適合的汎化に基づく情報検索システムの実験的研究

北海道大学　　吉岡真治　原口誠　大久保好章
{yoshioka,mh,yoshiaki}@ist.hokudai.ac.jp

背景

- 適合的汎化に基づく情報検索システム
  - 検索語と検索意図
  - 適合文書に含まれている語と検索語の対応関係のミスマッチ
- 部分マッチと完全マッチ
  - ユーザの検索語選択の問題

目的

- ブーリアン型検索と確率モデルの組み合わせ
  - ユーザが作成したブーリアン式の修正

確率モデルに基づく情報検索システム

- OKAPIをベースに作成
  - BM25のスコアリング式を用いて文書をランキング

$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

$$w^{(1)} = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}$$

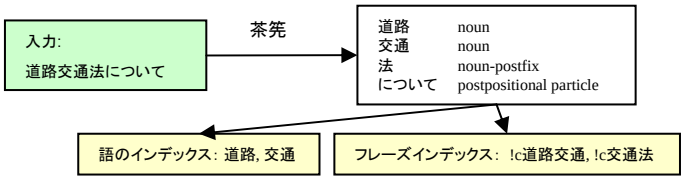
  - フレーズインデックスの重み付け
    - 辞書のエントリーによってスコアに影響

情報科学→ Word 情報, 科学 Phrase !c情報科学
情報処理→ Word 情報処理

  - フレーズインデックスの重みを修正　$qtf = c * qtf_c$

インデックスの作成

- 語とフレーズのインデックス
  - 語：茶筅を用いて主に名詞（名詞、未知語、シンボル）を選択
  - フレーズ：名詞の連接
    - 上記の名詞に加え名詞性の接辞を追加
- データベース: Generic Engine for Transposable Association (GETA)



関連文書フィードバック

- 擬似関連文書フィードバック
  - 初期検索の上位5件の文書を利用.
  - 文書中の語が少ないものは利用しない
- 検索語拡張
  - 関連文書中に存在する語を検索語に追加
  - 最大: 300語
  - Rocchioタイプのフィードバック

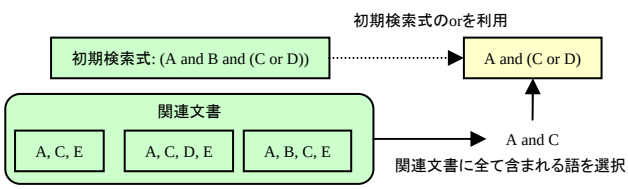
$$qtf = \alpha * qtf_0 + (1-\alpha) * \frac{\sum_{i=1}^{R} qtf_i}{R}$$

2つの検索モデルの特徴

	ユーザに対する仮定	検索される文書	検索式の可読性
確率モデル	ユーザは適切なキーワードを選択することが困難	全ての検索語を含む必要はない（部分マッチ）	多くの検索語を用いる場合、検索式と結果の関係を理解するのが困難
ブーリアンモデル	ユーザは適切なキーワードを選択可能	必要とされる検索語は必ず含む（完全マッチ）	検索式を満たしている理由を確認することが容易

ブーリアン検索式の修正（ABRIR）
Appropriate Boolean Query Reformulation for IR

- 初期に与えられたブーリアン検索式を関連文書の情報に基づき修正
  - 関連文書がブーリアン検索式を満たすように、ブーリアン式を修正



2つのIRモデルの組み合わせ

- 2つの方法
  - ブーリアン式を満たす文書を確率モデルでランキング
  - ブーリアン式を満たさない文書に対して、その非充足度合いに応じて、ペナルティを与える
    - ペナルティは、BM25の検索語の重みを利用して計算

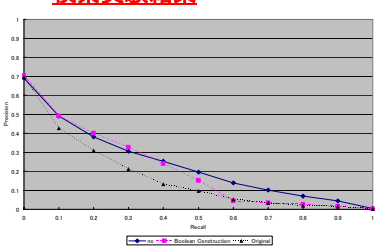
$$\beta \times w^{(1)} \times \frac{(k_3+1)qtf}{k_3+qtf} \qquad \beta: \text{parameter}$$

    - 全ての“and”要素に対して重みを計算し、総和を計算
    - “or”要素については、最も重みの高い検索語の重みを利用

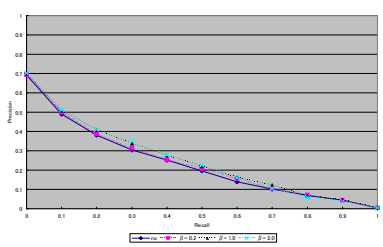
検索実験結果

NTCIR-4 Webタスク　サーベイ検索
(35課題　適合文書：3893文書中)

	ユーザ設定	ABRIR
関連文書でブーリアン式を満たすもの	1918	2380



R-P Graph for Different Boolean Query

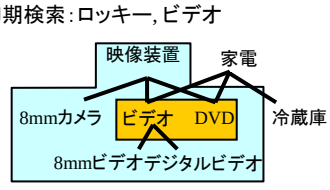


R-P Graph for Different β

適合的汎化に基づくブーリアン検索式の修正（ABRIR-AG）
ABRIR on Adaptive Generalization

- 適合的汎化に基づき、適合文書に網羅的に含まれる概念を用いてブーリアン検索式を修正

初期検索：ロッキー, ビデオ



関連文書

レンタルビデオ
映画「ロッキー」
….

DVDタイトルリスト
映画「ロッキー 」
….

Boolean式：ロッキー and (ビデオ or DVD)

まとめ

- ブーリアン型検索モデルと確率モデルの組み合わせ
  - 関連文書を用いたブーリアン検索式の修正の有効性を確認
  - 的確なブーリアン式の作成が困難な場合には、ブーリアン式の非充足度合いに応じてペナルティを与える方法が有効
- より良いブーリアン式の構築のために
  - 適合的汎化によるブーリアン検索式の修正
- 検索語の有効性に関する指標の検討
  - 検索語と関連文書の関係を分析し、検索語の有効性を検討
  - 検索語拡張の有効性の分析への利用
  - 検索課題の難しさなどを示す指標としての利用