第23回 Lucene/Solr勉強会
オンライン／初心者編 1
情報検索の基礎

7/20/2021 @kojisays
株式会社ロンウイット

はじめに

- 約2年ぶりの開催。
    - 内容が高度化、準備が大変。間隔が空くとますますハードルが上がる。コロナ禍で勉強会もオンラインが一般的に。オンラインで短めのものを、これまでよりも頻度高めで。
- 今後数回に分けて初心者向けの内容を用意。
    - 関口から初心者の方々へ。今回のみならず毎年？
- その後は事例や新機能の発表など。
    - 勉強会＝双方向で。

本日の内容

- 情報検索（全文検索）とその実現方法
- 転置インデックス
- リレーショナルデータベースとの比較
    - スコア計算
- 日本語の単語分割
- 検索エンジンの構成要素

検索エンジンと応用例

- 検索エンジン／情報検索（Information Retrieval）／全文検索（Full-text Search）とは？

  ⇒ 検索対象の文書テキストを、全文書・全フィールドに渡って横断的に検索

- 応用例

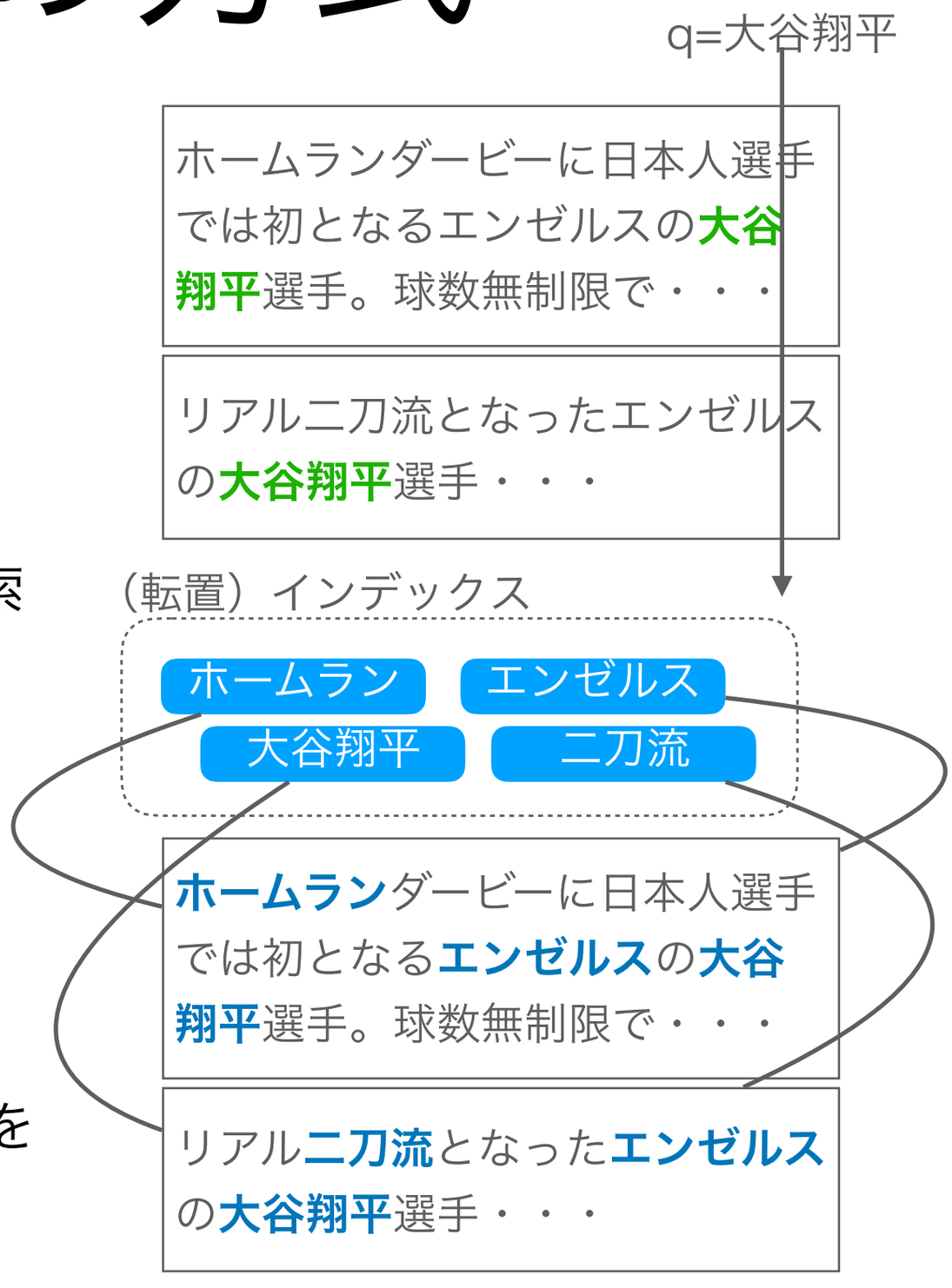
  ⇒ 図書館での図書検索、ECサイトの商品検索、企業内検索、Webサイト検索

- Googleが無料で使えるのに、なぜ企業はLucene/Solrを導入するのか。

  ⇒ 企業が保有するプライベートなデータはGoogleでは検索できない。いろいろカスタマイズしたい。便利な機能やアプリケーション特有の機能を追加したい。

全文検索の方式

- 順次検索方式
  - 文書の先頭から、クエリの文字列と順次比較する
  - LinuxのgrepコマンドやRDBのlike検索
- 転置インデックス方式
  - あらかじめ検索対象の文書からインデックスを作っておく
  - 文書に ラベル をつけ、ラベルで文書を引けるようにインデックスを作成



文書へのラベル付けの方法

検索対象文書

ホームランダービーに日本人選手では初となるエンゼルスの大谷翔平選手。

人手によるラベル抽出

検索対象文書

エンゼルス

ホームラン　大谷翔平

ダービー

コンピューターによるラベル抽出

検索対象文書

ホームラン　ダービー　に　日本人

選手　では　初　と　なる

エンゼルス　の　大谷翔平　選手

転置インデックスの作り方

手順（1）

文書ID　検索対象文書

1	カツオはサザエの弟
2	サザエはワカメの姉
3	ワカメはカツオの妹

手順（2）

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

手順（3）

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

転置インデックスの検索例1

元の文書

1	カツオはサザエの弟
2	サザエはワカメの姉
3	ワカメはカツオの妹

検索質問

q=サザエ

転置インデックス

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

→ 文書1と2を返す

転置インデックスの検索例2

元の文書

1	カツオはサザエの弟
2	サザエはワカメの姉
3	ワカメはカツオの妹

転置インデックス

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

検索質問

q=ワカメ　AND　姉

文書2を返す

RDB vs 検索エンジン

	リレーショナルデータベース	検索エンジン
特徴	関係モデル トランザクション管理 CRUD	高速な全文検索
テーブル	複数 正規化	単一 非正規化
検索	SELECT 関係論理演算	キーワード検索 （AND/OR/NOT、表記揺れ、類義語、・・・）
検索結果	集合 ソート	関連度（クエリと文書の類似度）順

※ 一般的な話であり、製品により異なります。

転置インデックスの検索例３

元の文書

1	カツオはサザエの弟
2	サザエはワカメの姉
3	ワカメはカツオの妹

転置インデックス

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

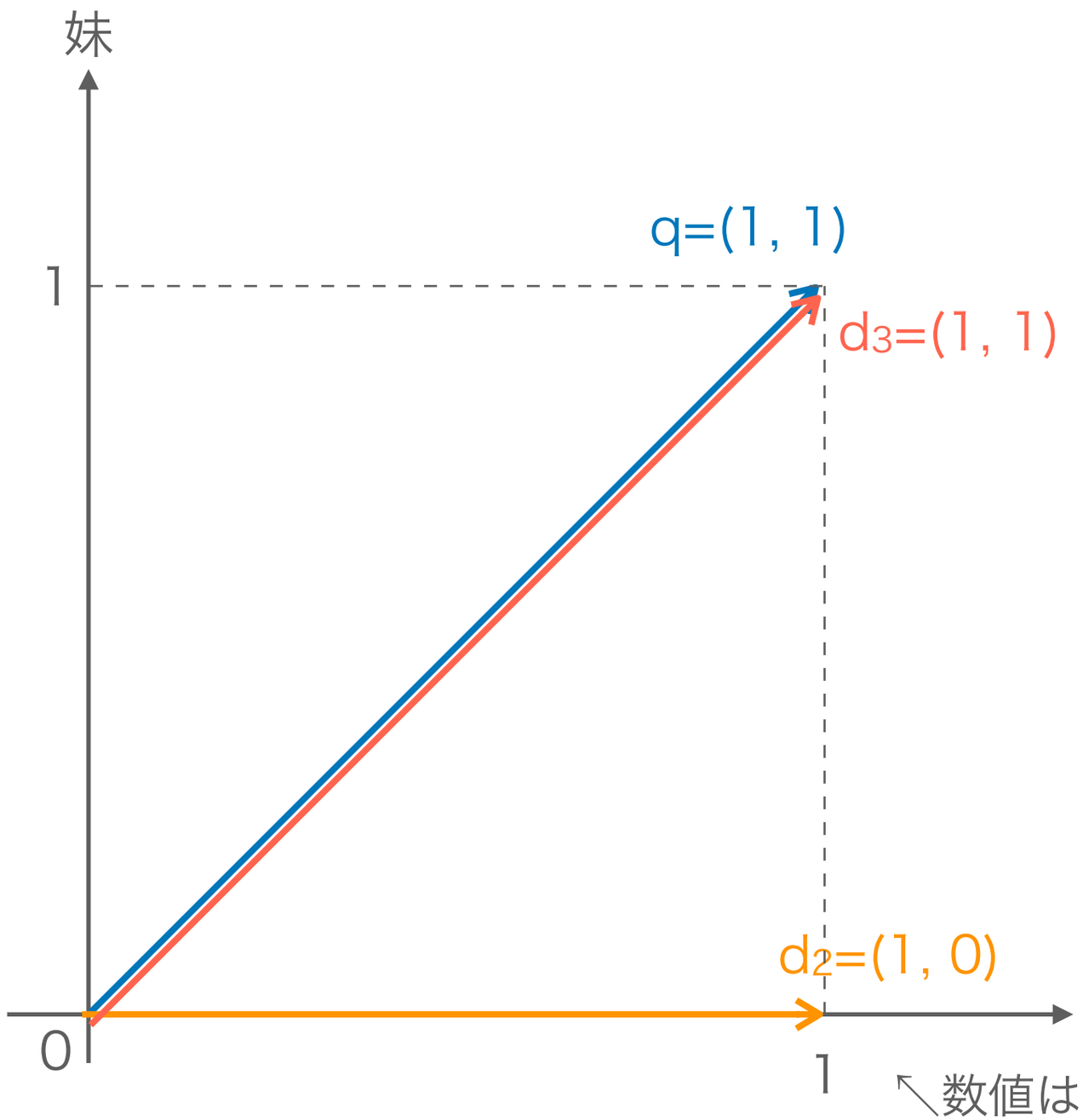
検索質問

q=ワカメ　OR　妹

文書3、2の順で返す

スコア*計算

*スコア＝クエリと文書の関連度
e.g. score(q, d)=cos(q, d)



q=ワカメ OR 妹

元の文書

2	サザエはワカメの姉
3	ワカメはカツオの妹

↖数値は単語の重要度（ここでは個数）
tf*idfなどが使われることが多い。

日本語の単語分割方法

- 形態素解析
  - 辞書に基づいた単語分割。
    - 検索結果は検索誤りが低く、検索漏れが発生しがち。

緊急事態宣言
↓
緊急 / 事態 / 宣言

- 文字N-gram
  - 機械的にN文字単位に分割。
    - 検索結果は検索漏れが低く、検索誤りが発生しがち。

緊急事態宣言　(N=2の場合)
↓
緊急 / 急事 / 事態 / 態宣 / 宣言

あいまいな単語境界

（例１）ここではきものを脱ぐ

ここ / では / きもの / を / 脱ぐ

ここ / で / はきもの / を / 脱ぐ

（例２）人間違い

人 / 間違い

人間 / 違い

形態素解析の検索漏れの例

検索質問

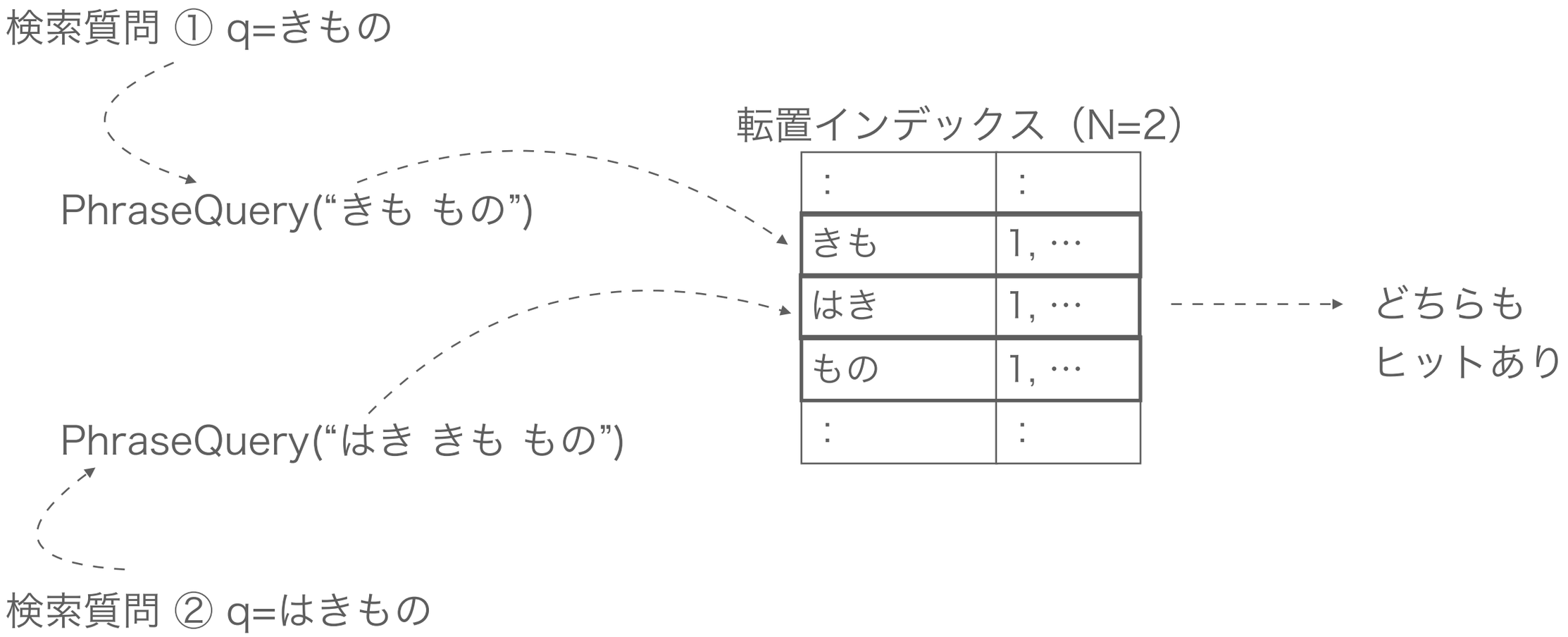
q=きもの

転置インデックス

ここ	1, …
で	1, …
はきもの	1, …
を	1, …
脱ぐ	1, …

ヒットなし

検索漏れがおきにくいN-gram



N-gramとて万全ではない

検索質問

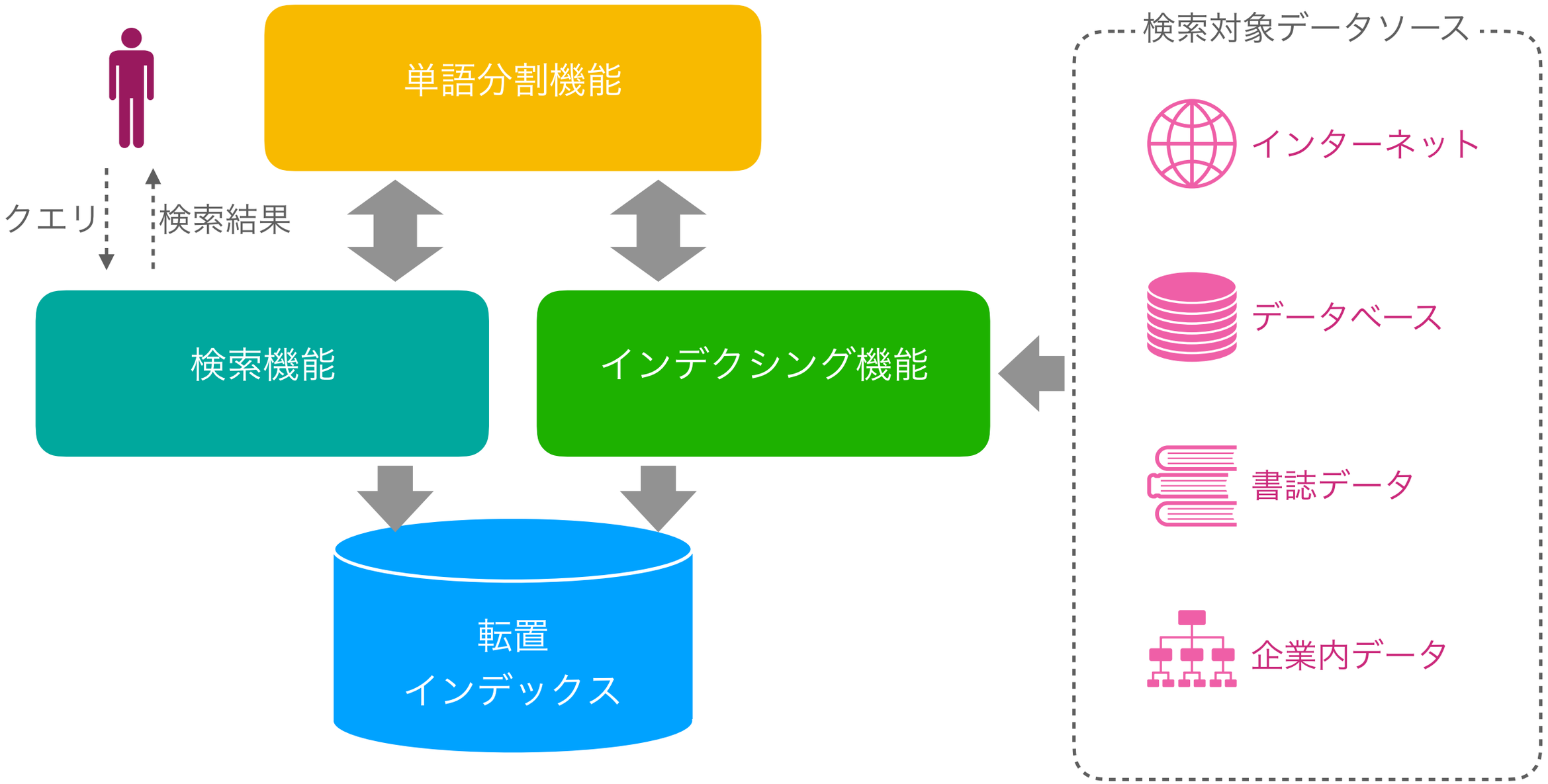
q=妹

転置インデックス（N=2）

:	:
の姉	2
の弟	1
の妹	3
:	:
カツ	1
ツオ	1
:	:

ヒットなし

検索エンジンの構成要素



まとめ

- NLPの一分野である情報検索ではさまざまな処理の単位が単語となるため、単語単位でいろいろなイメージを捉えることが重要。
- 日本語は単語境界があいまいであり、英語などとは処理の様相が最初からかなり異なる。
  - LuceneはAnalyzerの設計が大変よくできており、他のNLPツールで苦労するようなことがLucene/Solrでは起こらず、日本で広く普及した。

次回（予定）

- Apache Lucene/Solr 入門

- 日時未定

受講アンケート

- 次回以降の勉強会の参考とするため、ぜひ受講アンケートにご協力お願いします！

- 勉強会で発表できるネタをお持ちの方、本日の内容で質問のある方、次回以降で取り上げて欲しい内容のリクエストなどあれば、アンケートの自由記入欄にお書きください。

https://docs.google.com/forms/d/e/1FAIpQLSdYZ02PF2uJODrRfWVUSy643trUU_h1Sdm3VJ3g5zn6ckuKhw/viewform?usp=sf_link

開始時間（11:00）まで
ビデオと音声をオフ（ミュート）にして
そのままお待ちください。

第23回 Lucene/Solr勉強会
オンライン／初心者編 1