

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)てくてく / CEEK.JP

創成学類 フレッシュマンセミナー

2010/6/9

- 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

- 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

- 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(山本研)



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/>



小野永貴
図書館情報メディア研究科

- (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)

- 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



- 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



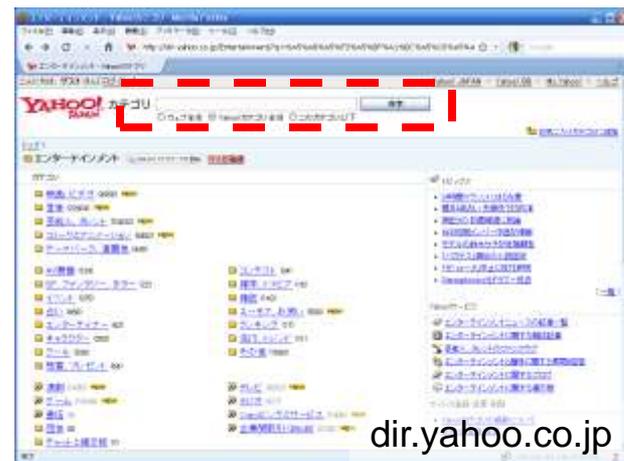
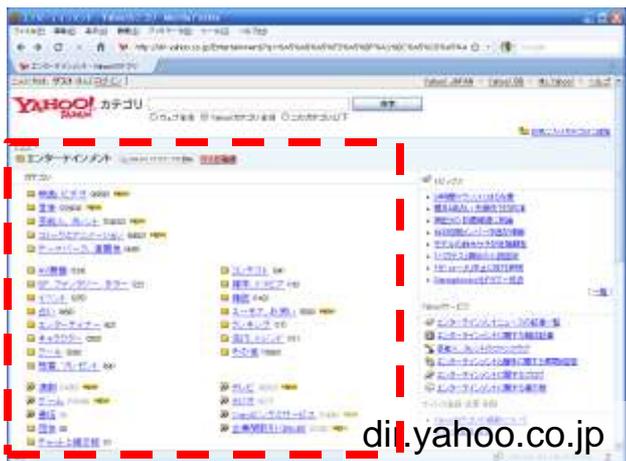
- ディレクトリ型 (Yahoo!)
 - サーファーマンが収集 (人力)
 - 信頼性の高いウェブサイト
 - 登録サイト数に限界



- ~~ロケット~~フルテキスト型 (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



↑
人力

←
ディレクトリ

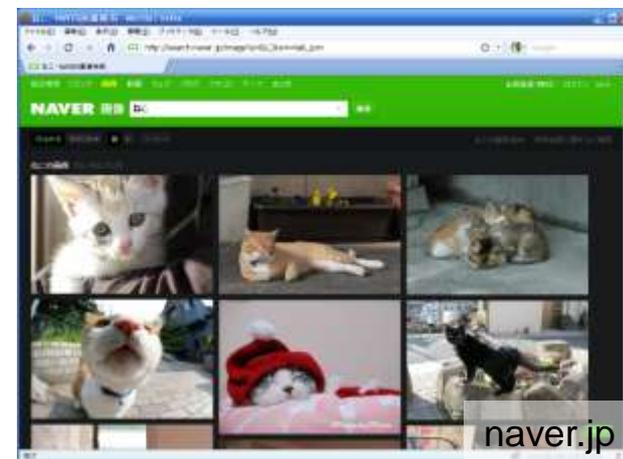
→
フルテキスト

ここをどうするか？

↓
ロボット



- 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



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



「検索」ボタンを押すと

筑波 - Google 検索 - Mozilla Firefox

ファイル(E) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)

http://www.google.co.jp/search?hl=ja&source=hp&q=筑波&btnG=Google+検索&aq=f&aqi=g1C

筑波 - Google 検索

ウェブ 画像 動画 地図 ニュース 書籍 Gmail その他

ウェブ履歴 | 検索設定 | ログイン

Google

筑波 検索

セーフサーチ: オフ

約 4,930,000 件 (0.35 秒) 検索オプション

他のキーワード: [筑波山ハイキング](#) [筑波大学附属小学校](#) [筑波大学付属病院](#)
[筑波大学大学院](#) [筑波技術大学](#)

筑波大学

つくば市。入試、学科の案内、研究組織の紹介、公開講座情報、キャンパスライフ紹介。
在学生の方へ - [入試情報](#) - [大学院](#) - [交通](#) - [キャンパスマップ](#)
[www.tsukuba.ac.jp/](#) - [キャッシュ](#) - [類似ページ](#)

筑波大学 | 交通・キャンパスマップ

交通・キャンパスマップ; [筑波キャンパス](#); [交通アクセス](#)・[キャンパスマップ](#)・[北地区](#)・[中地区](#)・[南地区](#)・[西地区](#)・[春日地区](#)・[大学施設一覧](#); [東京キャンパス大塚地区](#); [交通アクセス](#)・[キャンパスマップ](#); [東京キャンパス小日向地区](#); [交通アクセス](#) ...
[www.tsukuba.ac.jp/access/index.html](#) - [キャッシュ](#) - [類似ページ](#)
[www.tsukuba.ac.jp](#) からの検索結果 >

筑波サーキット

茨城県千代川村。財団法人・日本オートスポーツセンターの運営。施設案内、カレンダー、レース結果。
[自動予約システム](#)・[ログイン](#) - [アクセス](#) - [カレンダー](#) - [ライセンスストップ](#)
[www.jasc.or.jp/](#) - [キャッシュ](#) - [類似ページ](#)

(社)つくば観光コンベンション協会

筑波大学の研究成果の社会還元・普及事業として「ひらめき☆ときめきサイエンス～ようこそ大学の研究室へ～KAKENHI」が ... つくば観光コンベンション協会HPより、[筑波山周辺の駐車場状況を携](#)

完了 google.co.jp 66,245

The screenshot shows the University of Tsukuba website in a Mozilla Firefox browser window. The browser's address bar displays the URL <http://www.tsukuba.ac.jp/>. The website header features the University of Tsukuba logo and name in both Japanese (筑波大学) and English (University of Tsukuba). Navigation links include HOME, Site Map, Contact Us, and Campus Map. A main navigation bar lists various university sections: 大学案内, 学群・大学院・学内組織, 入試情報, 教育・学生生活, 研究・産学連携, 社会貢献・生涯学習, and 国際交流・留学.

The main content area features a large banner with the text "IMAGINE THE FUTURE. 開かれた未来へ。" (Imagine the Future. Open future). To the right of the banner is a purple box announcing the "Shinoda Goro 150th Anniversary International Symposium" (嘉納治五郎生誕150周年記念国際シンポジウム) held on June 12, 2010.

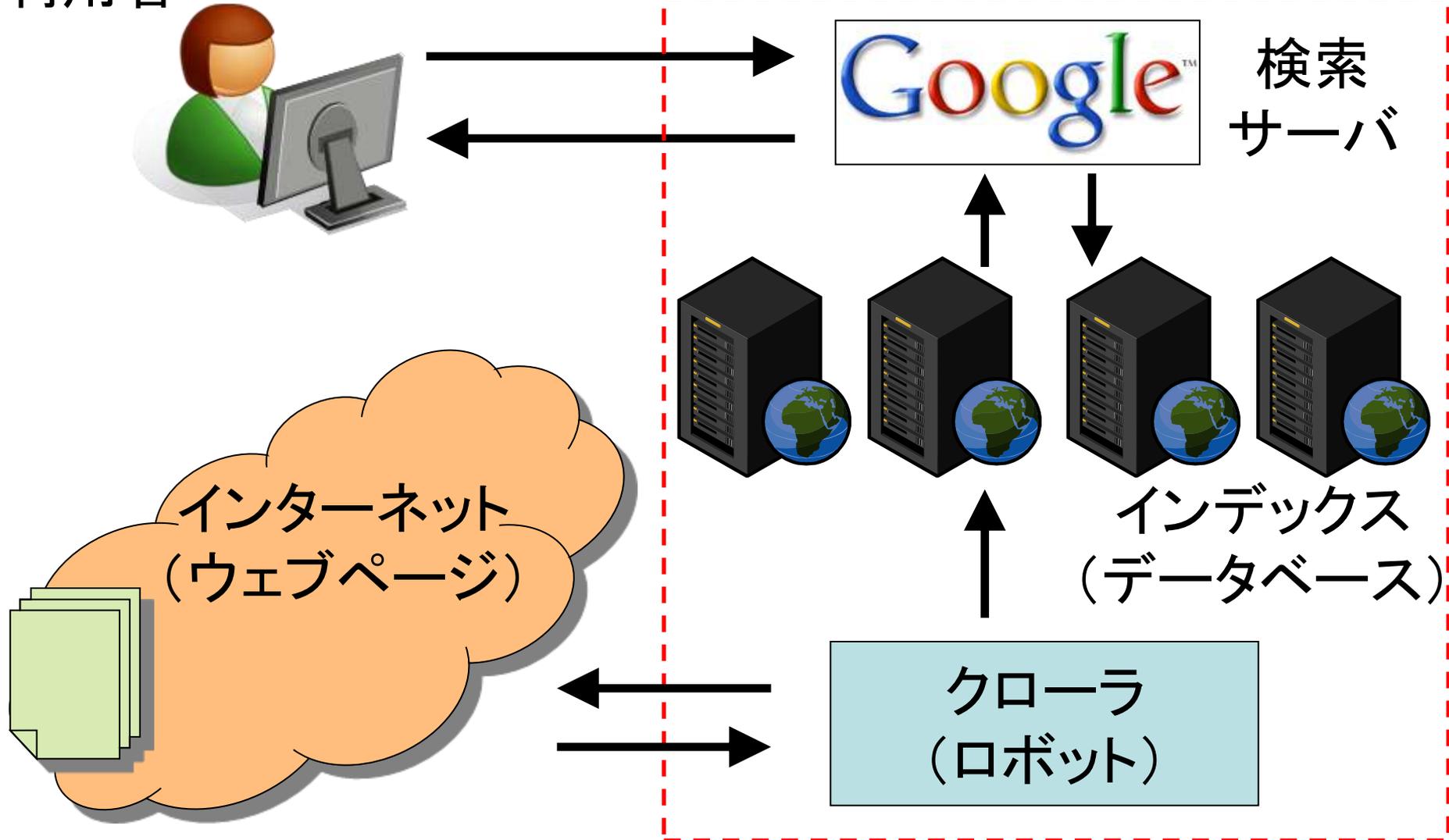
Below the banner are several sections:

- 最新情報 (Latest News):** A list of recent events and announcements, including a meeting on biodiversity, a molecular biology research center opening, a symposium on naturalists, and a campus relocation.
- イベント情報 (Event Information):** A list of upcoming events, including the 150th anniversary symposium and a bio-cafe.
- Left Sidebar:** Contains navigation links for prospective students, current students, faculty, and graduates, as well as a link to the faculty page.
- Right Sidebar:** Features the "TSUKUBA FUTURESHIP" logo, a list of projects like "TSUKUBA BRANDING PROJECT" and "PFI business initiatives", and a link to the "Women's Participation Promotion Room".

The browser's status bar at the bottom shows the page is "完了" (Completed) and the URL [tsukuba.ac.jp](http://www.tsukuba.ac.jp/).

利用者

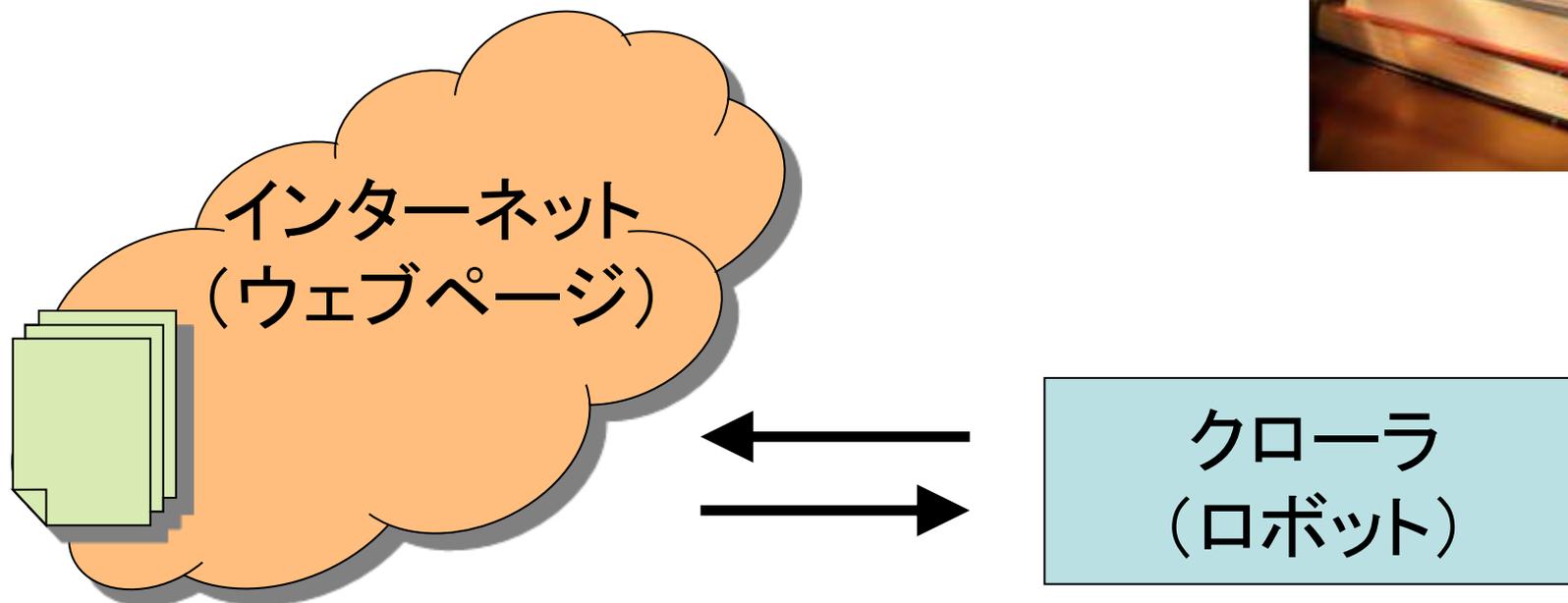
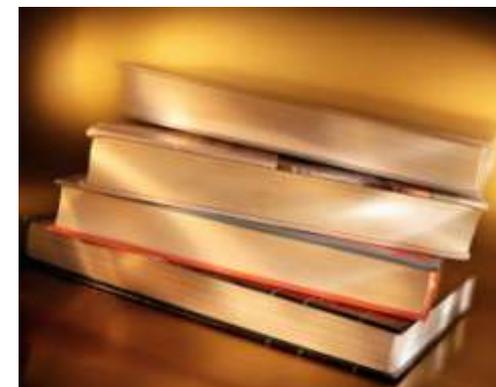
検索エンジン



- 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



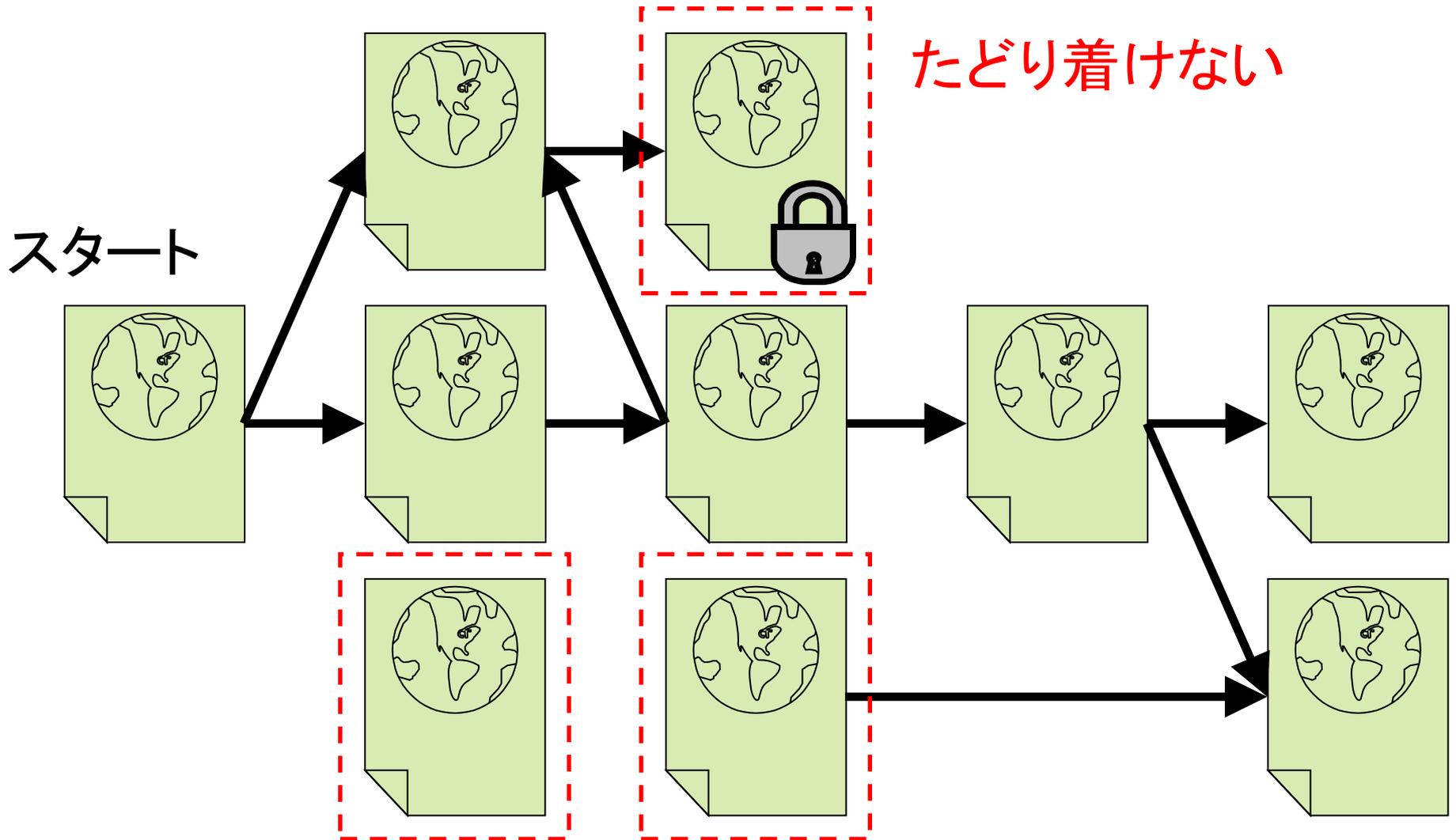
完了

tsukuba.ac.jp

```

次のソース: http://www.tsukuba.ac.jp/ - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) ヘルプ(H)

<div id="ul_topics"> 
    <!--// TOPICS S //-->
    <li><a href="http://www.tsukuba.ac.jp/research/first.html">いまさら聞けない、生物多様性ってなに？を考える会開催
    <li><a href="http://www.tsukuba.ac.jp/research/first.html">分子行動科学研究コア(最先端研究開発支援プログラム)サイト開設
    <li><a href="http://www.tsukuba.ac.jp/research/first.html">第1回ナチュラリスト養成講座開催
    <li><a href="http://www.tsukuba.ac.jp/research/first.html">筑波大学東京キャンパス(大塚地区)の仮校舎移転について
    <!--// TOPICS E //-->
  
```



- 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)



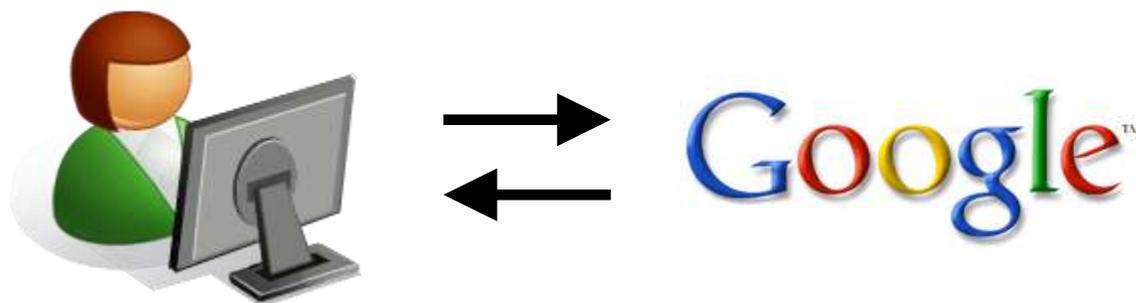
可愛い犬の写真



- 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）



大学 - Google 検索 - Mozilla Firefox

http://www.google.co.jp/search?num=100&hl=ja&safe=off&client=firefox-a&hs=2&js=org.mozilla%3Aja%3Aofficial&aq=大学&aq

Google

大学

約 296,000,000 件 (0.49 秒)

大学進学決定版 shingakunet.com リクルート進学ネット。全国の2300校の学校情報が満載！

他のキーワード: 東京大学 明治大学 京都大学 関西大学 立命館大学

大学 日本の大学
日本の大学はいかにいかなる角度から日本の全大学をご紹介。ナレッジステーションの人気ページです。
人文科学 - 首都圏 - 東京 - 佐賀県
www.gakkou.net/dagaku/ - キャッシュ - 類似ページ

大学 - Wikipedia
大学(だいがく)は学術研究および教育の最高機関。日本の現在の学制では高等学校もしくは中等教育学校卒業後、通常の課程による12年の学校教育を終了した者、またはこれと同等以上の学力を有する者を対象に専門的な高等教育をおこなうものとされている。...
歴史 - 日本の大学教育 - 世界の大学教育
ja.wikipedia.org/wiki/大学 - キャッシュ - 類似ページ

リクルート進学ネット / 大学・短期大学・専門学校情報
進学ネットでは大学・短期大学(短大)・専門学校の情報を紹介しています。仕事や資格、勉強したい内容から大学・短期大学・専門学校を探ることができます。学校見学会、オープンキャンパスや入試・出願情報も多数掲載！
shingakunet.com/ - キャッシュ - 類似ページ

大学のニュース検索結果

全日本大学野球: 北大初勝利、東海大コールド勝ち 1回戦 - 6時間前
全日本大学野球選手権38日、神宮球場と東京ドームで開催し、1回戦18試合を行った。4度目の出場となる北大は四国学院大を破り、1回戦大勝として初勝利を挙げた。13年連続出場の東海大は白熱六に八回コールド勝ちし好成績。情勢大は先発したエース...
毎日新聞 - 関連記事 45 件。 [立正大 初の遠征実り2部へ / 京都大学 - 日刊スポーツ - 関連記事 49 件。](#) [北大が初勝利の全国初勝利!! / 大学進学指 - 日刊スポーツ - 関連記事 10 件。](#)

東京都渋谷区 付近の大学の検索結果 - 場所を変更

西新宿 新橋西口 新橋 新宿三丁目 四谷三 新宿御苑前

国道大学 www.utsu.edu - 03-5467-1212 - レビュー (1)

東海大学付属望星高等学校 www.bosai.tokai.ed.jp - 03-3467-9111 - 詳細

google.co.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

2000年~

ページそのものを評価する試み

- HITS

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)

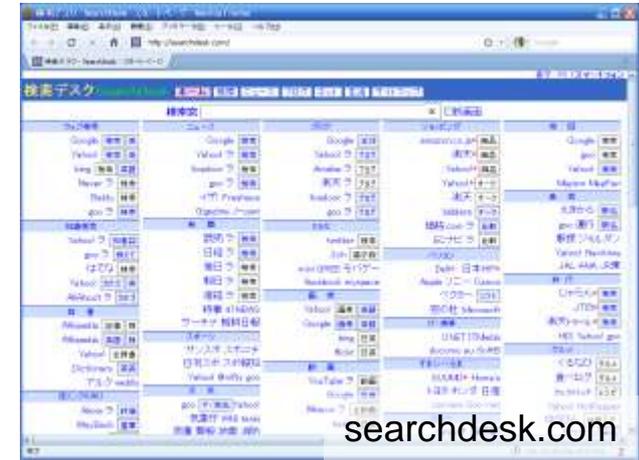
- CEEK.JP
 - 統合型メタサーチエンジン
 - 2002年8月開始
- 当時のウェブ検索エンジン
 - 定番が無い状態
 - 複数の検索エンジンを使い分ける



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)

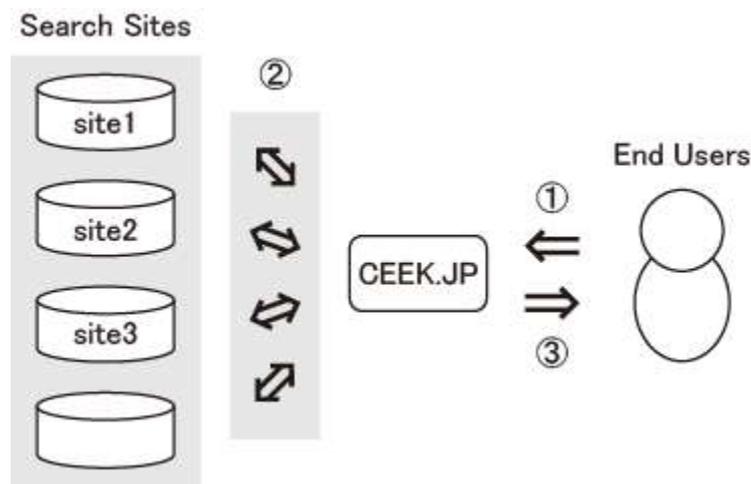
- 当時の検索方法

- 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

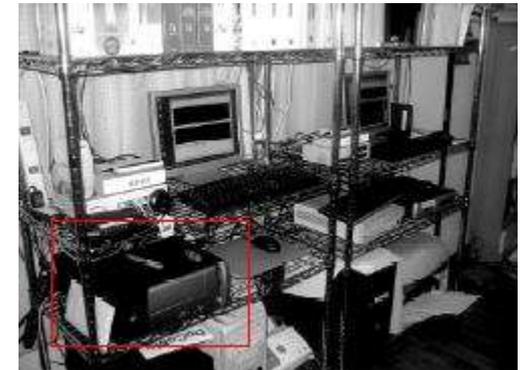
- 初期(2003年2月)

- 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

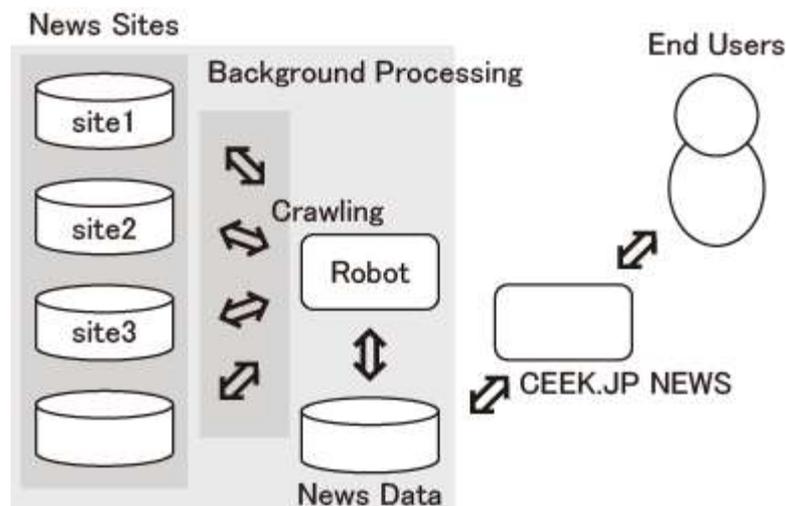
- サーバを準備しないと
 - 共有サーバでの運用は難しい
(クローラを動かす事が出来ない)
 - 専用サーバは高い
 - 自前でサーバを構築しよう！
「ママ、どうしておうちにサーバーがあるの？」
マイクロソフト社が作成したパンフレットのタイトル
- メリット
 - 暖房要らない :)
 - ハードウェアの自由度が高い
(メモリ, SSD)



初期(登宅)



現在(自宅)



- 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)てつくてつく

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)

- (有)てくてく
 - 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



- 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

- (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

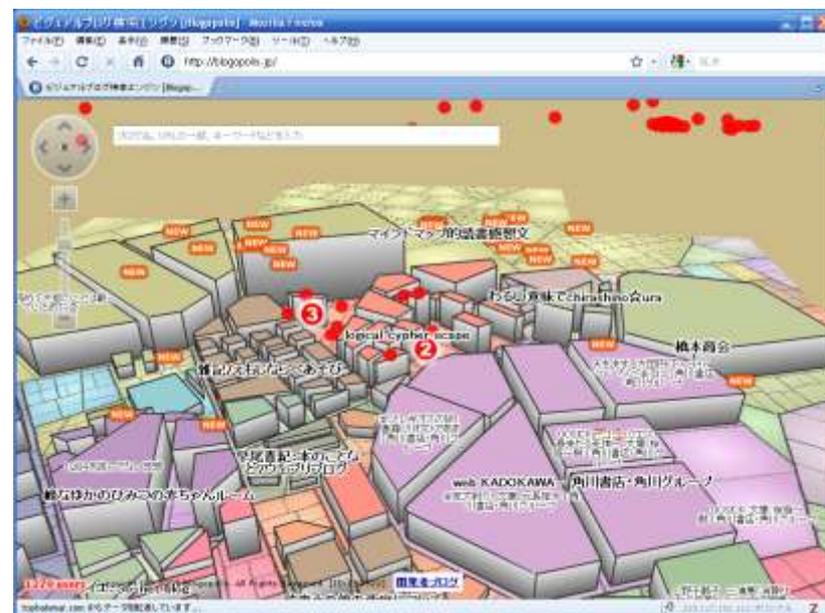
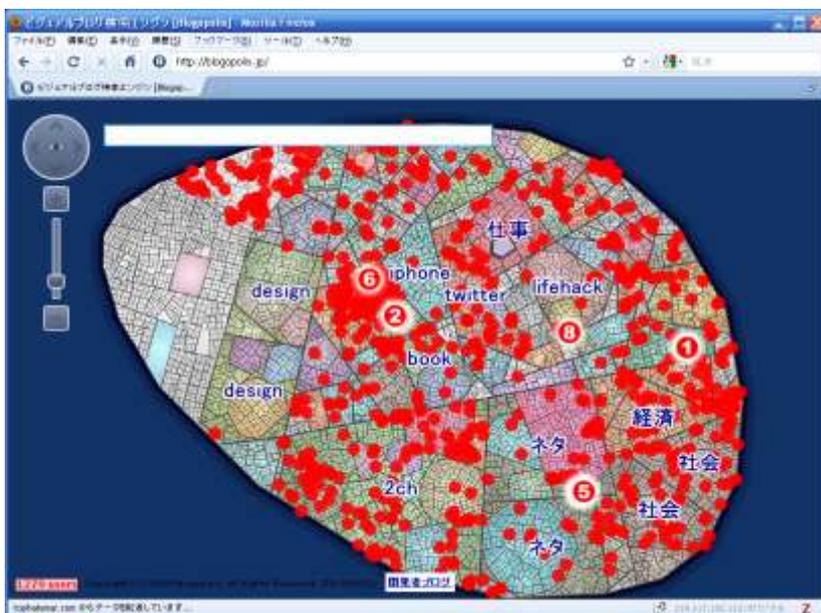
- 弊社のビジネスモデル
 - 運営サイトを見てシステム発注依頼
 - 保守契約等で細々と...
- 今後のビジネス
 - CEEK.JP NEWS 有償サービス
 - 検索エンジンを作る

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
 - Blogopolis
 - <http://blogopolis.jp/>



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

- ○○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すか？
 - 最初から可能性を狭める必要は無い

- 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

- 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

- 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