2005年6月18日
大図研オープンカレッジ

インターネット検索と大学図書館

兼宗 進
(一橋大学 総合情報処理センター)

今日の構成

- 第1部: インターネット検索の歴史
  - ■最初の10年間は図書検索の歴史をたどってきた
- 第2部: インターネット検索の現在
  - ■我々は、何から何を探し何を見ているのか?
- 第3部: インターネット検索のこれから
  - ■検索エンジン以降の動向
- 第4部: 図書館への期待
  - ■インターネット検索から何を学べるか

第1部: インターネット検索の歴史

最初の10年間は図書検索の歴史(-2000)

- ✧ リスト
- ✧ ディレクトリ
- ✧ 全文検索

WWWの歴史

- ハイパーテキスト
  - ■1945年頃からアイデア
- WWW(World Wide Web)
  - ■1989年: Tim Berners-Lee
  - ■1993年: WebブラウザMosaic
  - ■1999年: iモード、Windows98

http://www.seigyo.e.dendai.ac.jp/~comp_1e/learning_html/


http://www.w3.org/People/Berners-Lee/

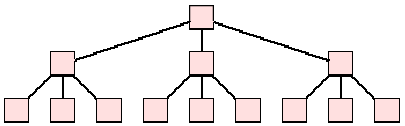
初期の検索(1)

- リスト(リンク集)

- 数百件を超えると見づらくなる

初期の検索(2)

- リストからディレクトリへ
  - ■線形配列から階層構造へ(n → n x n x ...)
- Yahooなど
  - ■人手で登録、分類。階層構造(十進分類との類似性)
  - ■特徴: サイト単位。(○)質。(×)網羅性と鮮度



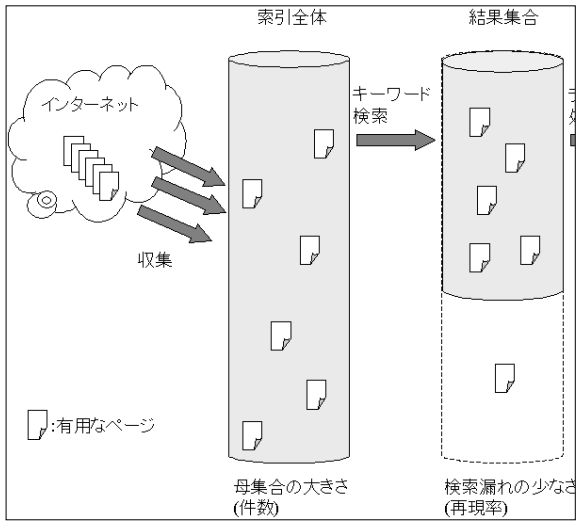
http://www.yahoo.co.jp/
http://www.google.com/dirhp?hl=ja

検索エンジンの登場

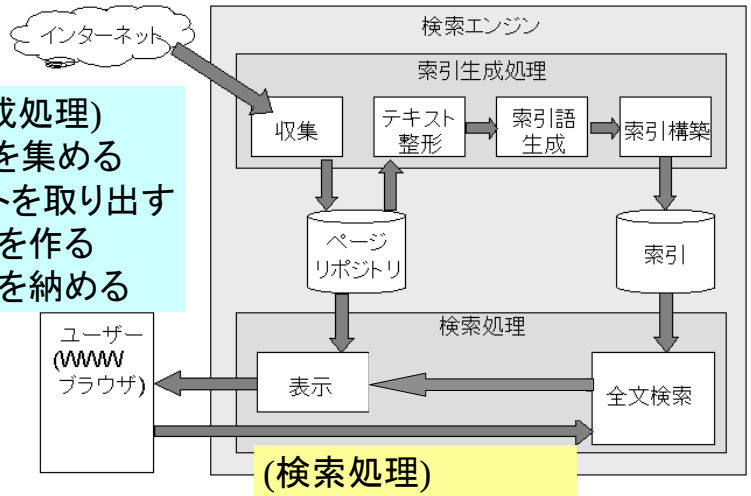
- サイト数の増大
  - ■ブラウザの普及(Netscape、IE) → 一般への普及
  - ■サイト数の増加→人手の登録が追い付かない! (限界)
    → 検索エンジンの登場
- 100万サイト以上
  - ■ディレクトリ: 人手による網羅的な登録は破綻
  - ■検索エンジン: 機械的に収集、ページ単位の検索
- 検索エンジンの出現
  - ■初期の検索エンジンについて基本的な原理を解説する

検索の流れ

- 旧来の情報検索のモデル
  - ■検索語で絞り込む



検索エンジンの構成



(索引生成処理)
1.ページを集める
2.テキストを取り出す
3.索引語を作る
4.索引語を納める

(検索処理)
1.索引語を検索する
2.要約を表示する

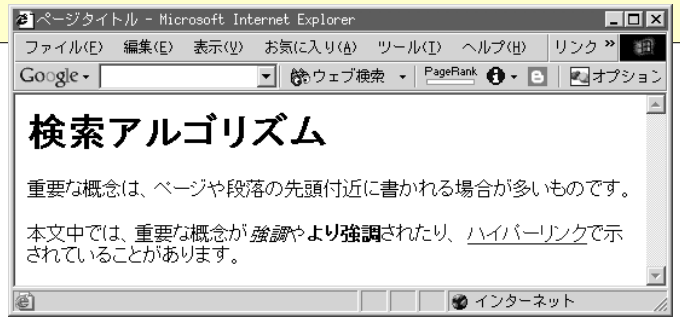
(1)収集

- WWWページ
  - ■誰もが勝手に公開
  - ■3億ページ(1999)→200億ページ(2004)?
    数えられない!!
- ページの存在を知るには
  - ■我々がブラウザでリンクをクリックするのと同じ作業をプログラム(robot, spider, crawler)が高速に行う。
    機械版ネットサーフィン
  - ■検索エンジンは数十億ページを収集して蓄える

(2)テキストの取り出し

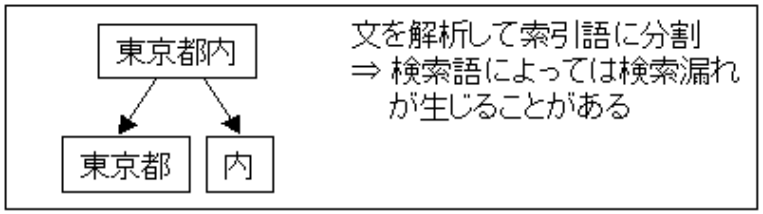
- 画面に表示される情報が中心

```
<html lang="ja">
<head>
<title>ページタイトル</title>
</head>
<body>
<h1>検索アルゴリズム</h1>
<p>重要な概念は、ページや段落の先頭付近に書かれる場合が多いものです。</p>
<p>本文中では、重要な概念が<em>強調</em>や<strong>より強調</strong>されたり、
<a href="link.html">ハイパーリンク</a>で示されていることがあります。</p>
</body>
</html>
```



(3)索引語の生成(1)

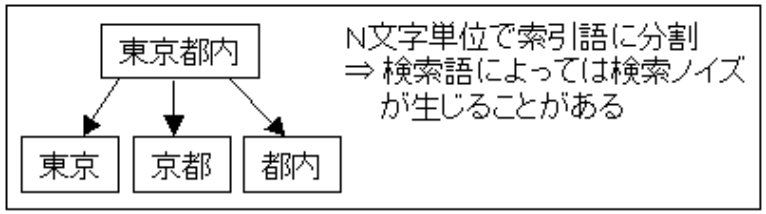
- 日本語: 単語の区切りがない→分割する
- 手法1: 形態素解析(単語ごとの分割)
  - ■辞書と解析プログラムで日本語を分割する



【検索例】
東京都 ⇒ 検索される
京都 ⇒ 検索されない
都内 ⇒ 検索されない(検索漏れ)

(3)索引語の生成(2)

- 手法2: N-gram(文字ごとの分割)
  - ■隣接するN文字で分割する



【検索例】 東京都 ⇒ 検索される
京都 ⇒ 検索される(検索ノイズ)
都内 ⇒ 検索される

第1部のまとめ

- 最初の10年間は図書検索と同じ
  - ■WWWの爆発的な増加→人手から自動索引へ

年	検索手段	図書検索との対応	キーワード
1990	リスト	図書原簿	一覧表
1994	ディレクトリ	分類目録	分類
1995	初期の検索エンジン	DB検索 OPAC	絞り込み
2000	検索エンジン	第2部	
2005	第3部		

第1部

- 初期の検索エンジン
  - ■絞れない → ゴミページがヒット →使い物にならない

第2部: インターネット検索の現在(2000-)

我々は、何から何を探し何を見ているのか?

✧ 現在の検索エンジン

現在の検索エンジン

- WWW利用の中心
  - ■数十億ページから数十万件ヒット
  - ■適当なキーワードを入れれば役に立つページが出る
    → なぜか?

Googleは80億ページ収集

マイナーな用語(メタデータ)でも3万ページヒット
(metadataでは200万件ヒット)

http://www.google.com/intl/ja/

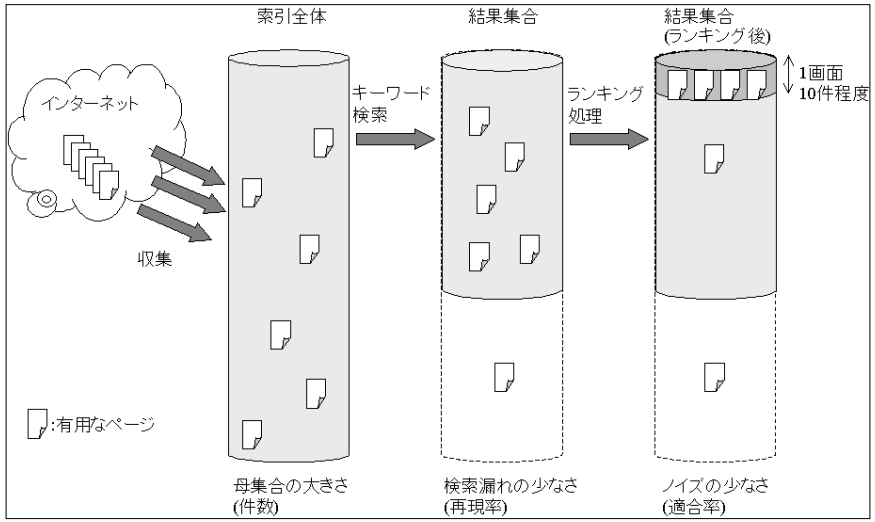
従来の情報検索との比較

- 検索の使われ方がまったく異なる

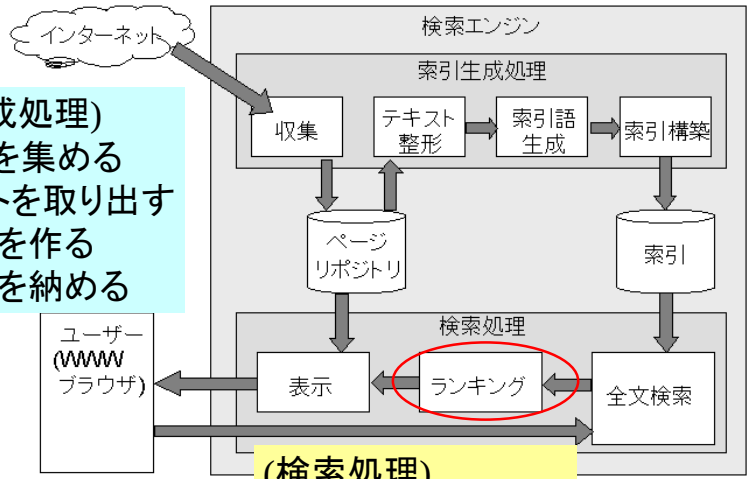
項目	従来の情報検索	検索エンジン
ユーザー	サーチャー	初心者
検索語	吟味した検索式	思い付いた1,2語
結果の閲覧	全件	先頭10件
絞り込み	する	しない
結果集合	数十～数千件	数万～数百万件
求める情報	網羅的	数件

検索の流れ

- 結果集合を絞らない
  - ■先頭数件しか見ない: 適切なランキングが命



検索エンジンの構成



(索引生成処理)
1.ページを集める
2.テキストを取り出す
3.索引語を作る
4.索引語を納める

(検索処理)
1.索引語を検索する
2.ランキングする
3.要約を表示する

ランキング

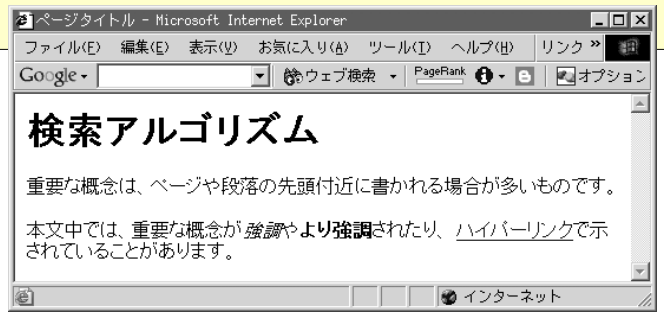
- よいページを先頭に表示できるかが勝負
- 「よいページ」とは?
  - ■さまざまな手法。検索エンジンごとの企業秘密
  - ■ページに点数を付ける(スコアリング)
  - ■3種類に分類してみた

基になる情報	スコアリング手法	類似の概念
データの特性	出現頻度、タグ、 出現位置、近接度	カーナビゲーション、 鉄道経路検索
ユーザーの行動	クリック人気	ベストセラー情報
ユーザーの推薦	リンクポピュラリティ	文献の引用情報

データの特性によるスコアリング

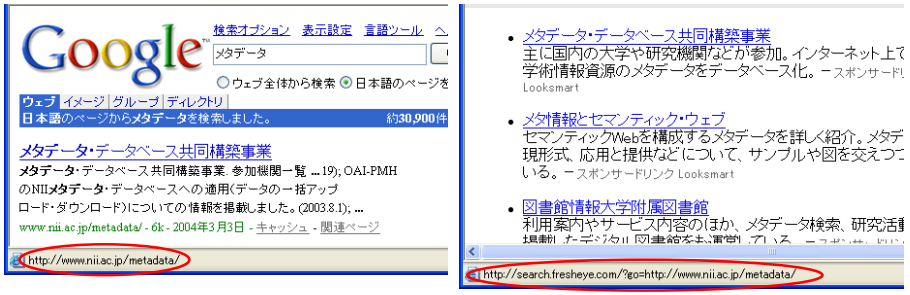
- HTMLを解析してスコアを付ける

```
<html lang="ja">
<head>
<title>ページタイトル</title>
</head>
<body>
<h1>検索アルゴリズム</h1>
<p>重要な概念は、ページや段落の先頭付近に書かれる場合が多いものです。</p>
<p>本文中では、重要な概念が<em>強調</em>や<strong>より強調</strong>されたり、
<a href="link.html">ハイパーリンク</a>で示されていることがあります。</p>
</body>
</html>
```



ユーザーの行動によるスコアリング

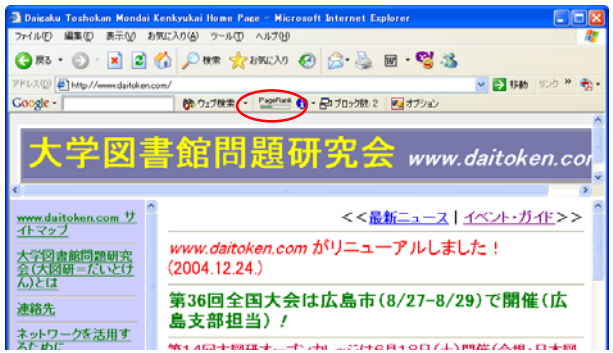
- ユーザーの行動観察からスコアを付ける
  - クリック人気
  - 行動を監視されるのはちょっと…?



http://www.google.com/intl/ja/
http://www.fresheye.com/

ユーザーの推薦によるスコアリング

- Googleの大躍進: 適切なランキング
- ページランクが基本
  - WWWページ中で他のページをリンク→推薦と見なす
  - 徳の高い(ランクの高い)ページからのリンクは有効
  - Googleツールバーで表示可能(例: 大図研はランク4)



第2部のまとめ

- 全文検索の時代
  - 素人でも満足する結果を得られる(画期的)
  - 成功の秘訣はランキング。その仕組みを見た
  - (参考)「検索エンジンの検索アルゴリズム」情報の科学と技術. 2004年2月号. pp.78-83
- 我々は、何から何を探し、何を見ている?
  - 何から: 収集した集合から(すべてのページではない)
  - 何を探す: キーワードを含むページ
  - 見るもの: 検索エンジンが選んだページ

第3部: インターネット検索のこれから(2005-)

検索エンジン以降の動向

- ✧ WWWの発展とRSS
- ✧ 検索対象の拡大

WWW生成の進化

- ■人手から自動生成へ

	従来	現在
HTML生成	人手	自動化(CMS)
更新	週単位	分単位
検索	検索エンジン	(今回の話題)

- ■更新が頻繁なサイトの例
  - ■ニュースサイト、Blog、日記
    → 検索エンジンの巡回が追い付かない！
- ■新しいニーズへの対応
  - ■アンテナ、RSS

アンテナの例(はてなアンテナ)

http://a.hatena.ne.jp/kameta/?gid=201511

RSSリーダーの例(Mozilla Thunderbird)

- ■メールソフトでRSS取得
  - ■WWWの更新が新着メールの形で通知される

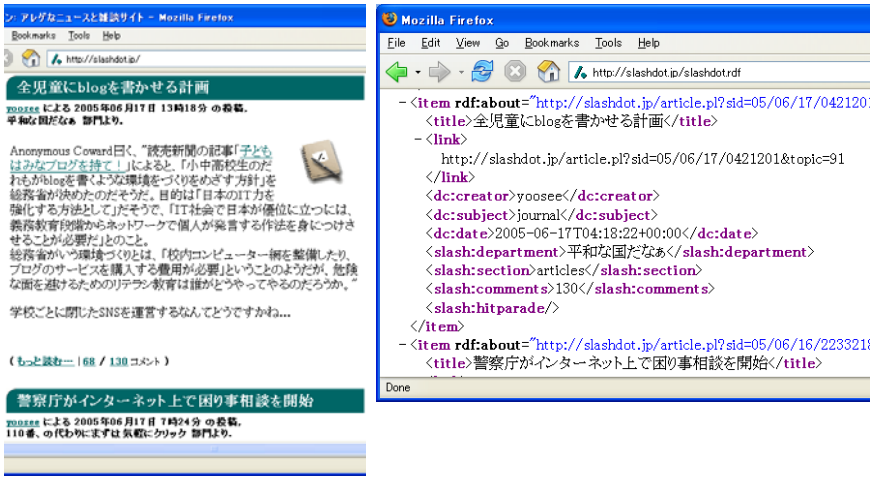
http://www.mozilla-japan.org/
http://www.mozilla.org/

メタデータ(RSS)による更新通知

- RSS
  - RDF Site Summary (など複数の規格)
  - HTML(収集を待つ) → HTML＋RSS(更新を公開)
  - 機械によるHTML生成→RSSを自動生成
  - Blog、Wikiなどが標準装備
- 基本構造
  - item: title, link, description
  - 名前空間にDublinCoreを使用可能(RSS1.0)

RSSの例(スラッシュドットジャパン)

- 記事のメタデータを公開
  - title, creator, date, subject など

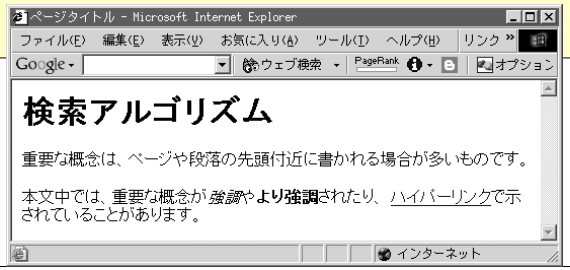


http://slashdot.jp/

(参考)WWW創成期からメタデータ

- HTML2.0(1995年)に存在
  - 広告などへの予期しない利用→頓挫(90年代後半)

```
<html lang="ja">
<head>
<title>ページタイトル</title>
<meta name="keywords" content="形態素解析,N-gram,索引">
<meta name="description" content="検索アルゴリズムのページです!">
</head>
<body>
<h1>検索アルゴリズム</h1>
<p>重要な概念は、ページや段落の先頭付近に書かれる場合が多いものです。</p>
<p>本文中では、重要な概念が<em>強調</em>や<strong>より強調</strong>されたり、
<a href="link.html">ハイパーリンク</a>で示されていることがあります。</p>
</body>
</html>
```



検索対象の拡大

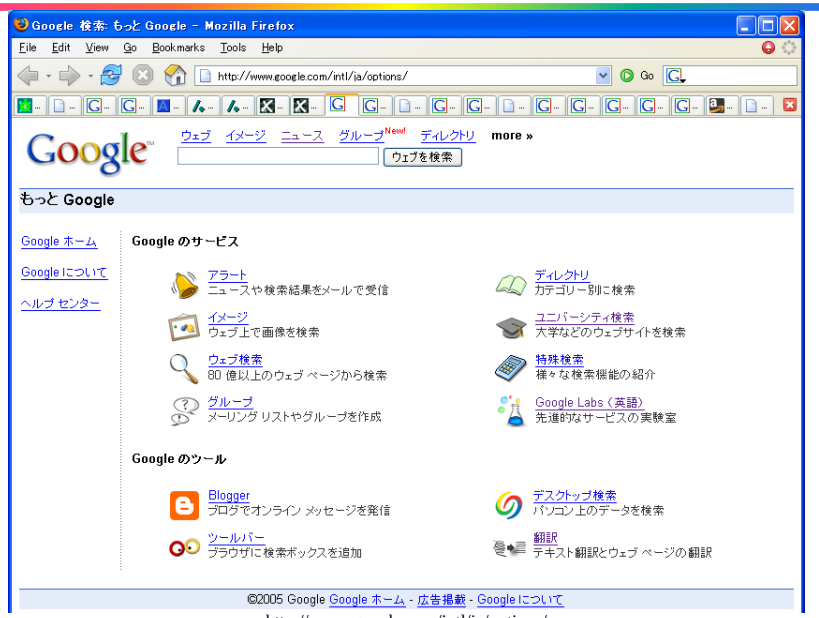
- WWW検索
  - テキスト + 画像 + PDF + PPT + ...
  - 新着(Alert)
- 文献情報
  - 学術文献(Google Scholar)
  - 書籍(Google Print)
  - 図書館(Worldcat)
- その他
  - 地図(Google Maps)
  - ニュース(Google News)
  - PC内のデータ(Google Desktop Search)

新着通知(Yahoo!アラート)

- WWWや各種サービスの更新を通知

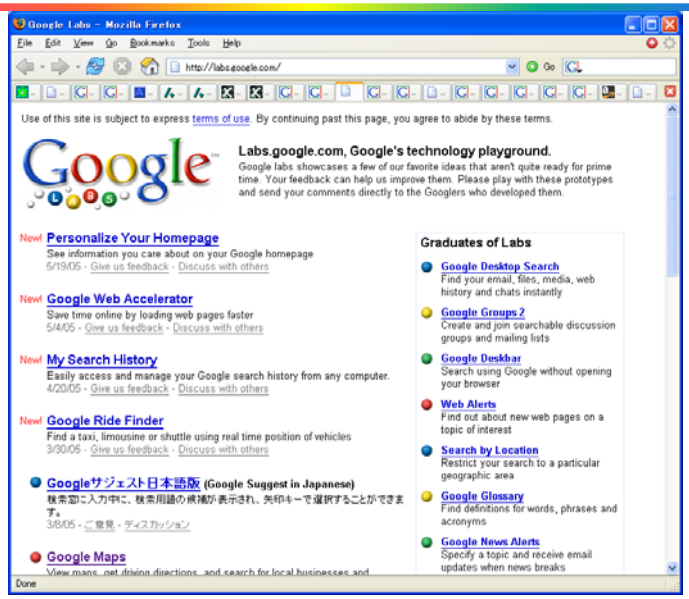
http://alerts.yahoo.co.jp/

Googleの各種サービス(日本語版)



http://www.google.com/intl/ja/options/

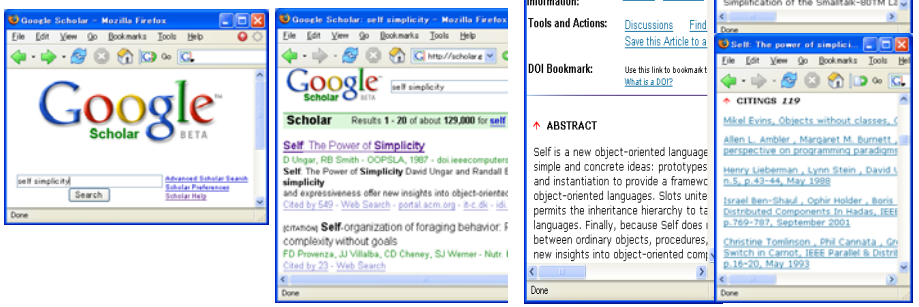
Googleの実験機能(英語版)



http://labs.google.com/

Google Scholar

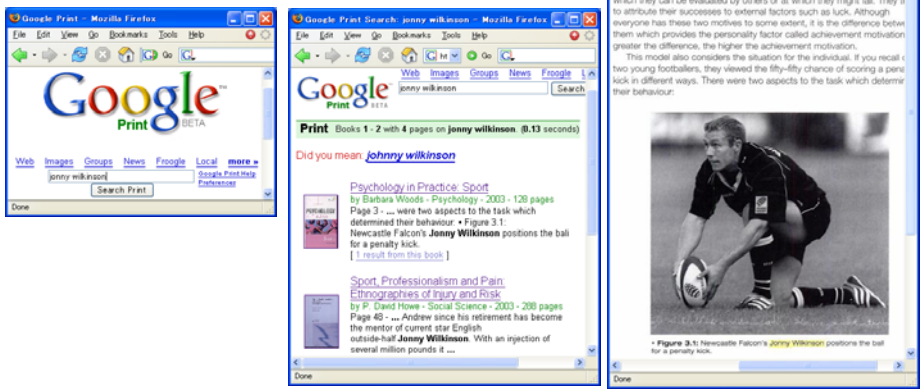
- 学術文献の検索
  - 文献検索→文献サイトへ
  - ACM(米国計算機学会)にリンクした例:
    - ➡書誌、抄録、参考、参照を閲覧
    - ➡PDFで本文参照(無料/有料)



http://scholar.google.com/

Google Print

- 書籍の本文を検索、表示
  - 出版社の許諾を得て電子化
  - 図のcaptionで検索された例:



http://print.google.com/

WorldCat

- 図書館の書誌を検索
  - OCLCが代表的な書誌をGoogle, Yahooに提供
  - Google:検索条件に site:worldcatlibraries.org を追加
    (例)「shirota yukari site:worldcatlibraries.org」
  - 大学の所蔵まで表示

http://www.google.com/

Google Maps

- 全米の地図と衛星写真を表示
  - 空港近くのホテルを検索した例:
    (飛行機まではっきり見える)

http://maps.google.com/

Google News

- 新聞社・通信社などのニュースを統合表示

http://news.google.com/nwshp?hl=ja

Yahoo!商品検索

- 通信販売サイトを検索、価格比較

http://psearch.yahoo.co.jp/

第3部のまとめ

- 基本的なWWW検索は完成
  - ■現在の検索エンジン: そこそこ便利に使えている
- 補完する方向へ
  - ■検索エンジンの苦手な部分(動的生成、更新頻度)
    → メタデータ(RSS)
  - ■資料提供の充実
    → 検索対象の拡大

第4部: 図書館への期待

インターネット検索から学べること

- ✧ 資料提供の範囲
- ✧ 検索技術の進化
- ✧ 図書館員への期待

資料提供の範囲

- 大学図書館の役割
  - ■蔵書を守ること?
  - ■研究・学習のサポート? → できることは多い

利用者のニーズ ＝ 資料の入手 ≠ 図書館からの提供

資料	提供している	していない
本	蔵書	訪問(大学、公共) 購入(書店、通販、古書)
論文	蔵書 電子ジャーナル 複写	購入(電子ジャーナル)
WWW	リンク集	検索(WWW)

利用者のニーズ

資料提供の範囲

- 大学図書館の役割
  - ■蔵書を守ること？
  - ■研究・学習のサポート？ → できることは多い

利用者のニーズ ＝ 資料の入手 ≠ 図書館からの提供

資料	提供している	していない
本	蔵書	訪問(大学、公共) 購入(書店、通販、古書)
論文	蔵書 電子ジャーナル 複写	購入(電子ジャーナル)
WWW	リンク集	検索(WWW)

利用者のニーズ

検索技術の進化

- インターネット検索:15年で劇的に進化した
  - ■検索エンジンのランキング(ページの推薦)
  - ■オンライン書店のレコメンデーション(書籍の推薦)
- 大学図書館は？
  - ■OPAC: 初期の検索エンジンと同じモデル(化石？)
    → なんとかしたい
- 図書館員への期待
  - ■図書館員の専門性: 目利きであること!!
    → 選書、推薦
  - ■この専門性を何とか活かしたい

全体のまとめ

- インターネット検索の進化
  - ■WWWの爆発的な普及(200億ページ)
  - ■図書検索モデル→検索エンジン (+ RSS)
- 検索の歴史から学べること
  - ■インターネット検索の本質は、適切な資料の推薦
  - ■大学図書館員の専門性
    → 資料提供範囲、選書/推薦を真剣に考えるべき
- 今後
  - ■メタデータ(RSS)と検索エンジンの図書検索に注目
  - ■いっしょに進めていきましょう！

(参考) http://kanemune.cc.hit-u.ac.jp/