簡易類似文書検索手法「ふわっと関連検索」の予備的評価と分析

高久雅生
物質・材料研究機構
TAKAKU.Masao@nims.go.jp

江草由佳
国立教育政策研究所
yuka@nier.go.jp

情報処理学会・第99回情報基礎とアクセス技術研究会・第150回データベースシステム研究会・合同研究会
2010年8月4日（水）（青山学院大学 東京青山キャンパス）

提案手法「ふわっと関連検索」

- 文書類似度に基づく文書検索
  - 検索キーワードを考えなくともよい
  - 文書特徴量をそのまま利用できる
- 文書類似度の計算エンジンを積んでいない普通のキーワード検索だけの検索ツールで文書検索したい！
- 一回の検索要求に対して、キーワード検索を複数回/複数パターン実行

利用サイト公開中
http://fuwat.to

- 今回は、評価実験の結果を報告

列車運行実績データの可視化

学[電]○稲川真範（千葉工大） 正[電]富井規雄（千葉工大） 牛田貢平（東京地下鉄（株））

Visualization of Train Operation Record Data

Masanori INAGAWA† Norio TOMII† Kohei USHIDA‡

†Chiba Institute of Technology 2-17-1 Tsudanuma Narashino, Chiba 275-0016 Japan
‡Tokyo Metro Co., Ltd. 3-19-6 Higashi Ueno Taito-ku Tokyo 110-0015 Japan

Several visualization methods of train operation record data obtained from train control systems are introduced. Visualization of the data is helpful to identify overall situation of occurrence of delays and the propagation to other trains and to confirm effectiveness of revision of timetables to make them more robust. We present several visualization methods together with our experience to apply them to actual train operation record data of Tokyo Metro, the largest subway company in Tokyo.

Keywords: train operation record, train operation, visualization, robustness

1. はじめに

近年，特に大都市圏においては，ラッシュ時に列車の遅延が頻繁に発生[illegible]り，それを受け[illegible]た[1]。文献[1]においては，遅延の発生もさることながら，最初に発生した遅延（初期遅延）が他の多数の列車に伝播して，運行計画の改善作業を，「概観」「特定」「提案」「予測」「評価」の５つのフェーズに分け，特にそのうちの「概[illegible]車運行実績データの表示形

http://www.tomii.cs.it-chiba.ac.jp/kashikaJRAIL.pdf

2. 列車運行実績データの活用

列車運行実績データの可視化

検索　検索オプション

ウェブ全体から検索　日本語のページを検索

カスタマイズ設定を表示

ウェブ　検索ツールを表示　**列車運行実績データの可視化** の検索結果 約 **2,750** 件中 **1 - 10** 件目 **(0.18** 秒**)**

[PDF] **列車運行実績データの可視化**

ファイルタイプ: PDF/Adobe Acrobat - クイックビュー

列車運行実績データの可視化を検討するに際しては，ユ．ーザにとまどいを与えないように，これまでに使用されて .... 可視化手法が有効に用いられうることを確認することがで．きた。6 おわりに．**列車運行実績データの可視化**手法を提案し，それらが列 ...

www.tomii.cs.it-chiba.ac.jp/kashikaJRAIL.pdf -

Twitter / hidetan: **列車運行実績データの可視化** http://www.t ...

Check out m.twitter.com! Skip to navigation; Skip to sign in form. Loader. Twitter.com · Login · Join Twitter! **列車運行実績データの可視化** http://www.tomii.cs.it-chiba.ac.jp/kashikaJRAIL.pdf #nhk about 2 hours ago from web ...

twitter.com/hidetan/status/8542455794 - キャッシュ -

Twitter / shiyu: 牛田さん共著論文「**列車運行実績データの可視化** ...

牛田さん共著論文「**列車運行実績データの可視化**」可視化の手法がクロマティックダイヤ図 http://www.tomii.cs.it-chiba.ac.jp/kashikaJRAIL.pdf #nhk.

twitter.com/shiyu/status/8542431922 - キャッシュ -

TweetBuzz - kashika

番組で出ていた牛田さん
車運行実績データの可
視化」可視化の手法が

tweetbuzz.jp/entry/9262701/www.tomii.cs.../kashikaJRAIL.pdf - キャッシュ -

スポンサーリンク

データ可視化で競争に勝つ
費用対効果・俊敏性・効率の向上!
ビジネスインテリジェンスで経営改革
spotfire.tibco.jp

富士通**可視化** 事例
IT投資による改善効果をご紹介。
木村佳乃がコスト削減をナビゲート
jad.fujitsu.com

DICOMデータなどを**可視化**
数値だけでは理解、把握できない
3次元**データ**を視覚**化**するソフトウェア
www.lightstone.co.jp/slicerdicer

広告掲載はこちら »

http://www.google.co.jp/search?q=%E5%88%97%E8%BB%8A%E9%81%8B%E8%A1%8C%E5%AE%9F%E7%B8%BE%E3%83%87%E3%83%BC%E3%82%BF%E3%81%AE%E5%8F%AF%E8%A6%96%E5%8C%96

ふわっと CiNii 関連検索

文章から検索 | ウェブページから検索

http://www.tomii.cs.it-chiba.ac.jp/kashikaJRAIL.pdf

論文検索

列車 運行 遅延 ダイヤ Ci 295件中 1- 20件目・・ 検索にかかった時間2.85秒

1. 第93回運輸政策コロキウム 都市鉄道の運行ダイヤ過密化に伴う列車遅延の波及に関する研究
   仮屋崎,圭司; 岩倉,成志; 森地,茂
   運輸政策研究, 2009/Win.
2. スイス連邦鉄道における接続を重視した新しい運行管理手法：戦略的施策から実際の運営の場に至るまでの余裕時分の活用手法(<特集>鉄道のスケジューリング問題)
   Laube,Felix; Luthi,Marco; 富井,規雄
   オペレーションズ・リサーチ：経営の科学, 2008-08-01
3. スプレッドシートを用いた協調推論型知識調整方式：列車運行予測の高精度化へ向けて
   江口,俊宏; 鶴田,節夫
   全国大会講演論文集, 1991-02-25
4. 列車運行予測のための協調推論型知識調整方式
   鶴田,節夫; 江口,俊宏; 松本,邦顕
   全国大会講演論文集, 1990-09-04
5. 論文部会報告 第46回鉄道サイバネ・シンポジウムの論文を査読して(列車運行システム、新しい交通システム、情報伝送・処理システム)
   平粟,滋人
   サイバネティクス, 2010/1
6. 近鉄 名古屋地区列車
   高道,和博
   鉄道と電気技術, 2010/1

http://fuwat.to/cinii?url=http%3A%2F%2Fwww.tomii.cs.it-chiba.ac.jp%2FkashikaJRAIL.pdf

「ふわっとCiNii関連検索」

最終更新日：2010年02月23日
(更新履歴)

ふわっと CiNii 関連検索

文章から検索　ウェブページから検索

論文検索

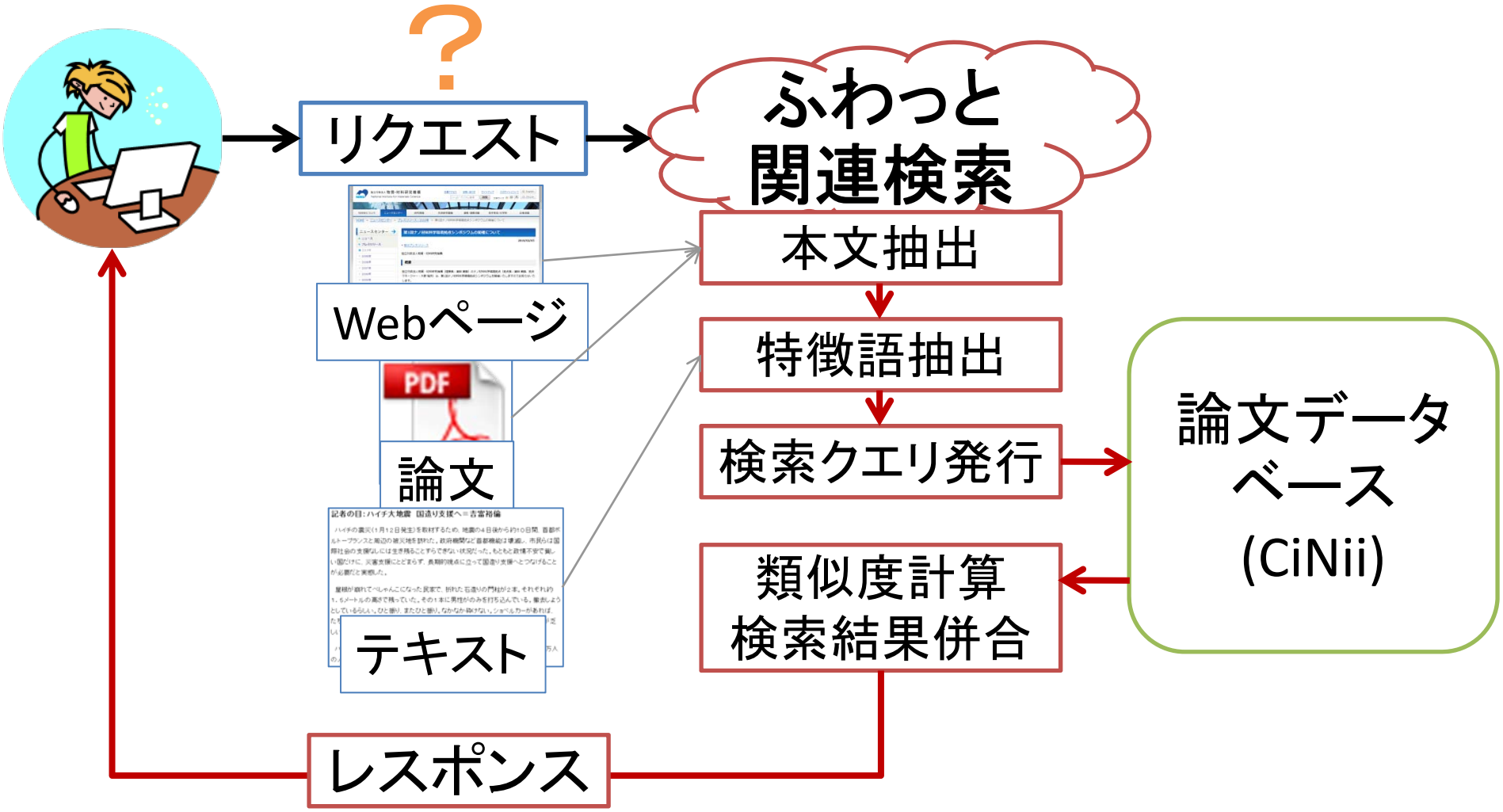
入力したテキストまたはウェブページに関連した論文をCiNiiで検索します。長いテキストやURLで指定したページからでも関連キーワードを自動的に抜き出して論文検索できるのが特徴です。

例: 朝日新聞社説 (元記事(asahi.com)), 日本経済新聞社説 (元記事(nikkei.co.jp)), 毎日新聞「記者の目」(元記事(mainichi.jp))

高久雅生 (Masao Takaku)
http://masao.jpn.org/, tmasao@acm.org

http://fuwat.to/cinii

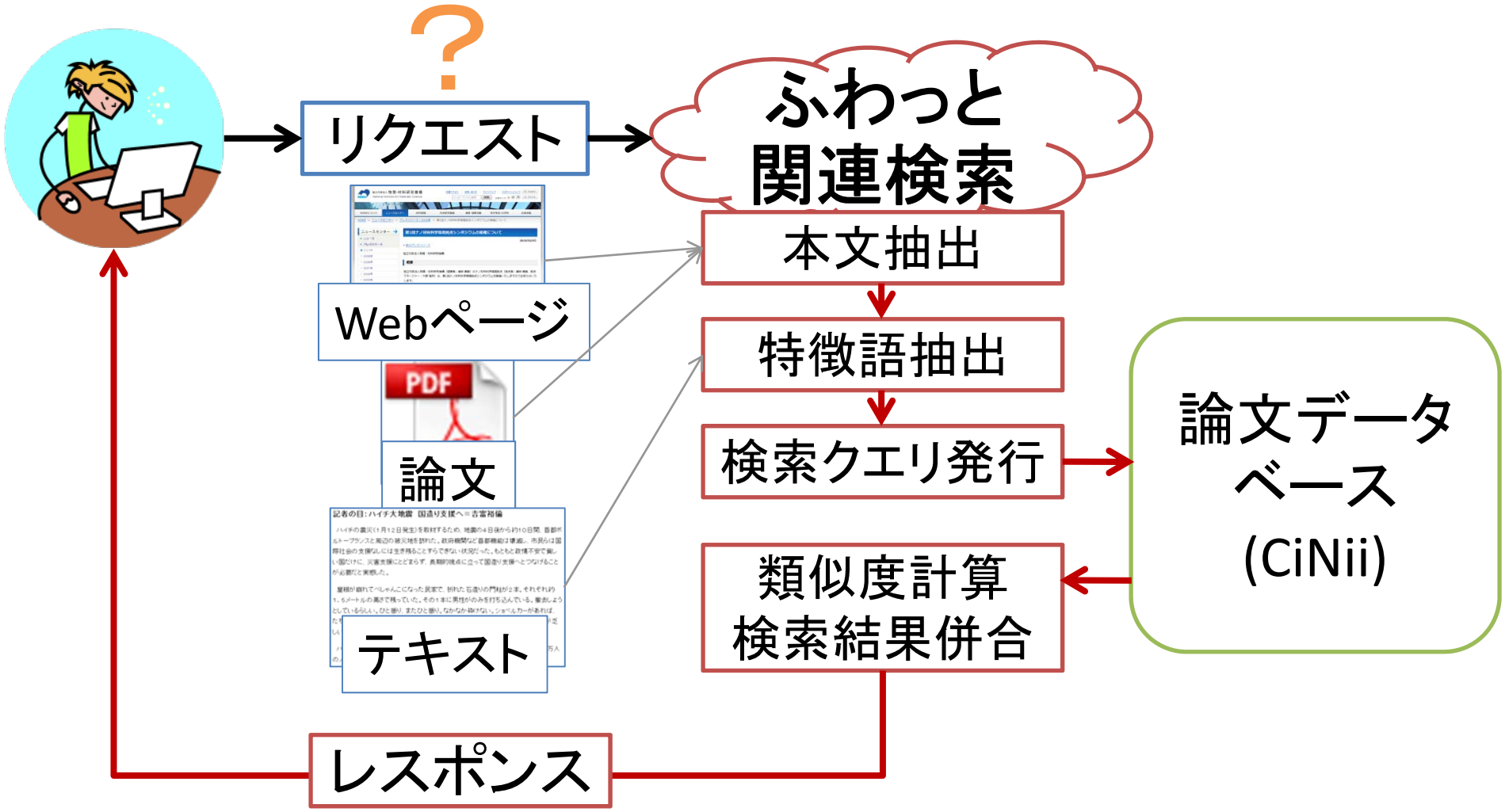
「ふわっと関連検索」システムの概要



実装: 本文抽出

- 類似検索の対象文書：Webページ（HTML）の場合
  - ヘッダ・フッタやナビゲーション部分など、特徴語抽出にノイズとなる部分を除去し、対象の本文テキストを取得
- PDFフルテキストの場合
  - 本文中のテキスト抽出
- 既存のツールを利用
  - Webページ (HTML): contentextract.rb
  - PDF: pdftotext (xpdf)

「ふわっと関連検索」システムの概要



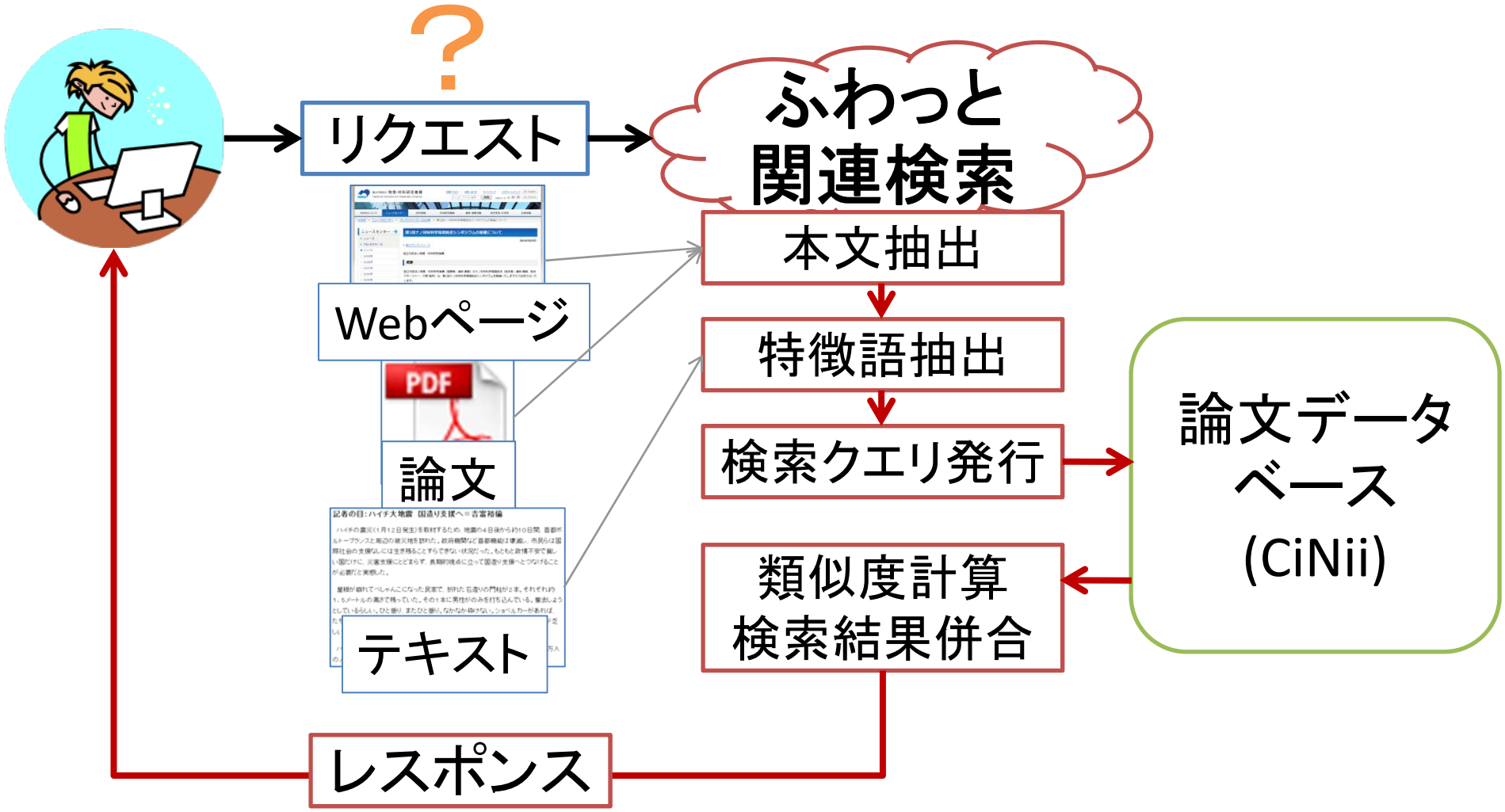
実装: 特徴語抽出

- テキスト中での出現語*w*に対して各語の重みを計算
- 2段階での簡易TF-IDFを採用
  - テキスト内での出現回数: TF(w)
  - テキスト中での単語の生起コスト: Cost(w) ※
  - データベース文書群中でのヒット文書数: DF(w)

1. テキスト中での特徴語上位n件 （Logcost）
   - weight(w) = TF(w)･LogCost(w)
2. データベース問合せ結果を使って特徴語スコアを調整 （IDF）
   - weight’(w) = weight(w) / log(DF)

※生起コスト Cost(w) : 形態素解析ツールMeCabのコスト値

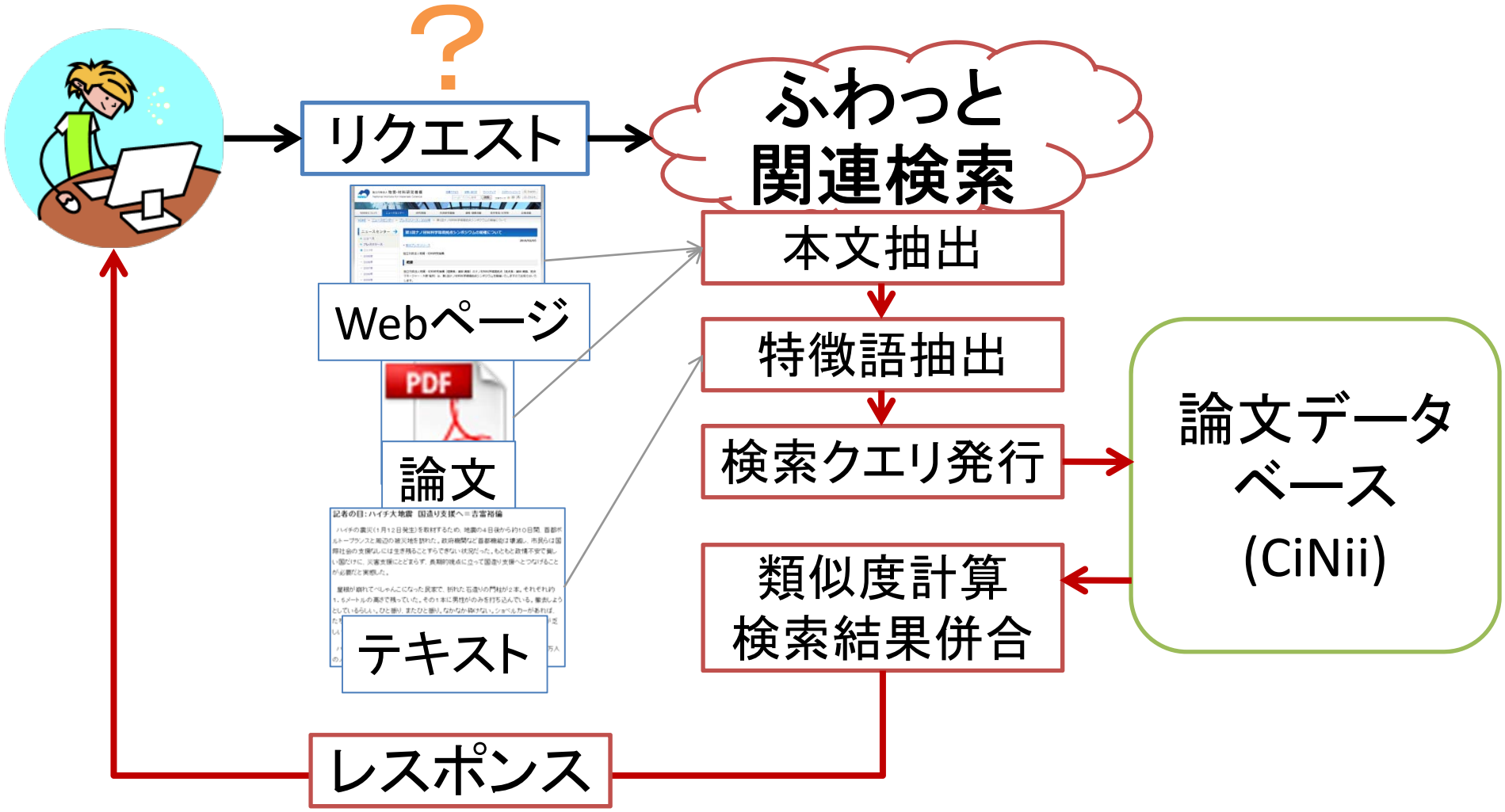
「ふわっと関連検索」システムの概要



実装: クエリ発行と検索結果併合

- 特徴語上位 n 件をクエリ発行
  - n回のクエリ発行
    - ゼロヒットになるものを除外
  - $(w_1,\ldots,w_n)$ の各語をAND結合し、クエリ発行 （AND）
    - ゼロヒットになってしまう場合は、重みの低い特徴語から除外して、$(w_1,\ldots,w_{n-1})$ の各語として、クエリ発行。
      - 以下、指定した表示件数が得られるまで続行。
  - ※「CiNii関連検索」においては n = 10 を採用
  - ※CiNiiが提供するOpenSearch APIを通じて検索
- 得られた検索結果を重複除去し、最終的な検索結果とする。

「ふわっと関連検索」システムの概要



評価実験

- 対象DB：CiNii
  - （CiNii Opensearch APIを経由して検索実行）
- 比較アルゴリズム：
  - 特徴語抽出、検索クエリ発行、検索結果併合、疑似適合フィードバックなどの派生手法を比較
- クエリ：新聞記事34件
  - 朝日新聞、産経新聞: Webサイト上のニュース記事
  - 2010-03-19 ～ 2010-05-09
- 検索タスク
  - レポート執筆等を想定した主題適合性
- 適合判定
  - 多段階適合度：A（適合）、B（部分適合）、C（不適合）
  - 判定者：1名

検索クエリ文書

文書ID	掲載紙	タイトル	掲載日
NP017	産経	通過駅？熊本、九州新幹線に不安　企業アンケ「プラス」は6割	2010年4月5日
NP018	産経	消費税論争勃発　その背景は？　民主執行部のバラマキ路線を牽制　「ポスト鳩山」の思惑も	2010年4月13日
NP021	産経	山崎さんが琴を演奏　国際宇宙ステーション	2010年4月12日
NP028	産経	「軍事対応」は1割以下　韓国艦沈没で世論調査	2010年5月7日
NP029	朝日	COP15政治合意「尊重」で一致　経済国フォーラム	2010年4月20日
NP032	朝日	光が織り成す夜の芸術　徳島でLEDフェスティバル	2010年4月18日
NP033	朝日	沖縄知事、県内移設反対の県民大会出席へ　普天間問題	2010年4月23日
NP035	朝日	巨大氷山衝突、南極の氷河もぎ取る　NASA撮影	2010年3月19日
NP049	朝日	バイエルンが先勝　リヨンに1－0　欧州CL準決勝	2010年4月22日

表1: 検索対象文書の新聞記事（抜粋）

比較対象アルゴリズム

- 特徴語抽出:
  - TF: weight(w) = TF(w)
  - **LogCost**: weight(w) = ∑LogCost(w)
  - IDF: weight'(w) = weight(w) / IDF(w)
- 検索クエリ発行: ※ n = 10 で実験
  - **AND**:
    - 特徴語上位n件をANDで連結し、当該ヒット件数が得られるまで、件数を減らしながらクエリを投げる。
  - Comb:
    - 特徴語の上位n件から、3語づつのすべての組み合わせを求め、それらのAND条件すべてを検索クエリとして発行
- 検索結果の併合:
  - Rerank:
    - 検索結果として得られた文書の「タイトル」「抄録」「掲載誌名」から特徴語ベクトルを抽出し、クエリ文書との特徴語ベクトルとのコサイン類似度順に並べ替える。
- 疑似適合フィードバック: ※ k=20, α = 0.25, 0.50, 0.75, 1.0, 2.0 で実験
  - PRF(α):
    - Rerank手法で得られた検索結果文書の上位k件を取り出し、特徴語ベクトルを生成し、正規化して、元の特徴語ベクトルに加えたものを検索クエリとして再検索を実行する。

      $weight'(w) = (1-\alpha)\, weight(w) + \alpha \cdot weight_d(w)$

実験結果と考察 (1/4)
- AND条件の場合 -

比較アルゴリズム	Prec@10
AND + TF	0.0794
AND + TF + IDF	0.0941
AND + LogCost	0.1059
AND + LogCost + IDF	0.0765
AND + Rerank + TF	0.1206
AND + Rerank + TF + IDF	**0.1382**
AND + Rerank + LogCost	0.1000
AND + Rerank + LogCost + IDF	0.1324

- TF ⇔ LogCost: 差はほとんど無い。
- IDF: 効果は限定的
- Rerank: やや効果あり。

実験結果と考察 (2/4)
- Comb条件の場合 -

比較アルゴリズム	Prec@10
AND + Rerank + TF + IDF	0.1382
Comb + Rerank + TF	0.2176
Comb + Rerank + TF + IDF	0.2324
Comb + Rerank + LogCost + IDF	0.2324
Comb + Rerank + LogCost + IDF	**0.2500**

- TF ⇔ LogCost: 差はほとんど無い。
- IDF: 効果は限定的
- Rerank: やや効果あり。
- Comb⇔AND: 強い効果あり。

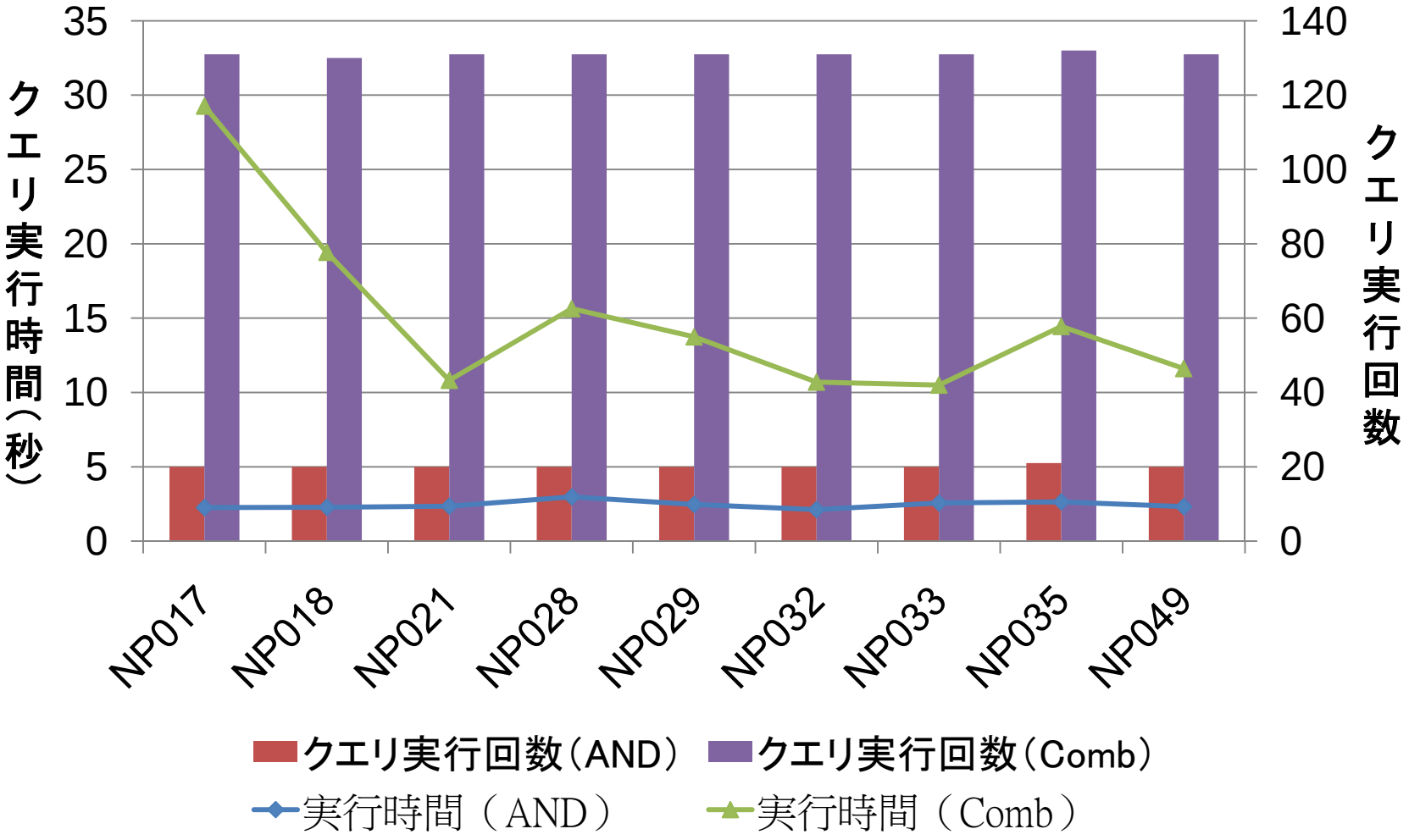
実験結果と考察 (3/4)
- PRF条件の場合 -

比較アルゴリズム	Prec@10
AND + Rerank + TF + IDF	0.1382
Comb + Rerank + LogCost + IDF	**0.2500**
PRF(25) + Comb + Rerank + LogCost + IDF	0.2353
PRF(50) + Comb + Rerank + LogCost + IDF	0.1971
PRF(75) + Comb + Rerank + LogCost + IDF	0.1059

- TF ⇔ LogCost: 差はほとんど無い
- IDF: 効果は限定的。
- Rerank: やや効果あり。
- Comb⇔AND: 強い効果あり
- Comb⇔PRF: α値によらず、ネガティブに作用

実験結果と考察 (4/4)

- 実行応答速度の面から -



まとめ

- 文書類似度に基づく文書検索:「ふわっと関連検索」
  - 検索キーワードを考えなくともよい
  - 文書特徴量をそのまま利用できる
  - キーワード検索しかサポートしてない検索ツールを対象とした類似文書検索ができるようになる
  - 一回の検索要求に対して、キーワード検索を複数回/複数パターン実行
- 新聞記事を対象とした評価実験
  - 適合度順ランキング10位時点での精度: 平均 0.25
- 今後の課題
  - 文書ジャンルの違いの影響
    - 論文・書籍、ウェブログ
  - 対象DBの違いによる影響
    - 書誌目録(NDL-OPAC、WorldCat)
    - レファレンス協同DB

利用サイト公開中
http://fuwat.to
(対応検索サービス募集中)