

OPAC検索ログによる キーワード補完計画 (第2回)

平成22年10月9日
図書系職員のためのアプリケーション開発講習会
前田朗



はじめに

- 大学OPACの検索ログを使い、学術向けのキーワード補完の仕組みを試作・評価する
- OPAC検索ログは、担当部署の許諾を得て入手
- キーワード補完を作る仕組みはいくつか考え付くが、まずはGETAssocを使い次のキーワードを連想検索させてみる。



次のキーワードを予測する

検索に使われるキーワード数を N とすると

- 直後のキーワードのみ着目する
 - $N-1$ の組み合わせが発生
- 同時に使われるキーワードを着目する
 - $N(N-1)/2$ の組み合わせが発生

同時に使われるキーワードに着目したほうがデータマイニングに使えるデータ量が増えるが、まずは「直後のキーワード」に着目してみる

簡略検索（キーワード欄のみ）の場合

The screenshot shows the OPAC search page with the following elements:

- Header:** 東京大学 OPAC THE UNIVERSITY OF TOKYO 図書館データベース
- Navigation:** 新着図書案内, 雑誌最新巻号案内, Webリクエストサービス, MyLibrary, ヘルプ, 問い合わせ ASK サービス, 新規検索
- Search Area:** キーワード検索 (Annotated with a green callout: 左から右にキーワードを拾う and a red arrow pointing to the input field)
- Filters:** 並び順 (dropdown), 一度に表示する件数 (20), 優先 (図書, 雑誌)
- Buttons:** 検索, クリア, ソフトウェア キーボード, 詳細検索へ, ヘルプ
- Left Sidebar:** English Version, 東京大学OPAC (checked), Webcat (学外), 検索対象 (図書, 雑誌, 特集記事), 所属キャンパス (全学, 総合図書館, 駒場図書館, 柏図書館, 法学部)

左から右にキーワードを拾う

詳細検索（複数検索項目）の場合

The screenshot shows the OPAC search page in Internet Explorer. A green callout box with the text "左から右 上から下にキーワードを拾う" (Pick up keywords from left to right, top to bottom) is positioned over the search criteria section. Three red arrows point from the callout box to the search criteria input fields: "あれや これや", "それや どれや", and "うえや しや".

東京大学 OPAC
THE UNIVERSITY OF TOKYO

English Version

東京大学OPAC
 Webcat (学外)

検索対象
 図書
 雑誌
 特集記事

所属キャンパス
(全学)
総合図書館
駒場図書館
柏図書館
法学部

全ての項目から AND
著者名に右の語を含む AND
出版社・出版者

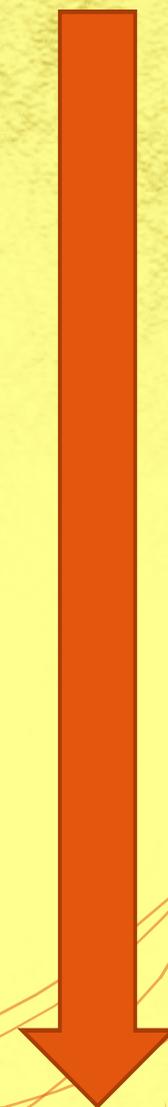
出版年 -
出版国 全て
言語 全て
分野 (全分野)
配架場所
並び順 一度に表示する件数 20
優先 図書 雑誌

検索 クリア ソフトウェア キーボード 基本検索へ ヘルプ

附属図書館ニュース
【耐震工事】
・法学部 長期閉室中(学内者のみ一部利用可能) [詳細](#)

ログデータの処理

1. キーワードの組を作成
 - 例)
 <author>前田/タブマイニング
2. キーワードの組をカウント
 - UNIXのsort+uniqコマンドで十分
3. GETAssocのitbファイル形式に整形
 - 例) GETAssocのサイトを参照
 @title=xxxxx
 #i=xxxxxxxxx
 #title=<keyword>ツール
 10 テスト
 3 図書館
4. GETAssocのインデックスを作成



ちょっとした問題

- GETAssocのUTF8文字列チェックにひっかかるデータがあった
- このデータのところでインデックス作成が異常終了してしまう..
- そこで次の対策をとった
 - インデックスの作成を2度実行
 - 最初の1回はエラーメッセージを出させる
 - 2回めが本番（問題データ排除済み）



とりあえずスルーした話

- データの正規化
 - 不要な記号をどこまで排除するか
- ISBNなど一意に文献を同定するIDの排除
- 文字化けしたデータの対応
- 長いフレーズでの検索
 - 東京大学OPACは自動で形態素解析を行うため
実質複数キーワードともいえる
- 評価の方法



次回にすすむ

