



OPAC検索ログを使った キーワード補完計画そのいち

平成22年9月20日

図書系職員のためのアプリケーション開発講習会

(平成22年9月24日

「マイニング探検会」向けに若干修正)

はじめに

- Web検索エンジンのキーワード補完
 - Yahoo! はAPIも公開
 - 検索ログを利用しているとのことだが、詳細は不明
- OPACでもそれに近いことを試してみる
 - キーワード補完ツールの作成
 - 学術分野に特化したキーワード補完の可能性を探る（Yahoo! 関連検索ワードや、人社会系・理工系の比較など）

テキストマイニングっぽいことを試してみよう！

OPAC検索ログの利用申請（要約）

- 利用目的： キーワード補完ツールの作成及び評価
- 利用範囲： 検索条件及び日時、IPアドレス等。IPアドレスはアクセス元の学部の判別に使用。
- 必要とするログデータは半年程度
- ログデータ本体は表に出さない
- 成果は随時一般公開可

現在のところ担当係による個別の許諾であり、特に規則があるわけではない

ログデータの入手

- 東京大学のOPACシステムでは、詳細な検索条件を記録したログを一日しか保持しない
- そこで、毎日の深夜に、講習会サーバにデータをコピー
- 実際に使えるデータが集まるのは、年末あたり（3ヶ月分程度のデータ）

解析しなくともわかること

- 思ったよりシングルキーワードが多い
 - ざっと眺めただけでも、フルタイトル検索が多い
 - 携帯なので単純に比較はできないが、以下に1語での検索が65%との情報を見ることができる。
 - 大阪市立大学携帯OPACの2005年のログ分析
<http://murakami.media.osaka-cu.ac.jp/papers/Kiyou06kanata.pdf>
- 個人的には、検索エンジンでは、複数キーワードを使うことが多いような...

推定：シングルキーワード

- 仮説 1
 - 「所在をさがす」ための利用が多い
 - テーマから本をさがす利用が少ない
- 仮説 2
 - 最初はまずシングルキーワードで様子見
- 仮説 3
 - 検索結果がランキングされていると思われる
- 仮説 4
 - 東大OPACではキーワードが形態素解析されるため、実質複数キーワードになっている

GETAssocによるキーワード補完

- あるキーワードの次に続く語を「連想検索」させる
- 検索項目も活かす（東大図書館 → <著者>前田 朗 とか）
- GETAssocの入力データ形式（itbファイル）で、次の書式を使えば簡単に単語・単語行列を作成可能（だと思う）
 - 数値[space] *value*

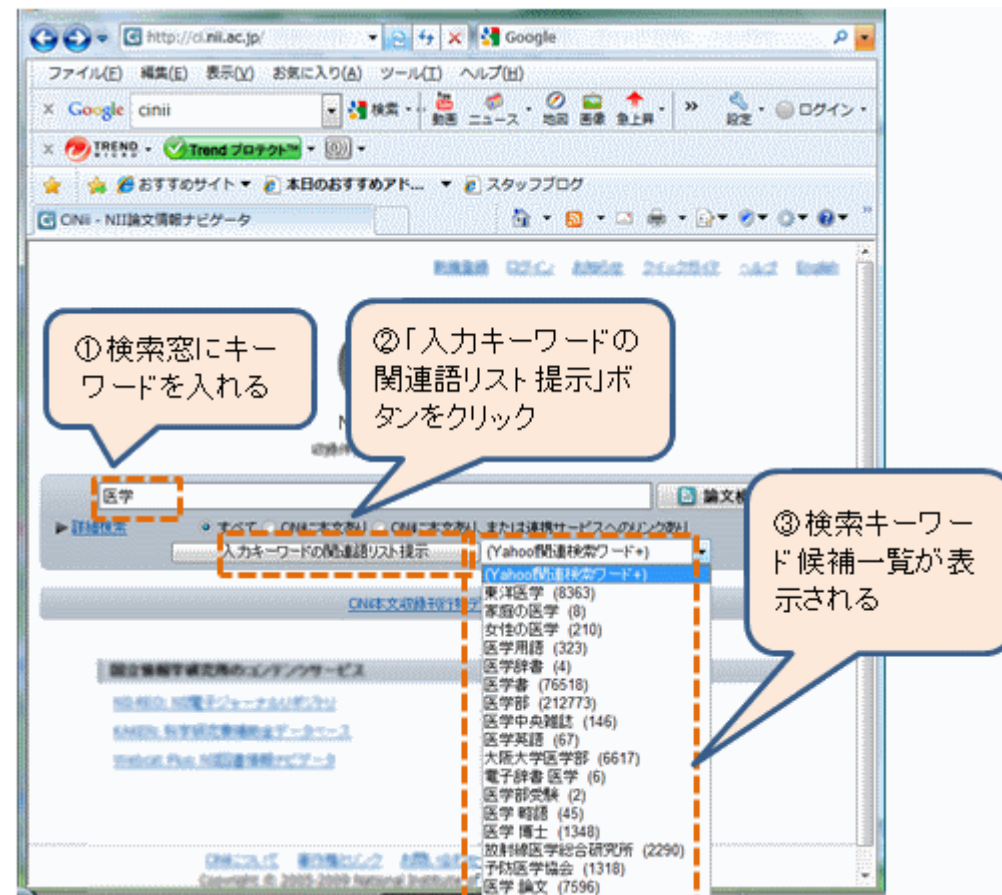
有効なキーワードの組をOPAC検索ログから得られるかが懸念される

接尾辞配列によるキーワード補完

- 単語入力中に次の語を推定させる
- OPAC検索でフルタイトル検索が多いのであれば、相性がよい可能性がある
- UT-KiwiやKiwiと同様のアルゴリズムが使えるかも（中川研究室に相談）

GETAssocではうまくいかなかったときの案

試作ツールのイメージその 1



CiNii With 関連検索ワードのようなツールにする？

試作ツールのイメージ 2

国会図書館件名標目連想検索 - Microsoft Internet Explorer

使い方:

以下に文章もしくはキーワードを入力し、「連想検索」ボタンをクリックしてください。

電子版『三十六歌選』絵巻 の公開
『三十六歌選』絵巻 の電子版を公開しました。
附属図書館と情報基盤センターでは、このたび
「三十六歌選」絵巻 の電子版を公開いたしました。

連想検索 リセット

Subject Heading	Score
三十六歌仙	1.816889e-01
絵巻	1.450949e-01
竹林	1.393866e-01
迦及変換	1.373348e-01
連歌	1.360169e-01
展示	1.357015e-01
絵入り本	1.332485e-01
科学技術情報	1.256611e-01
イントラネット	1.230102e-01

文章か
キーワード入力

入力から「連想」する国会図書館件名標目とそのスコアを表示
(Ajax使用)

国会図書館件名標目連想検索のようなツールにする？

試作ツールのイメージ 3

- Web APIにする
 - GETAssocを使えば、そのままWeb APIで公開可能



次回に続く...