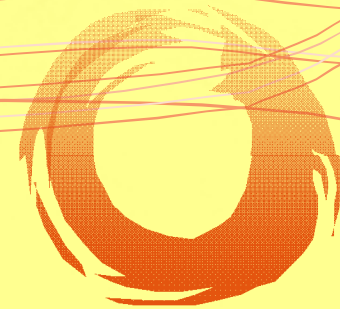


OPAC検索ログによる キーワード補完計画 (第3回)

平成22年11月6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



前回までのあらすじ

- 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
- 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。
- とりあえず、「動くもの」を作ってはみたが...

実のところ「動いてから」が本番

- 「8割の成果」であれば「動くだけ」
（2割の労力）でもいいけれど....
- 「残りの2割」の完成度で勝負するのが
図書館屋に限らずプロというものかなあ、
と思っているのです。
- ということで、泥沼の試行錯誤の世界に、
みなさまをご招待です。



チューニングの泥沼な世界

- プライバシーを守る

- キーワードから特定研究者の動向がわかる可能性もある
- 「連想検索」はそれをぼやかすことが可能
- フレーズは特定性が高いので形態素を使用する
- 使用項目（著者、フルテキストなど）も隠す

- それらしい結果を得る

- 現在のところ、個人の主観で評価するくらい
- 結果にノイズが少なければ、それなりの結果に感じるといえる → 「名詞に限定する」だけでもそれらしくなる
- GETAのパラメータを変更して試す（Webアプリ上にて変更可能）



Nグラムで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
- 前後 2 語のキーワードに着目する
 - N-1 以上、 $2(N)$ 未満
- 前後 3 語のキーワードに着目する
 - N-1以上、 $3(N)$ 未満
- 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

新規に
追加

同時に使われるキーワードだとデータ量が跳ね上がるが、前後 n 語であれば倍数で済む（現実的）

次回にすすむ

