集合知プログラミング
第4章　検索とランキング

サ　ミンソン

はじめに

- 全文検索エンジンの利用
  - ー単語のリストで膨大なドキュメントを検索可
  - ー検索結果のドキュメントを単語に関連する順序にランク付けして表示可能
- 本章で学ぶもの
  - ーPageRankアルゴリズム（Googleで使われている）
  - ークロール
  - ーインデックス
  - ーページの集合を検索するのに必要なステップ

４．１検索エンジン

- 検索エンジンを作るためのステップ

  ①ドキュメントを集める手段を作り上げる

  ②集めた物をインデックスする

  ③クエリを基にランク付けされたドキュメントのリストが返される

４．２シンプルなクローラ（１）

- クローラ

  ー既知のHTML文書の新しいコピーを要求し、文書中に含まれるリンクをたどり別の文書を収集するという動作を繰り返す

・シンプルなクローラの作り方

ページをダウンロードし、それをインデクサに渡す。そのページに含まれている、次にクロールするページたちへのリンクを探すためにパースする

ーこの機能のライブラリが存在

４．２シンプルなクローラ（２）

- ライブラリの紹介

ーurllib2：ページを簡単にダウンロードするためのライブラリ

ーBeautiful Soup

：構造化されたWebページの表現を作り上げるためのライブラリ、 壊れたHTMLのページにも強い

ークローリングの最中はどのようなページに出会うかわからないため、クローラを作るためには非常に役立つ

４．３インデックスの作成

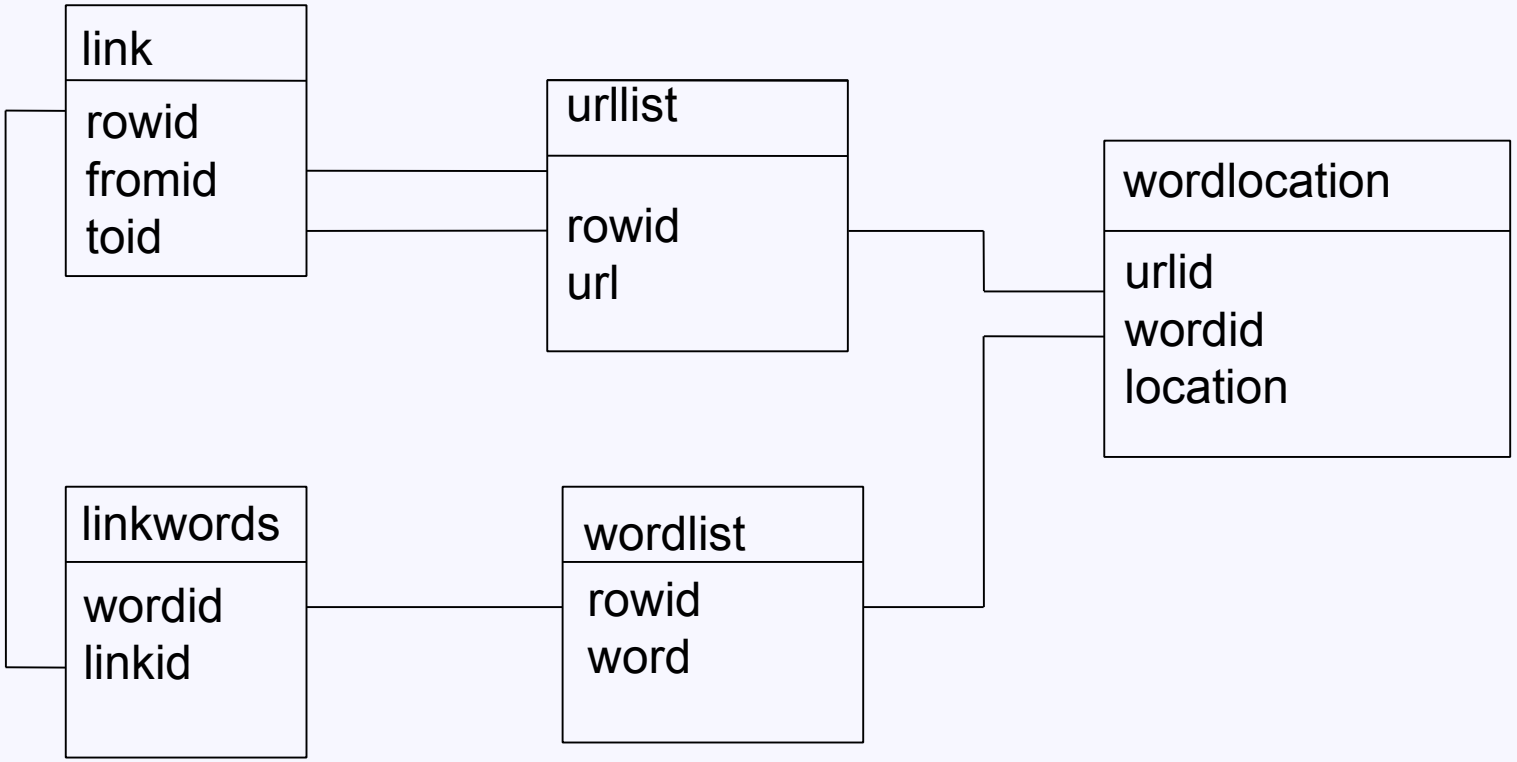
- 次のステップとして、フルテキストのインデックスのデータベースをセットアップする

インデックス：すべての異なる単語のリスト

　ーそれぞれの単語が現れる文書と、文書中で現れる位置を一緒に記録していく

４．３．１スキーマの設定

- 基本的なインデックスのスキーマは5つのテーブルで構成



4.3.2ページ内の単語を探し出す

- Webからダウンロードするファイルはhtmlで記述され、タグ、属性などインデックスには不要な情報が含まれている

- ステップ

①ページのテキスト部分を抽出する

②文字列をインデックスに保存できるよう、分割りされた単語たちのリストに分ける

４．３．３インデックスへの追加

③ドキュメント中での単語たちの位置についてのリンクを作り上げる

④ページ間のリンク関係を保存

⑤ページがデータベース中に既に存在するかチェック
→存在すればそのページに関連付けられた単語が存在するかチェック

・一度に一つの単語でしか検索できない　＋
ドキュメントは読み出された順に帰ってくる
→４．４（問い合わせ）で、複数の単語で質問できるように拡張