2004年9月11日
TP&D2004

インターネット検索の原理とメタデータの利用

兼宗 進
(一橋大学 総合情報処理センター)

目次

第一部: WWW検索の歴史

1. ディレクトリから検索エンジンへ
2. 全文検索の時代

1.ディレクトリから検索エンジンへ

- WWWの歴史
  - 1989年:WWW発明(HTML、URL)
  - 1993年:Mosaic、1994年:Netscape、1995年:IE
- 発展と現在の状況
  - リスト×
  - ディレクトリ○
  - 初期の検索エンジン(従来の情報検索)×
  - 現在の検索エンジン(Google以降)◎

検索エンジン以前

- リストからディレクトリへ
  - 線形配列から階層構造へ
  - ディレクトリは検索エンジンを補完する形で現役
- Yahooなど
  - 人手で登録、分類。階層構造(十進分類との類似性)
  - サイト単位。網羅性は低い

2.全文検索の時代

- WWWの爆発的な発展(1億サイト以上)
  - ディレクトリ: 人手による登録→網羅性は破綻
  - 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: 全盛期
  - 手軽で満足度高い(画期的!)
  - ド素人が情報検索をして、満足するデータを引き出せているのはなぜか?
- その理由と原理を明らかにする

検索エンジン

- WWW利用の中心
  - 数十億ページから検索～数十万件ヒット～10件表示

Googleは43億ページ収集

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)

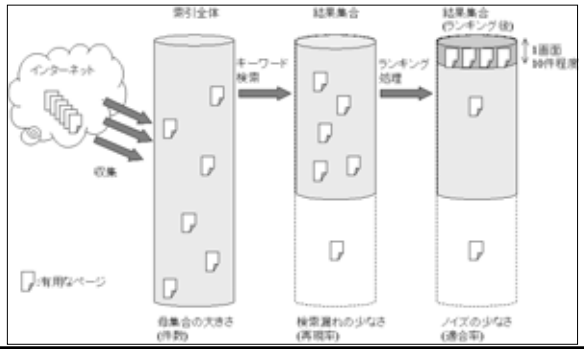
従来の情報検索との比較

- 検索の使われ方がまったく異なる

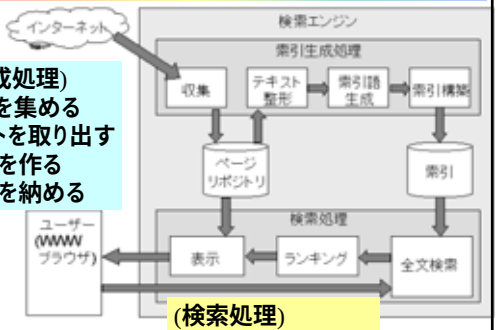
項目	従来の情報検索	検索エンジン
ユーザー	サーチャー	初心者
検索語	吟味した検索式	思い付いた1,2語
結果の閲覧	全件	先頭10件
絞り込み	する	しない
結果集合	数十～数千件	数万～数百万件
求める情報	網羅的	数件

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

(1)収集

- WWWページ
  - 勝手に公開
  - 100億ページ? 数えられない
- ページの存在を知るには
  - リンクをたどる
  - いくつかのリンク集(Yahooなど)から、「リンクをたどってページを読んでそこに含まれるリンクをたどる」という作業を延々と繰り返す
  - 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
  - 検索エンジンは数十億ページを収集して蓄える

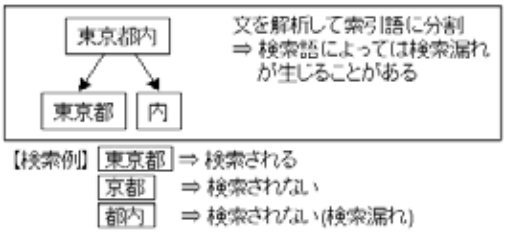
(2)テキストの取り出し

- 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

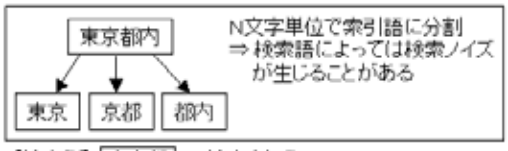
- 日本語: 単語の区切りがない→分割する
- 手法1: 形態素解析(単語ごとの分割)
  - ■辞書と解析プログラムで日本語を分割する



【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

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

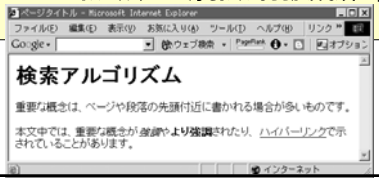


【検索例】東京都 ⇒ 検索される
京都 ⇒ 検索される(検索ノイズ)
都内 ⇒ 検索される

(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>
```



(4)ランキング

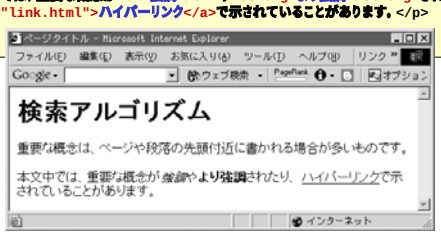
- よいページを先頭に表示できるかが勝負
- 「よいページ」とは?
  - ■さまざまな手法。検索エンジンごとの企業秘密
  - ■ページに点数を付ける(スコアリング)
  - ■3種類に分類してみた

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…?

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: 私図協はランク6)

第一部「WWW検索の歴史」のまとめ

- ディレクトリから検索エンジンへ
  - ■WWWの爆発的な増加→人手から自動索引へ
- 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)
  - ■適切な数件を表示。網羅性不要
  - ■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

第二部: 新しいWWWの流れ

1. WWW生成の進化と検索エンジンの限界
2. メタデータ(RSS)による更新通知

1.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	(第二部の話題)

- 更新が頻繁なサイトの例
  - ■ニュースサイト
  - ■Blog、日記
  - ■Wiki、掲示板

ニュースサイトの例

- 最新のニュースを公開
  - ■新聞社、ポータルサイト、…

Blogサイトの例

- 他のサイトを参照しながら日記やコメント
  - ■Webブラウザから記入
  - ■相互の連携: コメント、リンク、TrackBack、RSS

Wikiサイトの例

- WWWブラウザでページを編集
  - Wiki: 誰でも書ける。みんなで作っていくWeb

Wikiサイトの例

- WWWブラウザでページを編集
  - Wiki: 誰でも書ける。みんなで作っていくWeb

2.検索エンジンの限界とRSS

- 検索エンジンの索引更新
  - 巡回して索引に反映
    - ➡数十億ページの短期間での巡回は不可能
    - ➡巡回頻度に重みを付けて対応(巡回戦略)
  - 数日から数週間→遅すぎる!
- 新しいニーズの出現

ページの種類	ディレクトリ	検索エンジン	RSS
更新されない	◎	○	?
週単位の更新(A)	×	◎	?
分単位の更新(B)	×	×	○
DBから作成(C)	×	×	×
他からリンクされない	×	×	×

メタデータ(RSS)による更新通知

- 従来: 受身で待つ
  - 更新: 検索エンジンの巡回を待つ(数日から数週間...)
  - 索引: HTMLを解釈してもらう
- 新しい流れ: 自発的に発信
  - メタデータを公開(RSS)
  - 引用元へ報告(TrackBack)
  - 更新通知(Ping)

RSSの構造

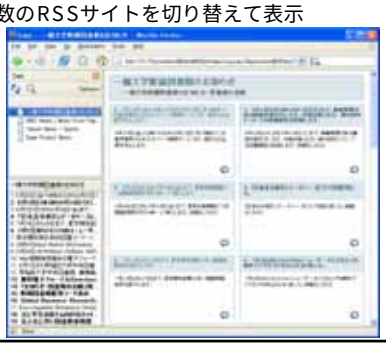
- RSS
  - サイトの見出しを公開
  - 機械によるHTML生成→RSSを自動生成
  - Blog、Wikiなどが標準装備
  - 名前空間にDublinCoreを使用可能(RSS1.0)
- 種類: 複数の規格が並存
  - RSS0.91/0.92: Rich Site Summary
  - RSS1.0: RDF Site Summary
  - RSS2.0: Really Simple Syndication
  - Atom: Atom Syndication Format

RSS1.0 (RDF Site Summary)

- RDFを採用
  - RDF: Resource Description Format
  - XMLでメタデータを記述するための枠組み
- 基本構造
  - channel: title, link, description, items
  - item: title, link, description
- モジュールによる拡張が可能
  - DublinCore, Syndication, Contentなど

RSSの利用(クライアント)

- RSSリーダー
  - ■アプリケーション、Webブラウザのプラグインなど
  - ■複数のRSSサイトを切り替えて表示



RSSをサポートするサービス

- RSSを補助する技術
  - ■TrackBack: 引用したことを相手に通知する
  - ■Ping: 更新情報を専用サーバーに送信する
- RSSの検索エンジン
  - ■Pingサーバーを利用。数分～数時間で巡回
  - ■BulkFeeds: 32万サイト(390万記事)
  - ■FeedBack: Blogに特化。8万サイト(330万記事)

第二部「新しいWWWの流れ」のまとめ

- 新しいニーズの出現
  - ■更新頻度の高いページが増加
  - ■RSSで検索エンジンを補完
- メタデータの復権
  - ■HTMLのメタデータ: SPAM行為で挫折
  - ■RSSのメタデータ: 広まりつつある

第三部: 図書館への応用

1. RSSの利用例
2. 新着資料への応用

1.RSSの利用例

- RSS: 更新通知に適したメタデータ
- 利用例
  - ■お知らせをRSSで公開(一橋大学附属図書館)
  - ■ベストセラー情報を公開(TRCほか)
- 共通する利点
  - ■従来: 定期的にWWWを見に行く必要があった(不便)
  - ■RSSで: RSSリーダーでチェックできる(便利)

(参考)他の方式との比較

- Alert(OPAC/電子ジャーナル)
  - ■WWW表示、メール通知: 疎遠か押し付けがましい
  - ■単独サイト: 複数サイトを見に行くのが大変
  - ■ユーザー: 登録しないと使えない
- OAI-PMH
  - ■メタデータ収集プロトコル
- Z39.50
  - ■横断検索。新着ではない

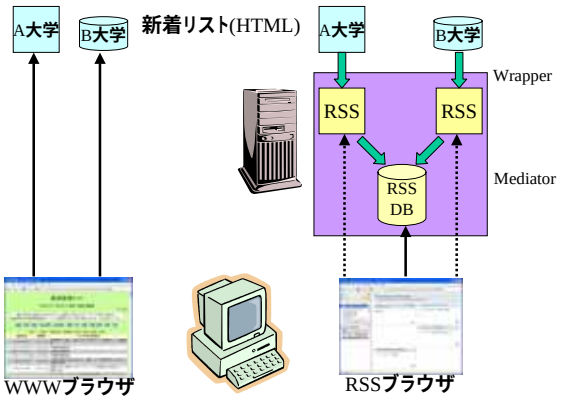
2.新着資料への応用

- 新着資料を扱うシステムを検討する
- 従来の新着一覧の難しさ
  - ■大学規模になると件数が多すぎる
  - ■個人ごとの興味に対応していない
  - ■複数の図書館の新着情報を統合して見られない
- RSSによる事例
  - ■農林水産省「新着資料案内」
  - ■所在場所(全国の図書室)ごとに表示
  - ■1日あたり数件

システムの試作

- 大学図書館の新着資料
  - 更新頻度: 日単位(B)
  - システムごとに形式が異なる
    - ➡ 一橋大: HTMLで公開(A)
    - ➡ 文教大: データベースから動的に生成(C)
- ラッパー・メディエータ方式で構成
  - Wrapper: 新着情報を統一形式(RSS)に変換
  - Mediator: 新着情報を統合的に提供
- ユーザーからのリクエスト
  - 検索条件を指定してリクエストする

システム構成図



考察

- 便利
  - ■Webほど疎遠でない。メールほどうるさくない
  - ■複数サイトの新着を扱える
- Alertと共通の課題
  - ■資料の選択条件(キーワード、分類、...)
  - ■OPACの新着リストに情報が少ない
- 今後の改良点
  - ■個人ごとの条件登録
  - ■図書館ごとのWrapperを容易に記述する仕組み
    (RSSを図書館システムの標準にしてほしい)

3.(最後に)図書館の資料検索について

- 図書館の資料検索を考える
  - ■リスト(目録)
  - ■ディレクトリ(図書カード)
  - ■初期の検索エンジン(現在のOPAC)
  - ✕ ■現在の検索エンジン(★存在しない★)

- 研究の必要性
  - ■現代の検索エンジンに相当する
    「手軽で満足度の高いOPAC」が存在しない
  - ■肝心の資料検索が取り残されている

まとめ

- WWWは進化→情報検索も進化
  - ■分類: ディレクトリ
  - ■全文検索: 検索エンジン
  - ■メタデータ: RSS

- メタデータの復活
  - ■RSSはメタデータ。注目に値する流れ
- 図書館への応用
  - ■可能性がある → 研究していきましょう!

(参考) http://kanemune.cc.hit-u.ac.jp/