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。

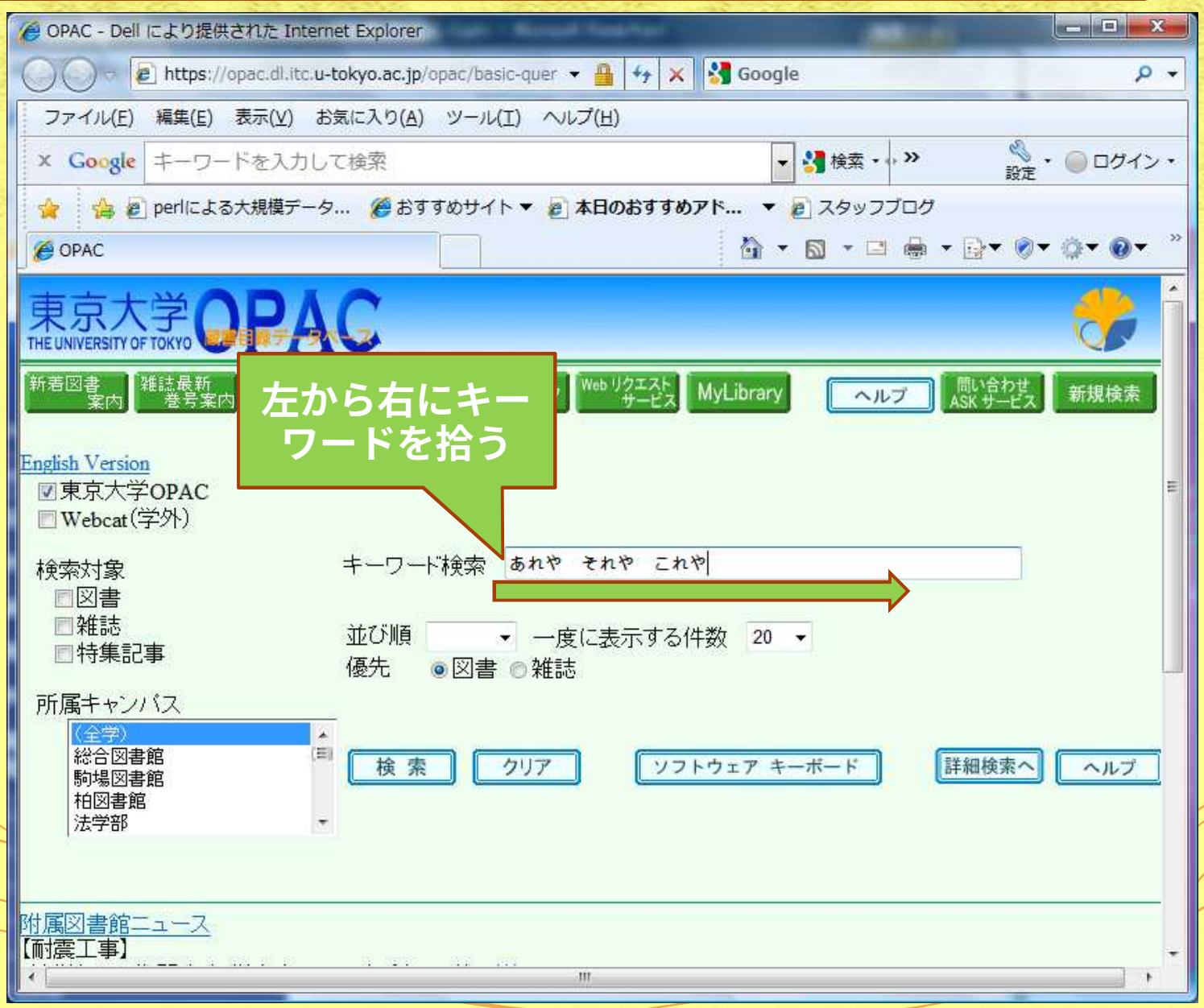
次のキーワードを予測する

検索に使われるキーワード数をＮとすると

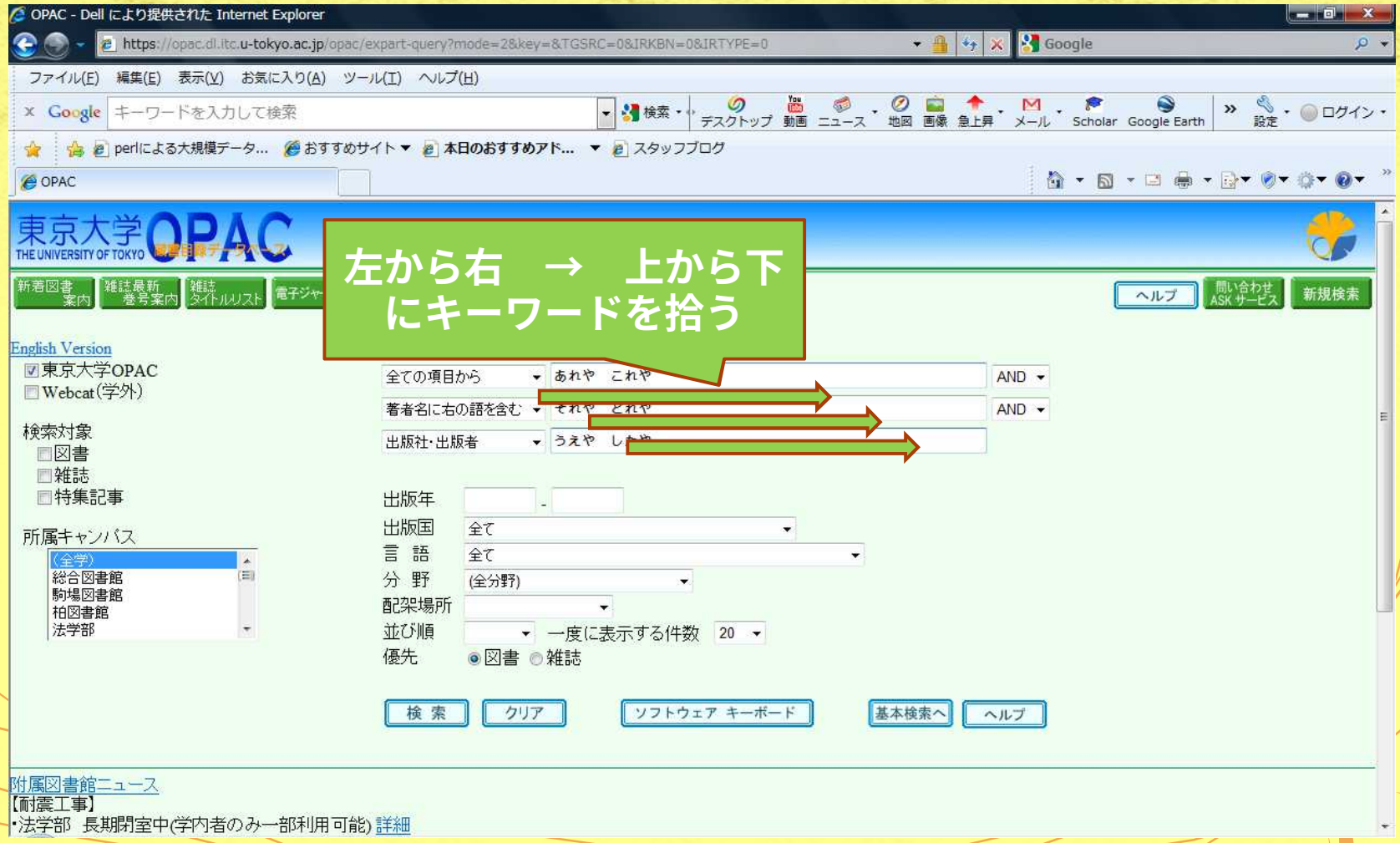
- 直後のキーワードのみ着目する
  - N-1 の組み合わせが発生
- 同時に使われるキーワードを着目する
  - N(N-1) / 2 の組み合わせが発生

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

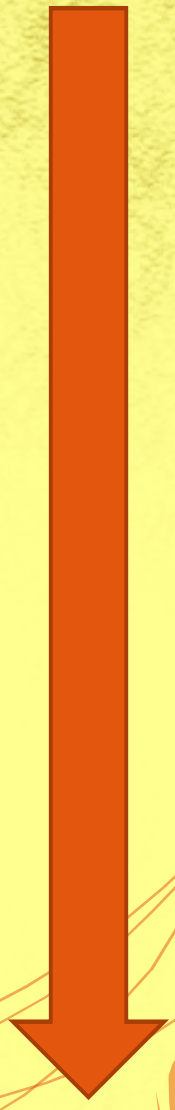


詳細検索（複数検索項目）の場合



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=xxxxxxxx
     #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（問題データ排除すみ）

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