Web検索エンジンと共に

吉田光男
システム情報工学研究科 CS専攻
（有）てっくてっく / CEEK.JP

創成学類 フレッシュマンセミナー

2010/6/9

アウトライン

Ceek.jp

- 自己紹介
- 基礎知識
  - Web検索エンジンの種類
  - ロボット型検索エンジン
- 事例
  - CEEK.JP
  - CEEK.JP NEWS
  - （有）てっくてっく
- 未来の検索エンジン
- アドバイスらしき何か

自己紹介

Ceek.jp

- 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（山本研）
  - 産学リエゾン共同研究センター 客員研究員
  - （有）てっくてっく
- 1984年生（25歳）
- 和歌山県出身

経歴

Ceek.jp

- 2003年 情報学類 入学
  - AC 入試
    - *図書館の情報をインターネットに解放する*
- 2005年 留年（2年生2回目）
- 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
- 2011年 進学できたらいいな…

所属(1/3)

- 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(山本研)

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

所属(2/3)

Ceek.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/

Shizuku Lab.

株式会社しずくラボ

小野永貴
図書館情報メディア研究科

所属(3/3)

Ceek.jp

- (有)てっくてっく

Web検索エンジン種類

総合科目「マルチメディアの舞台裏」を受講している方は居ますか？
（内容が重なってる）

拡大するウェブ

Ceek.jp

- 大量のウェブ
  - 26,000,000 unique URLs (1998)
  - 1,000,000,000,000 unique URLs (2008.07)

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

- 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

tsukuba.ac.jp


amazon.co.jp


twitter.com


youtube.com

ウェブの道しるべ

Ceek.jp

- 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）

海外では逆

yahoo.co.jp

google.co.jp

bing.com

検索エンジンの分類

Ceek.jp

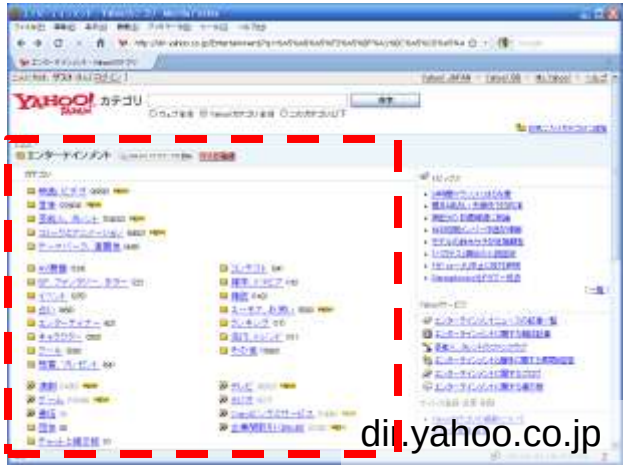
- ディレクトリ型（Yahoo!）
  - サーファーが収集（人力）
  - 信頼性の高いウェブサイト
  - 登録サイト数に限界

フルテキスト型

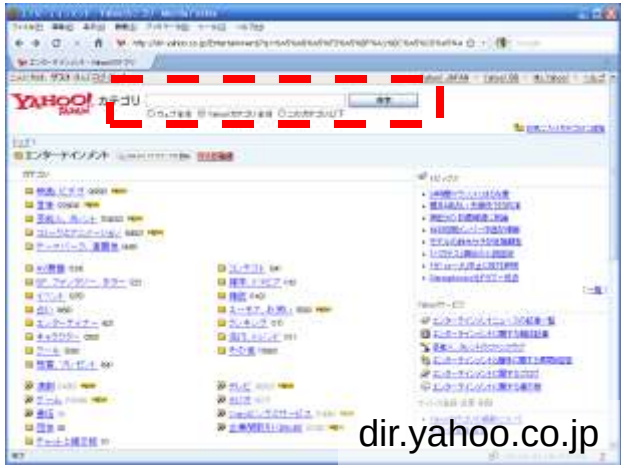
- ロボット型（Google）
  - クローラが収集（自動）
  - 大量のウェブサイト
  - 質の低いウェブサイトが混ざる

本当の検索エンジンの分類

Ceek.jp



人力



ディレクトリ

フルテキスト

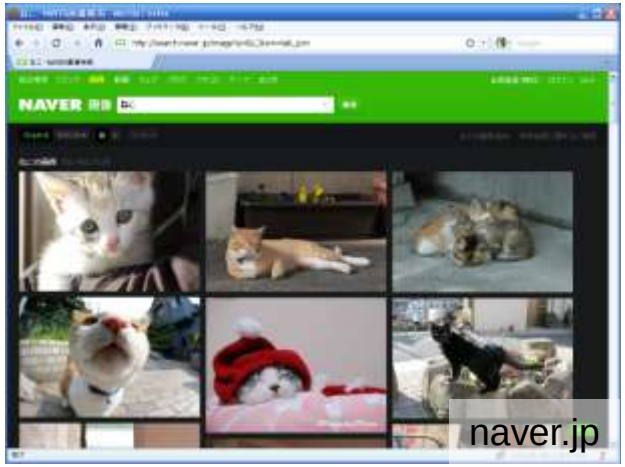
ここをどうするか？

ロボット

検索エンジンの種類

Ceek.jp

- 拡張するウェブ
  - 様々な種類のコンテンツ
- 何を対象とするか
  - 一般的なウェブページ
  - 画像, 動画
  - ブログ
  - ニュース
  - 論文



naver.jp

news.ceek.jp

ロボット型検索エンジン

ロボット型検索エンジンの仕組み

キーワードを入力して

Ceek.jp

「検索」ボタンを押すと

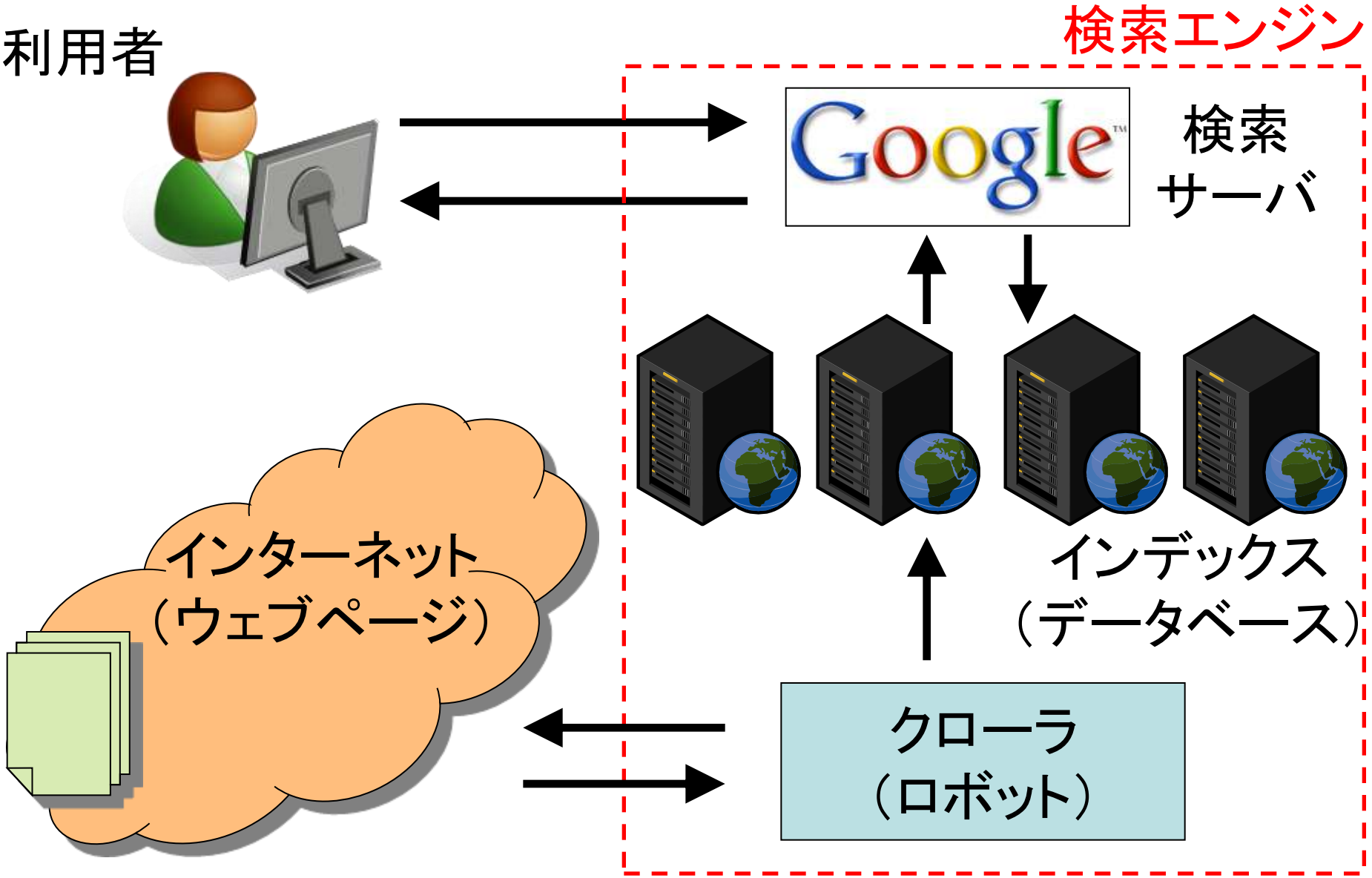
Ceek.jp

必要な情報にたどり着けた！

Ceek.jp

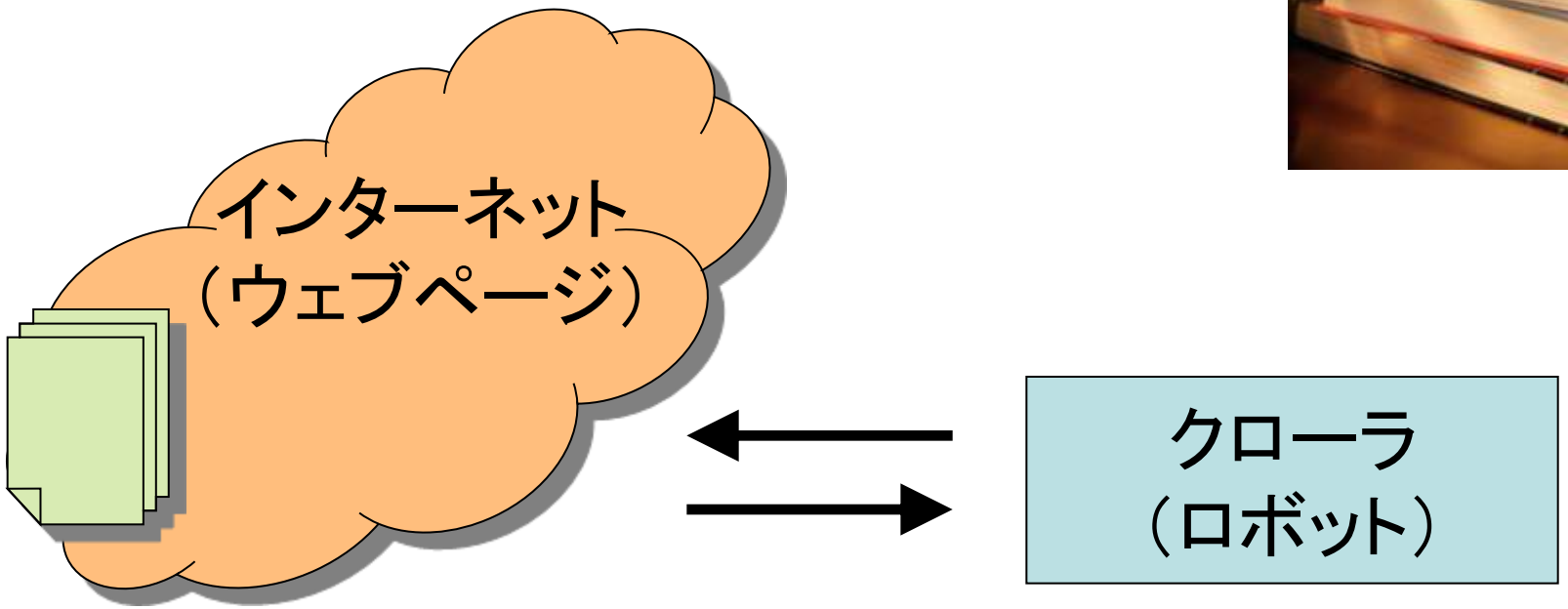
ロボット型検索エンジンの全体像

Ceek.jp



- リンクを辿りながらウェブページの収集
  - 未巡回のページを収集
  - リンク解析

繰り返し



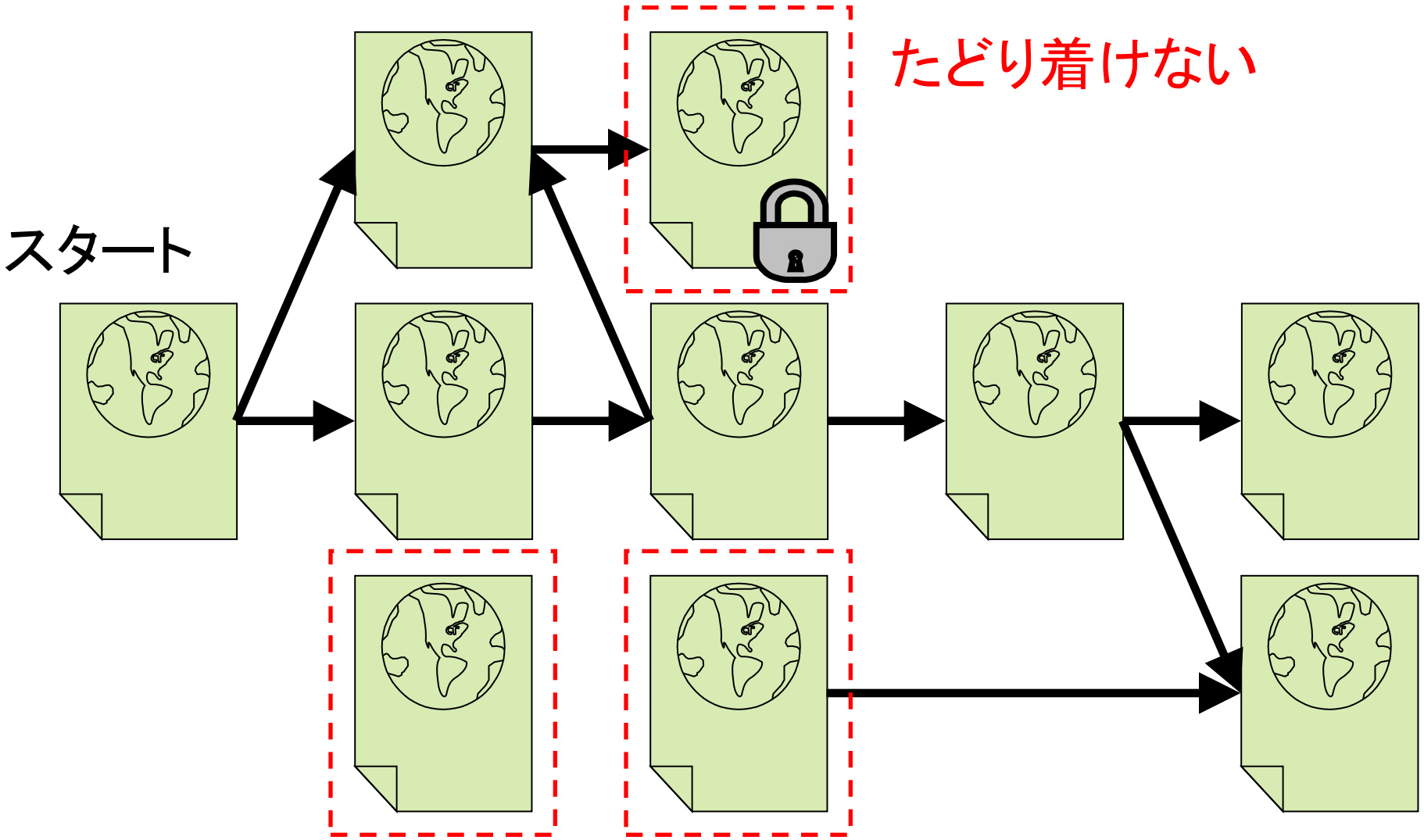
リンク

Ceek.jp

tsukuba.ac.jp

リンクの巡回

Ceek.jp



- インデックス
  - 対象のデータにアクセスする「目次」
  - ページに含まれる単語とそのページを記録

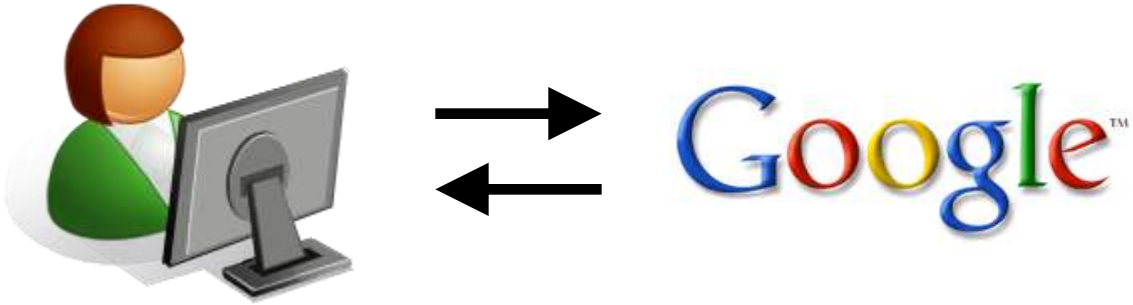
インデックス
（データベース）

可愛い犬の写真

検索サーバ

- データベースから検索結果を取得

- 検索結果を見やすく表示
  - 広告の表示
  - ランキング（表示順序）



広告の表示

Ceek.jp

google.co.jp

ランキング（表示順序）

Ceek.jp

- 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.

- キーワードとの類似性
  - TF-IDF

～2000年
スパムまみれになる

- ページの重要度
  - PageRank
  - HITS

2000年～
ページそのものを評価する試み

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（1/3）

Ceek.jp

- CEEK.JP
  - 統合型メタサーチエンジン
  - 2002年8月開始

google.co.jp

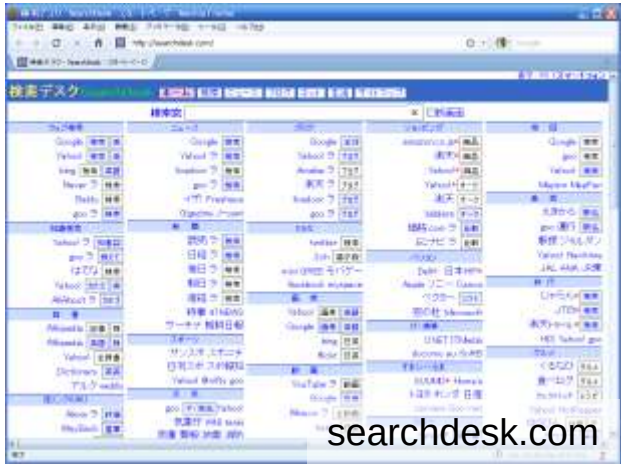
- 当時のウェブ検索エンジン
  - 定番が無い状態
  - 複数の検索エンジンを使い分ける

僕: …というサービスを考えているのですが、AC入試に効きますか？
某: 面白い！合格できると思います。
（2002年7月 某所）

CEEK.JP（2/3）

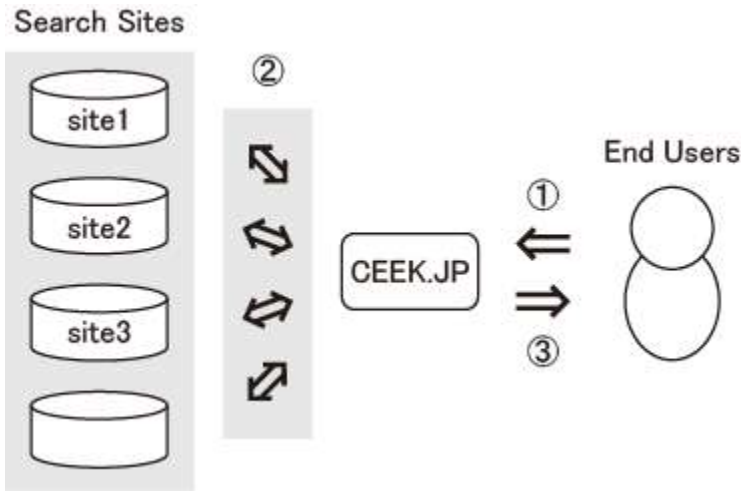
Ceek.jp

- 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

CEEK.JP（3/3）



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（1/6）

Ceek.jp

- 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

CEEK.JP NEWS（2/6）

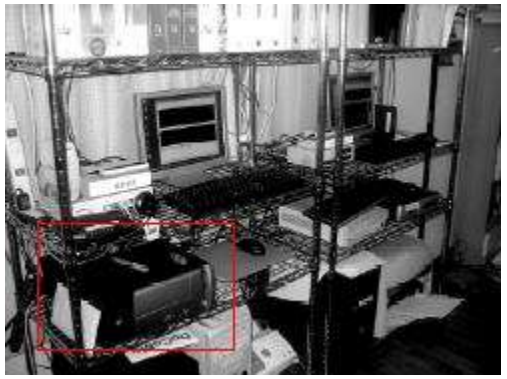
Ceek.jp

- 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の一種*
- 当時の（僕の）状況
  - 日本語のロボット型ニュース検索が存在しない
  - 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（クローラを動かす事が出来ない）
  - 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

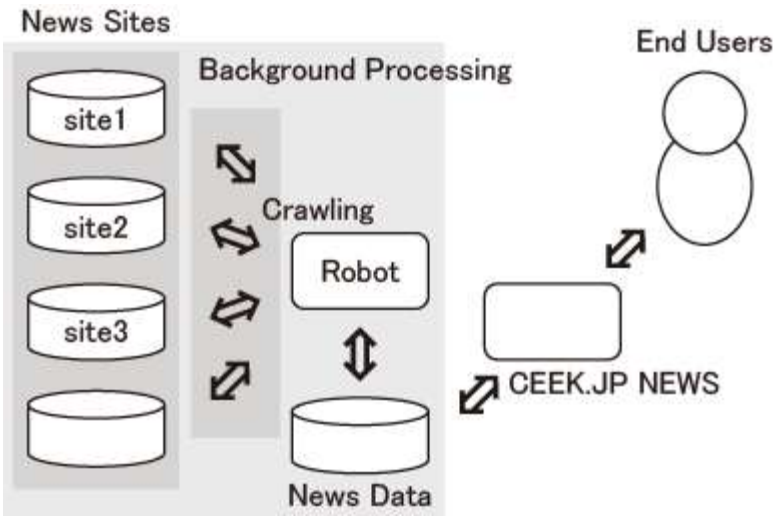
- メリット
  - 暖房要らない :)
  - ハードウェアの自由度が高い（メモリ, SSD）



初期（登宅）

現在（自宅）

CEEK.JP NEWS（4/6）



- 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）

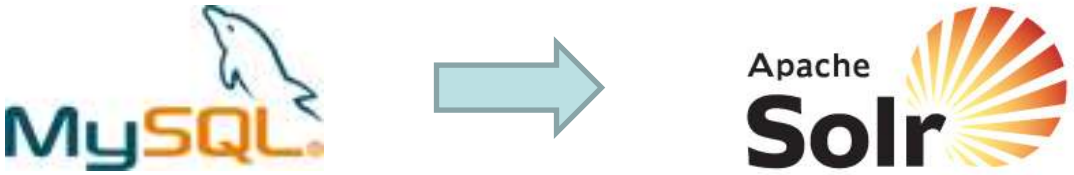
CEEK.JP NEWS（5/6）

Ceek.jp

- 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（バイトを募集するかも）
  - 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）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。
（インタビュー記事より,『日経クリック』2007年2月8日 p.21）

てっくてっく(1/5)

Ceek.jp

- (有)てっくてっく
  - 2006年3月 設立
- 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
- 有限会社？
  - 決算の公開義務がない

てっくてっく(2/5)

Ceek.jp

- 学生に適切な報酬を
  - 学生の技術力は低いか？
  - 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…
  - 日本人学生は高い報酬を避ける(責任回避？)
  - 外国人は沢山来る

- 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

てっくてっく(4/5)

Ceek.jp

- (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

てっくてっく(5/5)

Ceek.jp

- 弊社のビジネスモデル
  - 運営サイトを見てシステム発注依頼
  - 保守契約等で細々と…
- 今後のビジネス
  - CEEK.JP NEWS 有償サービス
  - 検索エンジンを作る

未来の検索エンジン
ウェブ検索エンジンの課題

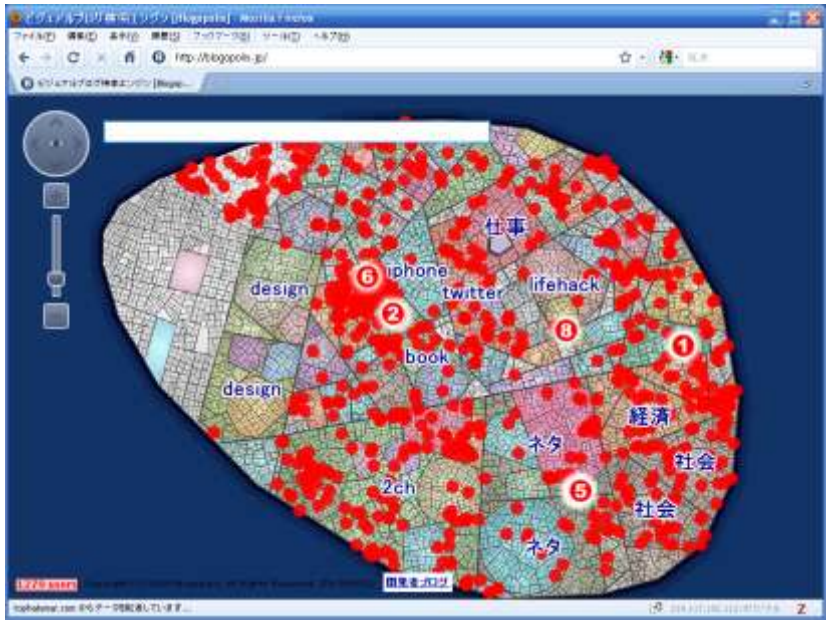
キーワードや求めているものがはっきり分かっていなくても、探し出す検索の仕組みを作りたい。それが本来の検索のはず。

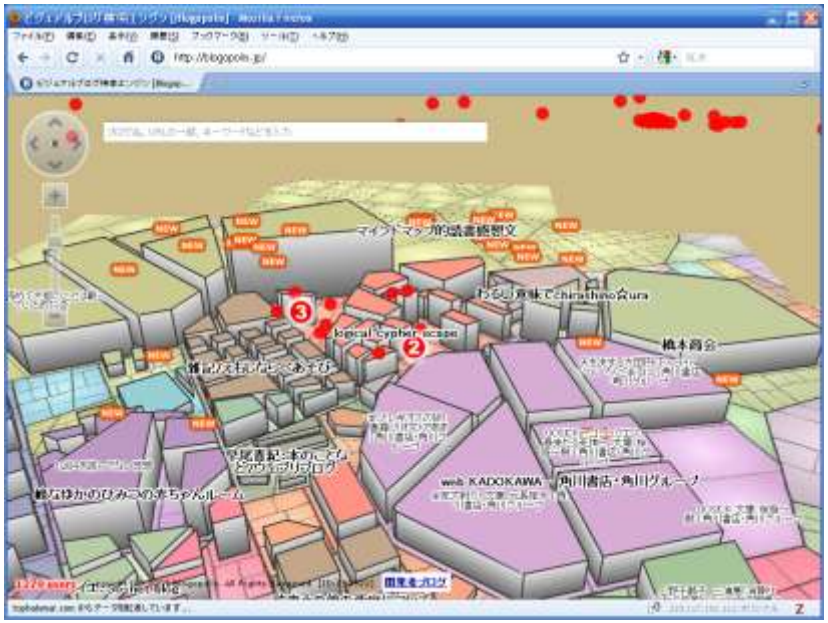
（インタビュー記事より,『日経産業新聞』2010年5月18日 6面）

電話帳メタファからの脱却

Ceek.jp

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





ロボット + 人

Ceek.jp

- ロボット + 人
  - NAVER（ネイバー）
  - http://www.naver.jp/

検索エンジンの課題

Ceek.jp

- キーワード検索からの脱却
  - 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」
- ページ単位からの脱却
  - ショート・コンテンツ
  - コンテンツの分割
- 単一結果からの脱却
  - 個々に応じた結果を
  - 対話

前後の文脈がわからない

アドバイスらしき何か

なんたらかんたら(1/2)

Ceek.jp

- ○○が学べない
  - 学べます
  - ただし、あなたが動けばの話ですが
- ブランドの確立
  - 「筑波大学の○○です」以外を言えますか？
  - 学内に留まっていませんか？
    - 最初から可能性を狭める必要は無い

なんたらかんたら(2/2)

Ceek.jp

- 会社作りたいです
  - どうぞ
  - 無責任に背中を押します
  - 「いいね」「面白いね」しか言わないよ
- 起業は在学中にするのがお勧め
  - 卒業後も継続できそう → 続ける
  - 失敗した！ → 就職

サービスを開発していて

Ceek.jp

- 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

おわりに

おわりに

Ceek.jp

- 先を見すぎて足下を見失わないように
  - 来るかわからない未来よりも今を楽しむ
- 天井を作らないように
  - 夢は諦めた瞬間に終わる
- 気楽に
  - 時にはどうしようもない事があります

ceekz@mibel.cs.tsukuba.ac.jp
「吉田光男」で検索

何かありましたらお気軽に。

@ceekz