全文検索BOF
イントロダクション

かずひこ
NaCl (ネットワーク応用通信研究所)
日本Rubyの会

パネラー紹介

- 竹迫良範さん (Namazu Project)
- 平林幹雄さん (Hyper Estraier開発者)
- 西田雄也さん (Rast開発者)
- 高尾宏治さん (Rast開発者)
- 大沢和宏さん (検索サイト運営者)
- かずひこ (コーディネータ)

自己紹介

グラフィックデザイナー (ゆうな: 右)
オープンソースプログラマー (かずひこ: 左)
という夫婦の夫の方です

→ google: かずひこ

なぜ全文検索？

- 情報が多すぎる
- 整理・分類は難しい
- テキストマイニング

何を検索するか？

- ローカルの文書
- 電子メール
- ブラウザのキャッシュ
- サイト内検索
- RSS 検索
- インターネット上の情報

どうやって見つけるか？

- 分かち書き
- N-gram
- キーワード抽出
- 概念検索
- 言語ごとの処理
- 正規表現の拡張
  - (.)[^¥1]¥1[^¥1] (一日一善)
  - 意味のパターン (慈父悲母)
  - 韻の検索、だじゃれ検索

どうやって結果を評価するか？

- TF･IDF法
- ベクトル空間
- PageRank
- 個人の好みに応じた重み付け

どう見せるか？

- サムネイル表示
- 要約の作成
- カテゴリ分類
- 検索結果同士の関係を視覚化
- OpenSearchで二次利用

いつ更新するか？

- 定期的
- オンデマンド
- 更新中の検索
- 更新時の障害対策

どこで使うか？

- 単体で検索システム
- アプリケーションに組み込み
- 速度、スケーラビリティ

などなど...

- 全文検索は奥が深い
- だからハックしがいがある!