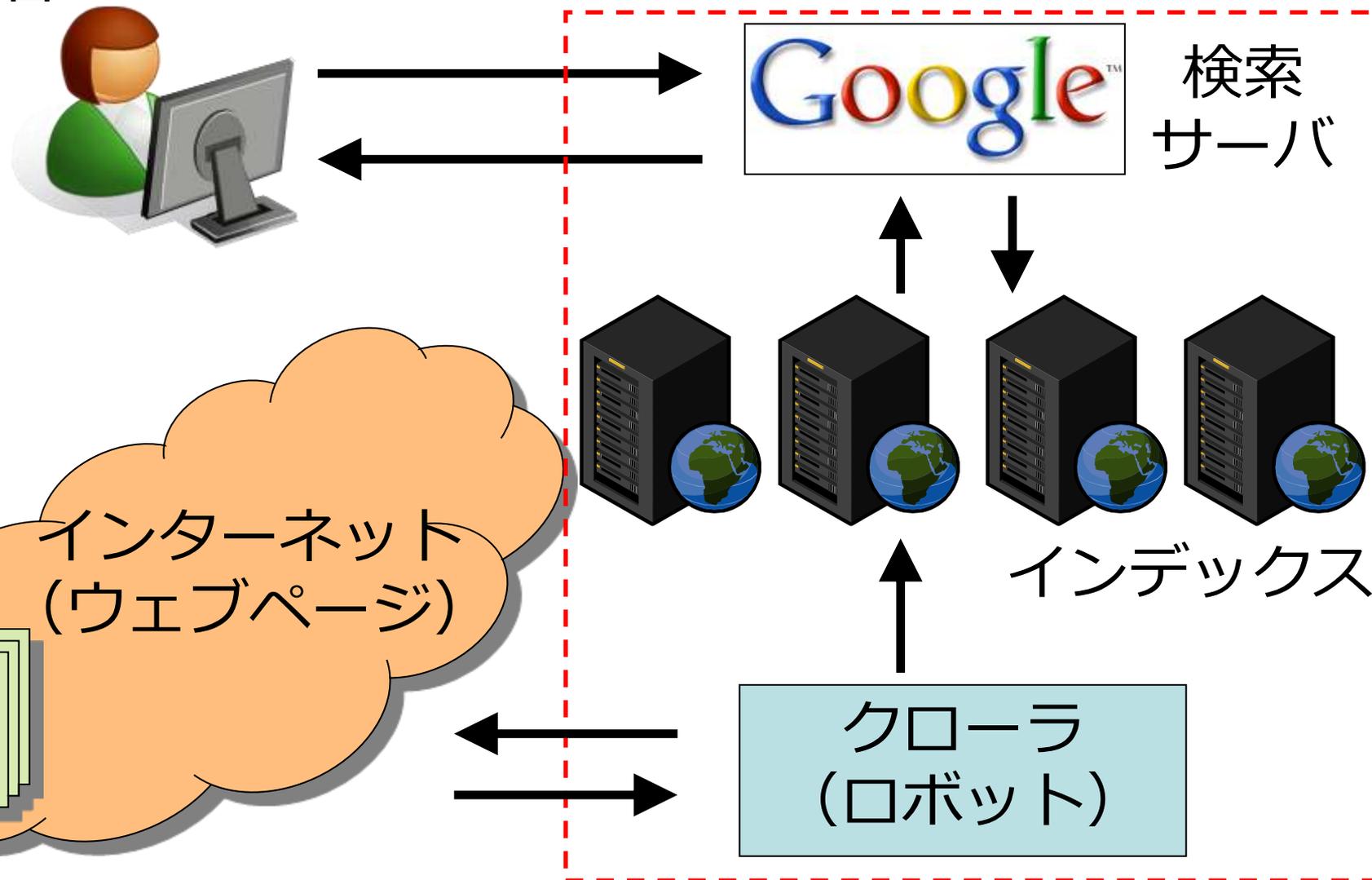
- ~~ロボット型~~ (Google)
  - クローラが収集 (自動)
  - 大量のウェブサイト
  - 質の低いサイトが混ざる





利用者

検索エンジン



# Web検索エンジンの課題 未来の検索エンジン

キーワードや求めているものがはっきり分かっていなくても、探し出す検索の仕組みを作りたい。それが本来の検索のはず。

(『日経産業新聞』 2010年5月18日 6面)

- 内容との乖離

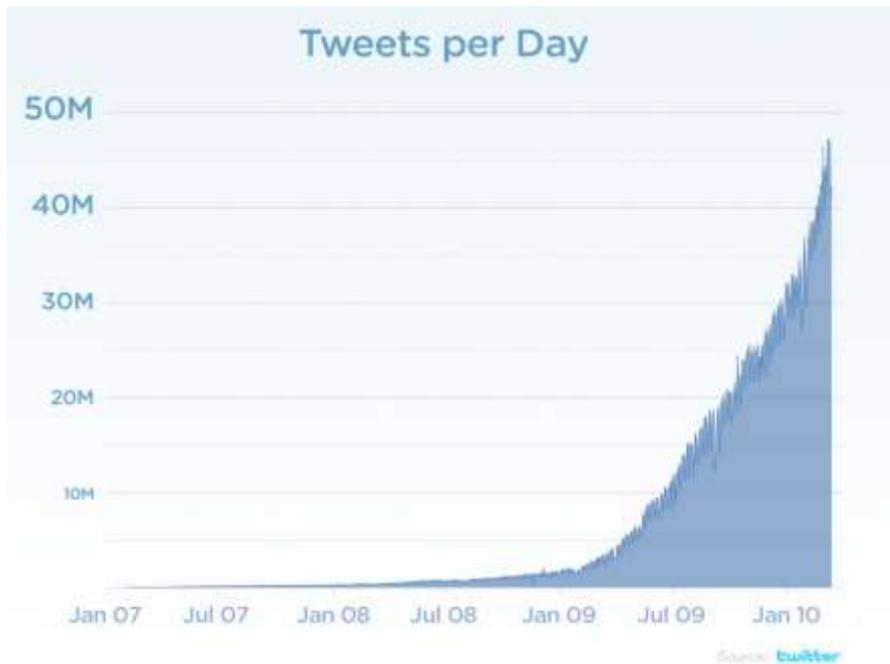
可愛い犬ですね！



- 概念で検索

- あるキーワードが含まれるウェブページ
- 「元気の出るウェブページを見たい」

- マイクロブログの普及  
- ショート・コンテンツ



ツイートの測定結果  
(Twitterブログ, 2010)



前後の文脈がわからない

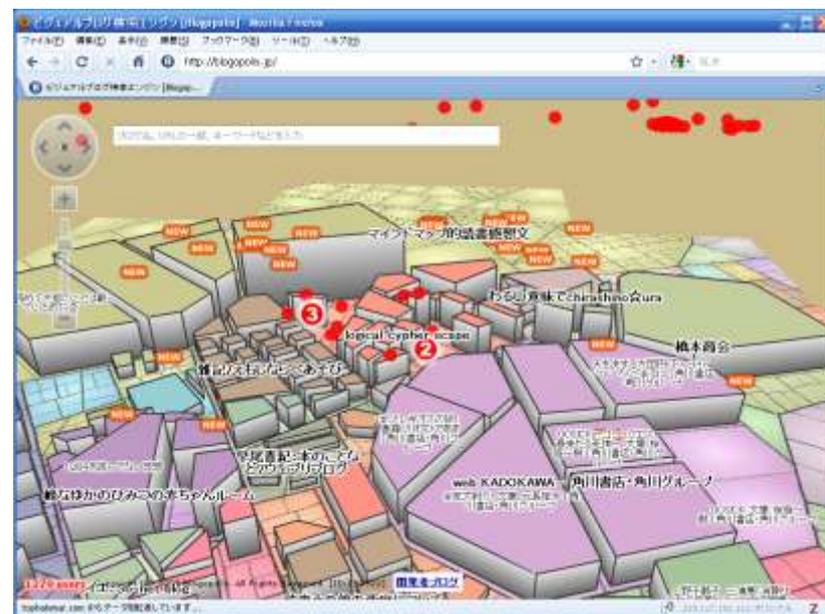
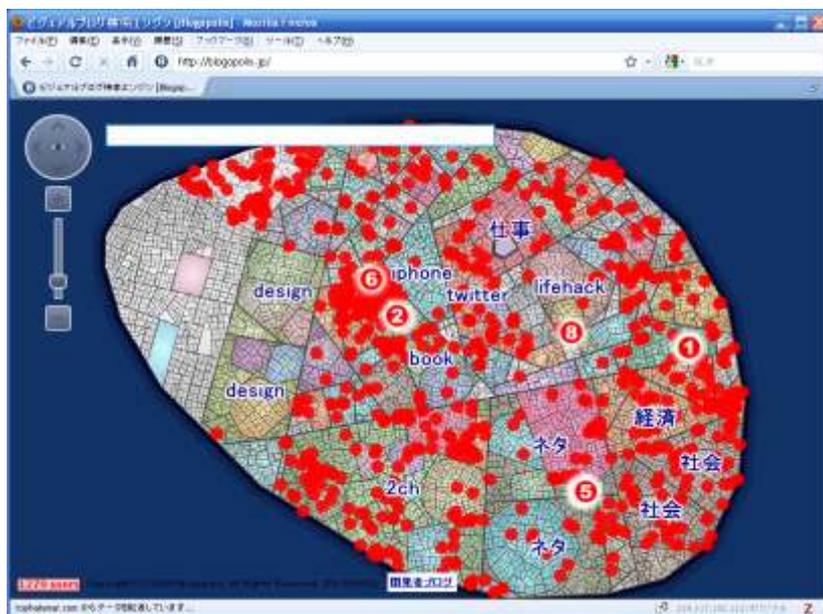
- 大量の検索結果を全部みる？
  - 平均 2.35 (米国 1999)

Amanda Spink, Jack L. Xu.

Selected results from a large study of Web searching: the Excite study.  
Information Research, Vol.6, No.1, 2000.

- ランキングアルゴリズム
  - ベクタスペースモデル, PageRank
- 新しいコンテンツ
  - ブログ, ニュース
  - ショッピング

- 電話帳メタファからの脱却
  - Blogopolis
  - <http://blogopolis.jp/>



# CEEK.JP

## 統合型メタサーチエンジン

「検索」の技術が躍進してこそ、人類の英知の利用が促され、より良い社会になる。

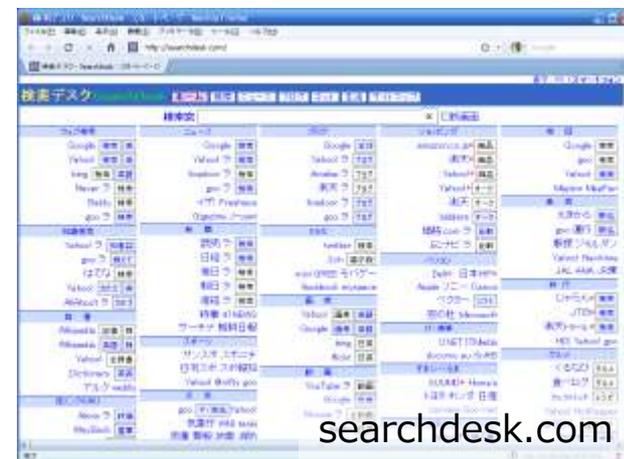
(CEEK.JP ヘルプページ)

- CEEK.JP
  - 統合型メタサーチエンジン
  - 2002年8月開始
- 当時のウェブ検索エンジン
  - 定番が無い状態
  - 複数の検索エンジンを使い分ける



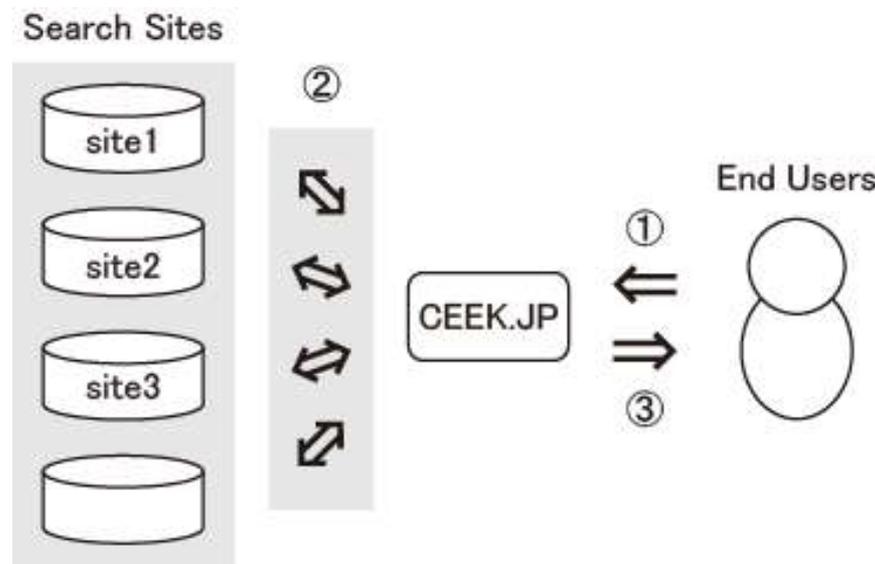
- 当時の検索方法

- 検索サイトAで検索
- 検索サイトBで検索
- 検索サイトCで検索
- 検索サイトDで検索 ...



- CEEK.JPのねらい

- 作業の自動化
- 検索サイトA~Dの結果を1ページにまとめる



1. ユーザがキーワードを入力
2. CEEK.JP が複数の検索エンジンで検索
3. 検索結果を統合して表示

# CEEK.JP NEWS

## ロボット型ニュース検索エンジン

- CEEK.JP NEWS

- ロボット型ニュース検索エンジン
- 2004年6月開始 (β版: 2003年11月)
- 日本語では最も歴史がある

- 初期 (2003年2月)

- CEEK.JPと同じ仕組みで

- ニュースソースの少なさ
- 反映の遅さ
- 権利問題



- 需要の予想

- 海外でBlogがブームに

- ニュース検索の需要が伸びると予想

- 狭義のBlog (Wikipedia)

- ウェブページのURLとともに覚え書きや論評などを加えログ (記録) しているウェブサイト的一种

- 当時の (僕の) 状況

- 日本語のロボット型ニュース検索が無かった

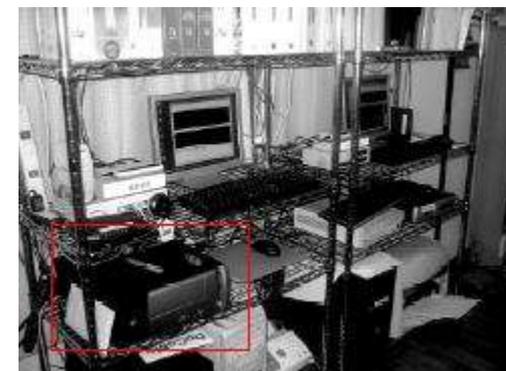
- ウェブ検索エンジンを作りたい

- サーバを準備しないと
  - 共有サーバでの運用は難しい  
(特にクローラ)
  - 専用サーバは高い
  - 自前でサーバを構築しよう！

「ママ、どうしておうちにサーバーがあるの？」  
マイクロソフト社が作成したパンフレットのタイトル

- メリット

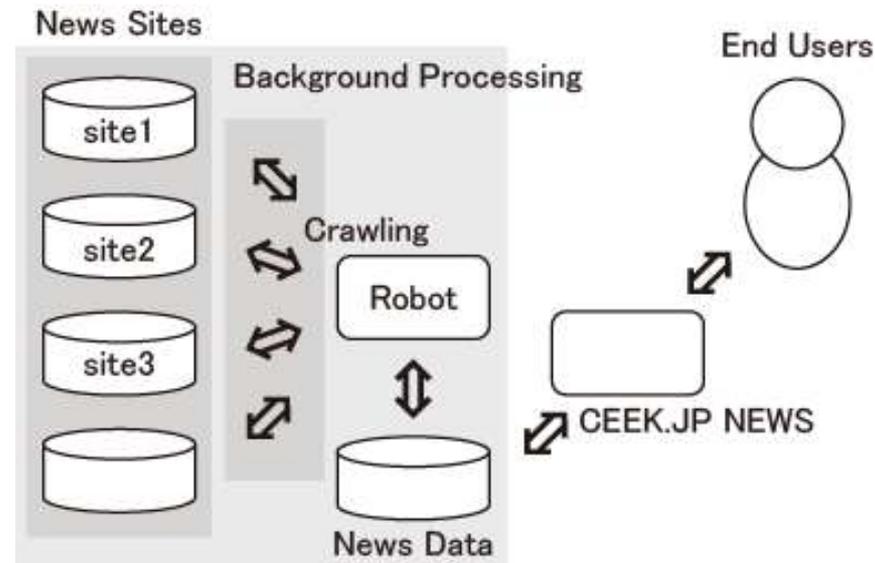
- 暖房要らない :)
- ハードウェアの自由度が高い  
(メモリ, SSD)



初期 (登宅)



現在 (自宅)



- 前述のロボット型検索エンジンと同様
  - リンクは1ページのみ辿る
  - サイトに応じた解析エンジン  
(タイトルや記事本文の抽出)

- サービス理念
  - システムは情報の取捨選択をしない  
(表示順序は日付順)
  - 機械化が進むと人間は馬鹿になる  
(と僕は信じている)
- 検索エンジンのパワー
  - 興味がわかる
  - 未来の予測

- 今後の展望

- 対象サイトの増強
- バックエンドシステムの変更  
(ファセット検索の実現)



- 有料検索サービスの提供
  - 解析結果の提供
  - API の提供
- 未来年表の自動作成

# 有限会社てっくてっく

大学に入ってまっ先に思ったのは、人材がもったいないということ。

（『日経クリック』 2007年2月8日 p.21）

- 有限会社てっくてっく
  - 2006年3月 設立
  - 変な会社名…
- 目的
  - 学生に適切な報酬を
  - サービス運用の法的リスク軽減
  - 新しい働き方を目指して (あとで)
- 有限会社
  - 決算の公開義務がない



- 学生に適切な報酬を
  - 学生の技術力は低いかな？
  - 受注金額の50%を報酬に  
(ただし、継続雇用を保障しない)
- 失敗…
  - 高い報酬を避ける (責任回避?)
  - 自己マネジメントを行わない
    - 安定雇用を望む

- サービス運用の法的リスク軽減
  - 個人: 無限責任 / 法人: 有限責任
  
- (旧) 著作権法
  - ウェブ検索エンジンの法的リスク
    - クローリング (複製権)
    - インデキシング (翻案権、同一性保持権)
    - 検索結果表示 (公衆送信権)
  - 黙示の許諾論
    - 回避手段を取らない場合は許諾したと見なす

- (改正) 著作権法 (2010年1月1日施行)
  - ウェブ検索エンジンの例外追加
  - 第47条の6、政令第299号、文科省令第38号
- ライントピックス事件  
(平成17 (ネ) 10049)
  - 読売新聞が記事見出しの無断複製により著作権を侵害されたと提訴
  - 知財高裁は記事見出しの著作権を認めなかったものの、不法行為を認め賠償命令

- ビジネスモデル

- 受託開発

- 運営サイトを見た会社が依頼
- ウェブサービス全般

- 保守業務

- 会社運営コスト（人件費等）の70%をカバー

- おかげさまで

- 初年度からずっと黒字

# キャリアを考える



一例です。自身のキャリアは自己責任です。

- PCに触る
  - 1995年 (小学5年)
- ソフトウェア検索エンジン (K-Search)
  - 2000年 (高校1年)
- 無料Webスペースの提供 (K-Server)
  - 2001年 (高校2年)
- CEEK.JP 開始
  - 2002年 (高校3年)

- フリーランスで活動開始
  - 2003年 (大学1年)
- アルバイトしてみる
  - 2004年 (大学2年)
- フリーランスで凄く儲かった
  - 2005年 (大学2年2回目)
  - 価格.com の類似サイト (提供終了)
- 会社創業 (2006年)
  - 池嶋俊 (現在はGoogle) と2人で

- 受託開発の方向性

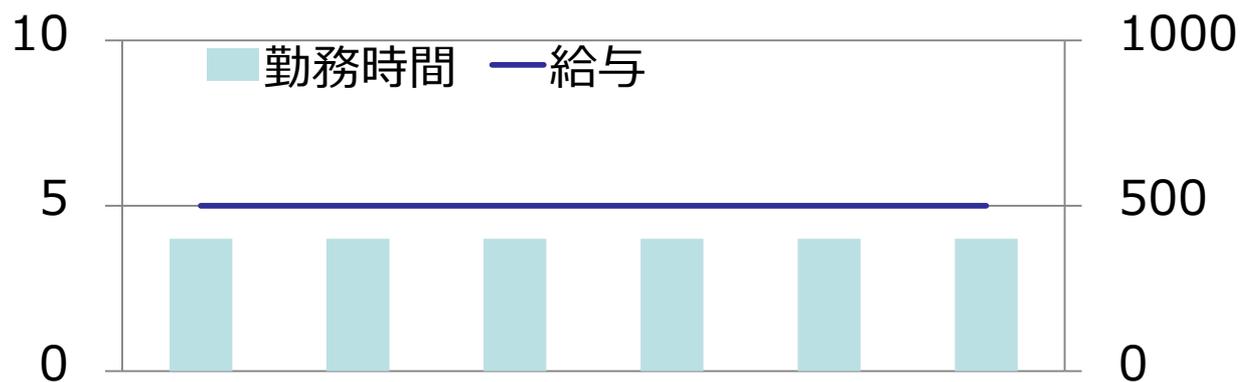
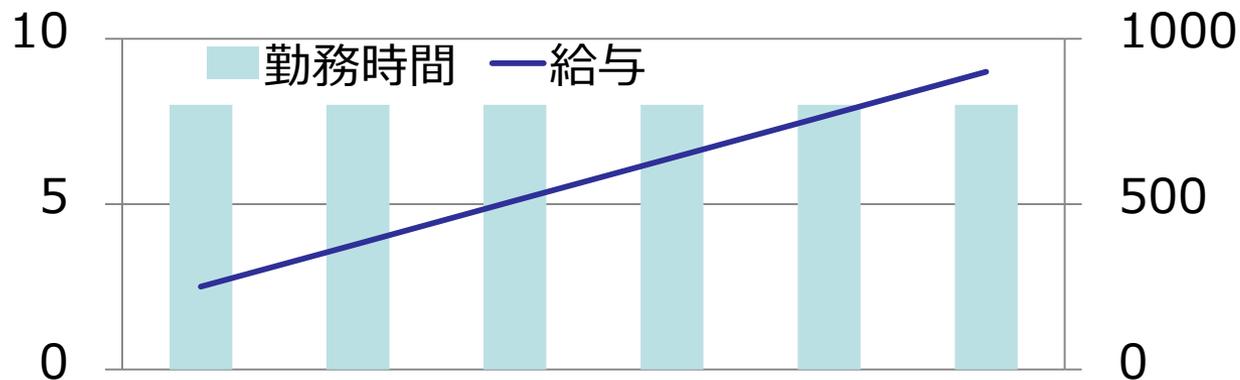
- ウェブとネイティブソフトのハイブリッド
- 両方出来る会社はそんなに無い
  - 吉田: ウェブ
  - 池嶋: ネイティブソフト



ONPOO.net

柔軟な視聴設定ができ、P2Pで音楽を配信するサービス

- どのように仕事をしたいか



- 通常の給与モデル
  - 勤務時間は一定で給与が上がる
  - 勤務時間に選択肢がほぼ無い
- 目指したいモデル
  - 勤務時間を短くしてほどほどの給与
  - 自由な時間を増やす
    - お金を稼ぐために使う時間を減らす
- 起業するしかなさそうですね…

- 5年で〇億貯める！
  - 5年後には隠居生活
  - 身体壊しそう…
  - (かつてのベンチャー企業)
- ほどよくほどほど
  - 保守業務
    - 年間保守料の目安: 受託金額の30%
    - 品質の高いソフトウェアを納品しましょう
  - ウェブサービスの提供
    - 課金、広告収入

- お金はあまり使わない
  - 食事
  - 図書
  - サーバ



- 大学院生なので研究します
- 研究と仕事は両輪でありたい
  - 仕事上の課題を研究で解決する
  - 学术界との課題にずれが出るが…
    - 工学のゴールはビジネスなのでは？
- Webコンテンツ抽出
  - ニュース検索の運営コスト軽減

YOMIURI ONLINE 読売新聞

宇宙の夢をコミュニケーション。  
松本零士さんが若田さんとブログ対談

政治

「麻生政権がけつぐち」石原幹事長代理が強い危機感

自民党の石原幹事長代理は8日午後、支持者を集めて前向のホテルで開いたセミナーで講演。相次ぐ失言などで麻生首相の求心力が低下していることについて、「国民の心奪取り回。民主党に政権運営をやらせてみてくれ」とも述べている。自民党の国会議員の7〜8割は、麻生政権で（麻生）選挙をやって与党で、られるか疑問を持っている状態だ。政治的にも経済的にもけつぐちにあると述べ、強い危機感を示した。

次期総選挙に関しては、「選挙戦は11月までの速になるのでは？」と、こんなことにはなかった」と口を開いた。自民党役員が「首相総辞職」に言及するのは異例。

石原氏の発言について首相は8日朝、記者団に内容を聞いて、「い」から分かることだけ言った。

2008年12月05日 23時32分 読売新聞

YOMIURI ONLINE トップへ

月々3,000円からの国際協力  
小さなことから未来につなぐ「プランジャパン」支援  
書集集中  
www.plan-japan.org

政治 経歴記事

- 小沢代表、選挙管理内閣の首相「失態しきれない」(12月2日 00:18) 読売 選挙・麻生政権
- 岐阜市長が辞職、市電乗車の私立移管問題で民意問(12月1日 21:54)
- 「うしろから鉄砲玉撃つな」吉賀選対委員長が麻生批判一喝(12月1日 21:06) 読売 選挙・麻生政権
- 麻生さん「アヤマ」野党のヤジに肉向て反撃(12月1日 20:09) 読売 選挙・麻生政権
- 中川元幹事長呼びかけの議員会合に57人、政界再編の布石(12月1日 20:03) 読売 選挙・麻生政権
- 麻生さん向う、失言しないコツ「かつ」～伊吹元幹事長(12月1日 19:44) 読売 選挙・麻生政権
- たばこ税増上げ見送り、与党で反対論強まる(12月1日 19:00) 読売 選挙・麻生政権
- 原日教組の自民有志議員、議員連立～中山氏も賛同(12月1日 21:12)
- 雇用・能力開発機構廃止へ、行革相出陣野郎が合意(12月1日 23:00)
- たばこ増税で与党内の調整難航、首相は慎重姿勢に(12月1日 21:02) 読売 選挙・麻生政権

編集長のおすすめ

選挙中町…見合いは遅くはない? 選挙中町作局…選挙中町-菅正人

PERSONALIZE SI.COM WITH YOUR FAVORITE PRO AND COLLEGE TEAMS. IT'S FAST AND FREE!

SI.COM MLB

EXTRA MUSTARD FANATIION SI BUILT FANTASY DAN PATRICE SWHSUIT SI PHOTOS SI KIDS VIDEO TAKEZIE

INTRODUCING THE ALL-NEW TL

Unconventional Wisdom: Busting the myth of the salary cap

Small-market teams love salary caps. Or rather, they think they do. At least on paper, cap stop teams in New York, Boston and Chicago from oligopolizing the free-agent market, and should therefore help level the economic playing field. And, to a certain extent, they do; a small-market team in a capped league is more likely to acquire or retain top-tier talent, but there's a catch. That same small-market team will need to win, and keep winning, just to stay financially viable. And sometimes, winning might not even be enough.

Let's say, in some far-off universe, MLB owners and players actually did agree on a salary cap. With it would come the normal provisions: a salary floor at around 75-85 percent of the cap, and a guaranteed percentage of total industry revenues for the players. Since the players have been taking in about 45 percent of revenues the past few years, we'll keep it at that figure (the other three major sports leagues, which are all capped, each pay out over 50 percent).

Using 2008 as an example, the 30 teams took in about \$6 billion (not including MLB Advanced Media revenue), for an average of \$200 million per team. Forty-five percent of that (the players' share) is \$90 million, which we'll use as the midpoint between our floor and cap. If we want to make the floor 75 percent of the cap (a low-end figure, relative to the other leagues), we can use \$77 million and \$103 million, respectively.

With a \$103 million cap, nine teams would have been affected last year, and a total of about \$286 million would have had to be sliced off the top. Since total salaries have to remain at existing levels, the bottom 21 teams would have had to take on the burden, which had previously been placed on the Yankees, Red Sox, et al. On the other end, 14 teams would have been under the payroll floor, by a total of \$251 million. Even discounting the Marlins' \$22 million payroll, the other 13 teams would have had to spend an average of \$13 million more just to meet the minimum. Some of these teams might be able to afford it, most wouldn't.

Nobody likes seeing the Yankees sign big-name free agents every single offseason, but is it actually good for baseball as a whole?

More MLB

Latest MLB News

- Can Atlanta contend in the East with its new... and old... rotation?
- Bumens: Time for a salary cap?
- Rodonski at court house of Clemens grand jury: Mets might not permit Santana to pitch in WBC
- Piggins, Angels avoid arbitration with \$5.8M deal

MLB Truth & Rumors

- 2009年度の対外発表
  - 論文誌1件、口頭5件、ポスター4件
  - 燃え尽きた…
- やりたい事と違うような…
  - 研究のための研究に
  - 本当にやりたい事は？
    - ユーザにプロダクトを使って貰いたい！

- 手軽に開発・手軽に利用
  - ネイティブアプリにはダウンロードの障壁
- 上流から下流まで
  - アイデア、設計、開発、宣伝、運用 …
  - 手応えを感じる事ができる
- フィードバック得やすい
  - アクセス解析

- 難しいのは「継続」すること
  - サービスの開発は比較的簡単
  - 継続する事が難しい
  
- 継続を容易にする2点
  - 自身が使う事
  - メンテナンス・フリーである事

- 継続する事は自身の名を売る事
  - 「静岡大の○○です」以外は言えますか？
  - 続けていれば誰かの記憶にひっかかる
  - 継続は信用の基
- その結果
  - お仕事が来たり
  - インタビューされたり
  - 講演依頼が来たり
  - 自信がついたり

- 在学中に起業するのがお勧め
  - 卒業後も継続できそう
    - 続ける
  - 失敗した！
    - 就職
- 経験の落とし穴
  - 社長の経験は社長でしか積めない  
(副社長と社長の溝)
  - 経験・知識にゴールはない
- IT起業であればお手軽

- 何かサービスを作りましょう
  - 紹介するものが無ければ営業も難しい
  - 受動的な営業を目指す
- 自身の能力を知りましょう
  - 予定時間と作業時間の記録
- 安売りは止めましょう
  - 一度下げた受託金額を上げるのは困難
- 売り切りにならないように
  - 保守契約、再利用

おわりに

- 学生の特権
  - 失敗した際のリカバリが容易
  - 自身のブランドを持ってみませんか？
- 便利な世の中
  - IT起業であれば準備いらず
  - ウェブサービスで多人数にリーチ
- 一歩踏み出した先に
  - ワクワクする未来があるはず

**とりあえず一歩前に！**

# 「吉田光男」で検索

## ceekz@mibel.cs.tsukuba.ac.jp

何かありましたらお気軽に。

@ceekz