Apache Lucene入門

株式会社 ロンウイット

RONDHUIT

WWW.RONDHUIT.COM

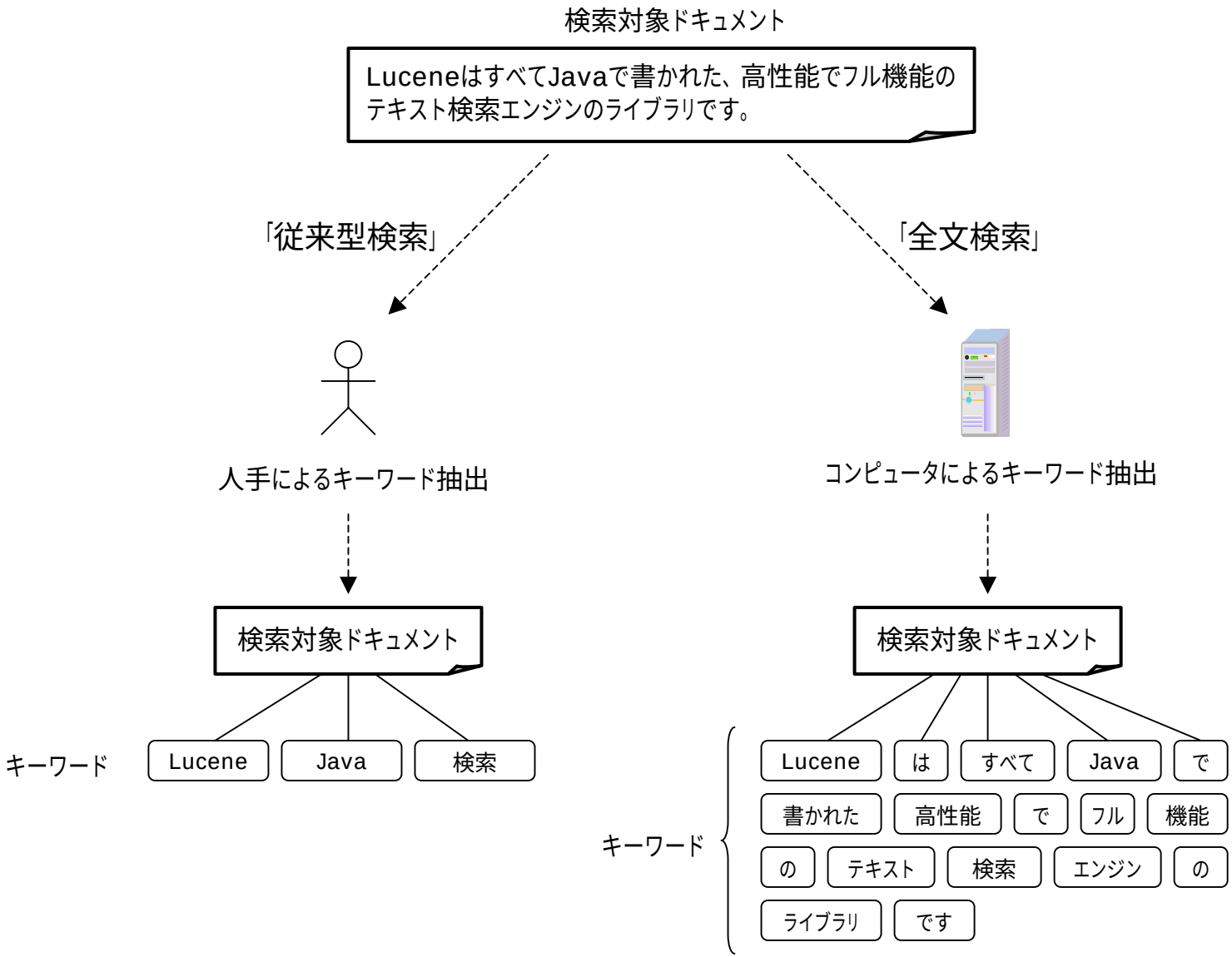
会社案内

RONDHUIT

- 株式会社 ロンウイット ／ RONDHUIT Co., Ltd.
- 〒100-0005
  東京都千代田区丸の内1-1-3 AIGビル9F
- 代表者：関口宏司
- 資本金：300万円
- 設立：2006年5月2日
- 業務内容：
  - Webサイト内検索の導入サービスの提供
  - アプリケーション開発（フレームワーク構築）
- 社名の由来：
  - 丸八通り、仏語で「丸（Rond）」「八（Huit）」

全文検索

RONDHUIT



全文検索 (cont'd)

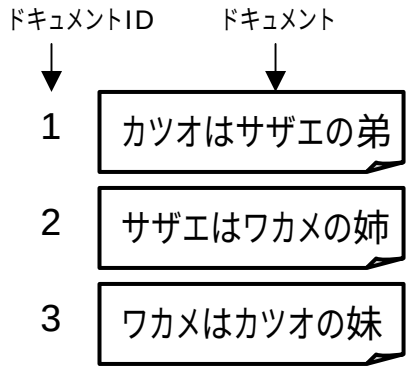
RONDHUIT

- 従来型 (非全文型) 検索
  - 統一されたキーワード抽出、品質のよいインデックス
  - コストがかかる
- 全文検索
  - 大量のドキュメントを安価に処理
  - キーワードがうまく抽出できない場合がある
  - インデックスに「雑音」が含まれる
- 全文検索の方式
  - 順次検索方式
    - インデックスを作らない
    - ドキュメントの先頭から、検索質問語の文字列と順次比較する
    - 例：UNIXコマンドのgrep
  - 転置索引方式
    - あらかじめ検索対象のドキュメントからインデックスを作成
    - 例：Lucene、Namazu、Google、Yahoo!、・・・

転置索引の作成方法

RONDHUIT

(a)



(b)

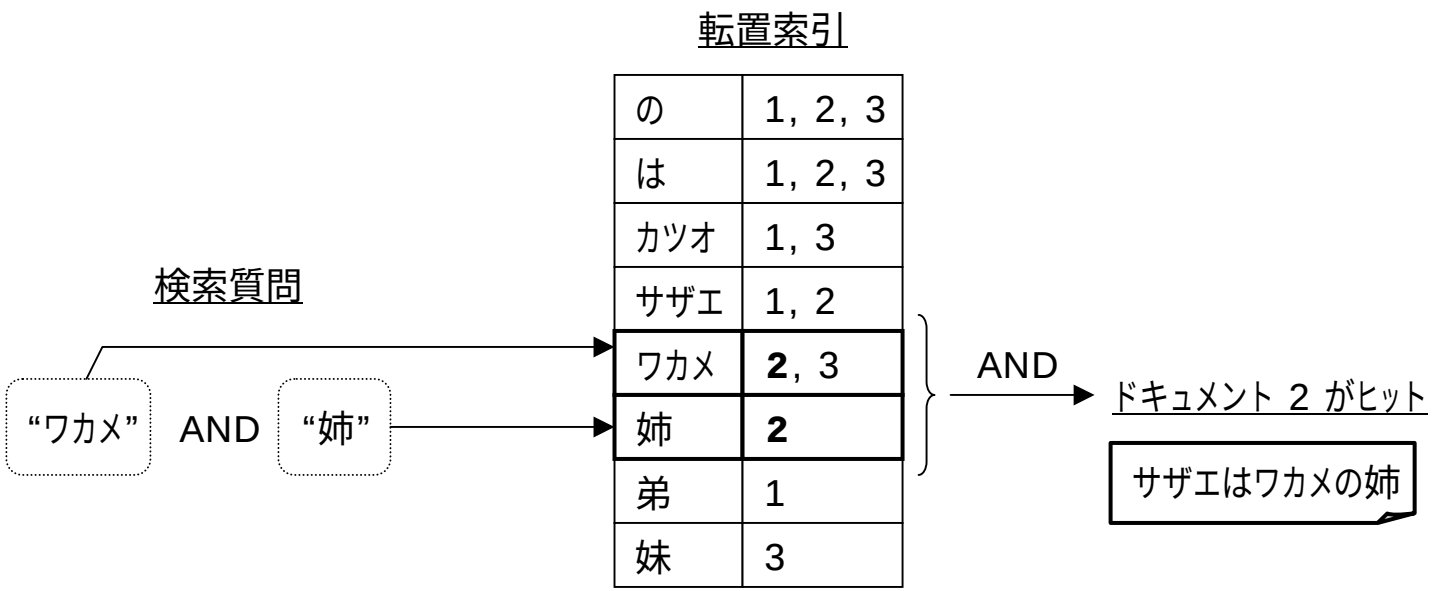
カツオ：1、は：1、サザエ：1、の：1、弟：1、
サザエ：2、は：2、ワカメ：2、の：2、姉：2、
ワカメ：3、は：3、カツオ：3、の：3、妹：3

(c)

の	1，2，3
は	1，2，3
カツオ	1，3
サザエ	1，2

ワカメ	2，3
姉	2
弟	1
妹	3

転置索引の検索方法


転置索引
検索質問
“ワカメ”
AND
“姉”
の
1，2，3
は
1，2，3
カツオ
1，3
サザエ
1，2
ワカメ
2，3
姉
2
弟
1
妹
3
AND
ドキュメント 2 がヒット
サザエはワカメの姉

全文検索の各方式の特徴

RONDHUIT

方式	長所	短所	適用例
順次検索方式	•インデックスを使わないので余計なメンテナンス作業が不要 •インデックスを使わないので「今あるドキュメント」の内容をリアルタイムに検索できる	•大量のドキュメントの検索には不向き •多数のユーザから繰り返し検索される状況下ではかなり非効率	•UNIXのgrepコマンド
転置索引方式	•大量ドキュメントを保有し、多数のユーザから繰り返し検索される状況下でも効率的に処理できる •大規模な検索に向く	•インデックスをメンテナンスしなければならないため、「今あるドキュメント」とインデックスの内容に差異が生じる場合がある •インデックスのサイズが巨大になる	•多数のユーザから利用されるアプリケーション •インターネットやイントラネットなどのコンテンツ検索機能

日本語全文検索システムの問題

RONDHUIT

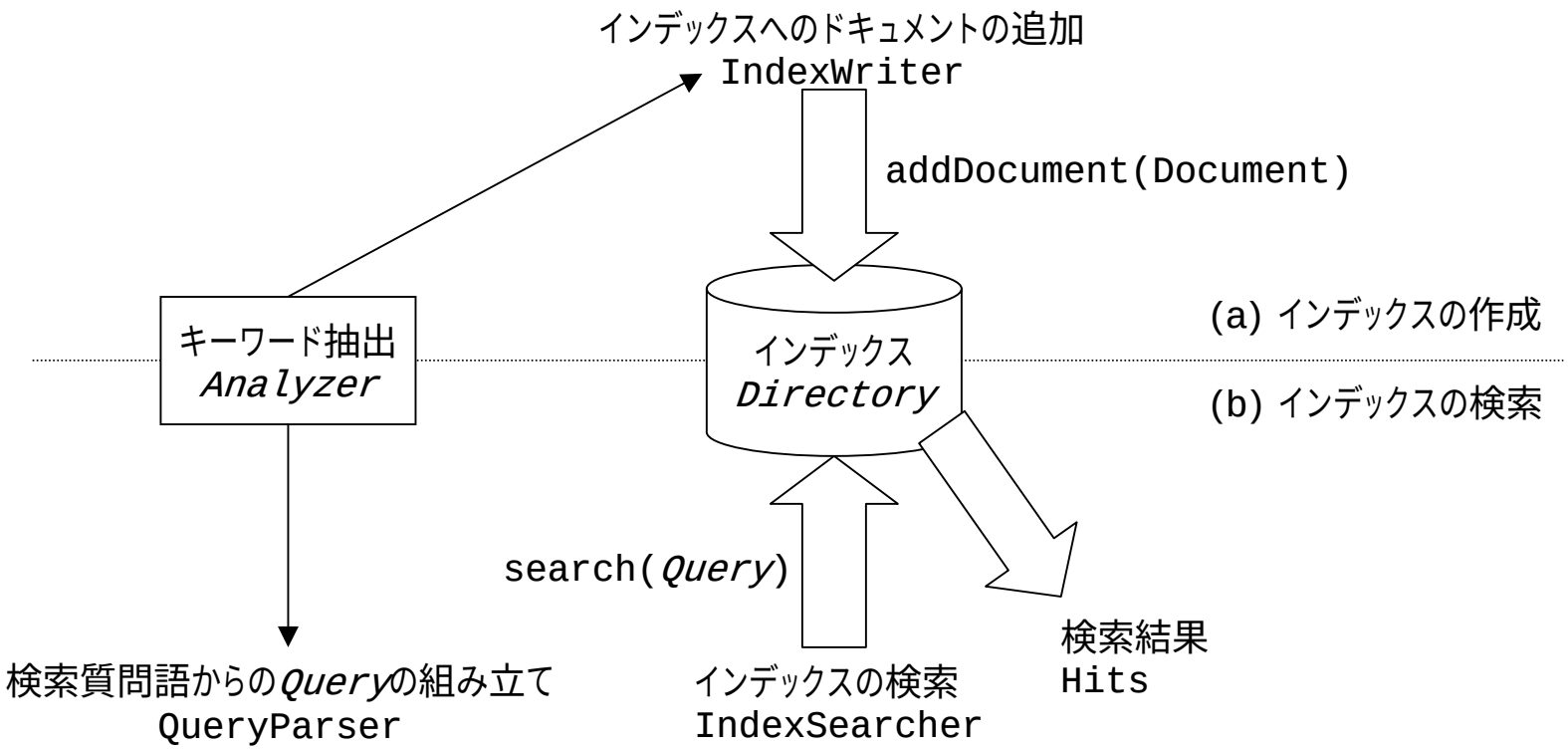
- インデックスのサイズ
- インデックスの作成にかかる時間
- ヒットしすぎる
  - 検索結果一覧の表示順序（ランキング）のスコアを計算
  - Google ⇒ PageRank
  - Lucene ⇒ tf*idf（*Similarity*抽象クラスのJavadoc参照）
- 日本語テキスト処理
  - 英語などと違って、単語を識別するのが困難
  - 形態素解析
    - 辞書を用いる方式が主流 ⇒ 流行語に弱い
    - JapaneseAnalyzer
  - N-gram
    - Nの大きさによりノイズや検索漏れ
    - CJKAnalyzer（「東京都」⇒「東京」「京都」）
  - 表記の揺れ
    - 「インタフェース」「インターフェイス」、「引っ越し」「引越し」

Apache Lucene

RONDHUIT

- 全文検索システムを構築するためのオープンソースのJavaライブラリ
- 数百万件の文書を高速に検索。大規模システム向き
- 適用例
  - 大規模ポータルサイト
  - 大規模検索エンジンポータル
  - 大規模ショッピングサイト／ショッピングモール
  - オークションサイト
  - 図書館の蔵書検索サービス
  - ナレッジマネジメントシステム
  - コールセンター／ヘルプデスクシステム
  - (テキストマイニング)
    - 問題発見・予測
    - コールセンター／ヘルプデスクシステムと組み合わせる

Lucene API – 概観



Luceneのデモ（インターネット）

RONDHUIT

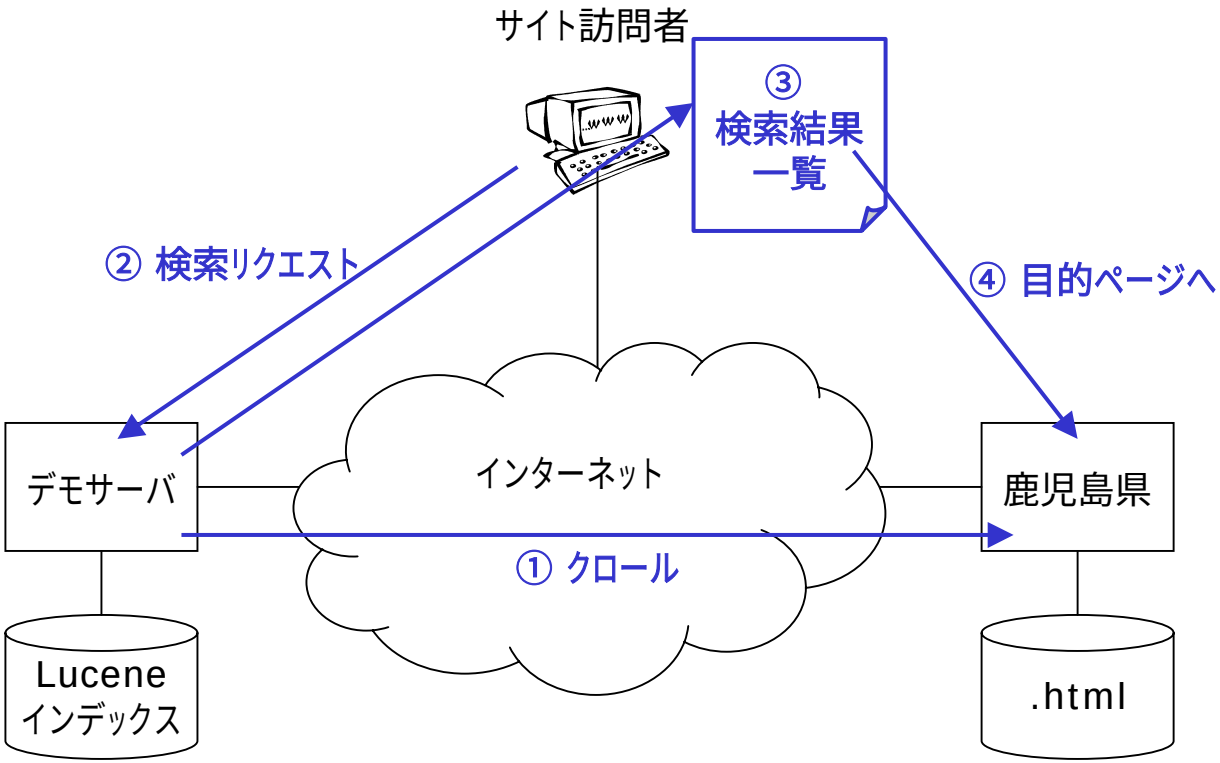
- RONDHUITサイト内検索サービス
  - 新規設置
  - Namazuリプレース
    - 検索語の要約／強調表示がうまくいかない
  - Googleリプレース
    - インデックスとオリジナルページとのズレ
    - ライバル企業の広告が表示されてしまう
  - サイト内検索1ヶ月無料お試し実施中

    http://demo.rondhuit-search.com/ssdemokag1000/

- インクリメンタルサーチ
  - Lucene+Ajax

    http://demo.rondhuit-search.com/etcajax/

サイト内検索デモのしくみ



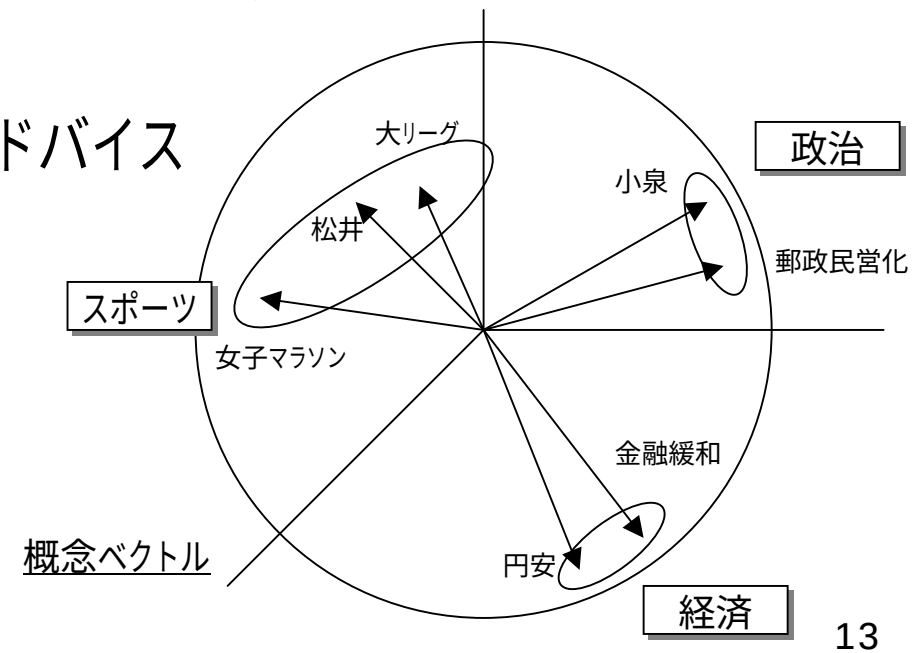
Luceneのデモ（書籍のサンプル）

RONDHUIT

1. コンテンツ「サザエさん一家」の検索
   - Luke ― インデックスブラウザ
2. 日本語Analyzerによる単語抽出
   - JapaneseAnalyzer/CJKAnalyzer
3. Webアプリケーション
   - PDF/Word/XML/HTML/RDBテキストデータの透過的検索
4. Lucene＋Ajax＝インクリメンタルサーチ！
   - 前方一致検索の利用
5. 「もしかして」キーワードアドバイス

※本デモは書籍のサンプルには含まれません。

   - あいまい検索の利用
6. テキストマイニング
   - TermFreqVector



Questions?

RONDHUIT

WWW.RONDHUIT.COM