筑波大学
University of Tsukuba

大規模科学データに対する データ工学的アプローチ

天笠俊之，川島英之，北川博之
筑波大学大学院システム情報工学研究科
計算科学研究センター

筑波大学 University of Tsukuba

自己紹介

- 名前
  - 天笠俊之
- 所属
  - 筑波大学大学院システム情報工学研究科
    - 北川データ工学研究室
  - 筑波大学計算科学研究センター
- 研究テーマ
  - データ工学
  - データベース，データベースシステム
  - XMLデータ，XMLデータベース
  - 科学分野におけるデータ工学の応用

筑波大学
University of Tsukuba

筑波大学計算科学研究センターのあらまし

筑波大学計算科学研究センター研究部門

筑波大学 University of Tsukuba

サイエンス

- 素粒子物理研究部門
- 宇宙・原子核物理研究部門
- 量子物性研究部門
- 生命科学研究部門
- 地球環境研究部門

CS

- 高性能計算システム研究部門
- 計算情報学研究部門
  - データ基盤分野
  - 計算メディア分野

筑波大学 University of Tsukuba

大規模科学データへの取り組み

- データ工学（データベース，データマイニング）の観点から，科学データの管理・運用，知識発見
- 主なコラボレーション
  - 地球環境研究部門
    - 大規模気象データベースGPV/JMA
    - 気象データ分析
  - 素粒子物理研究部門
    - ILDG / JLDG
  - 産業技術総合研究所
    - GEOGridプロジェクト
      - 大規模異種衛星センサデータ
  - 国土交通省国土技術政策総合研究所
    - 河川シミュレーション

筑波大学
University of Tsukuba

本日の話題

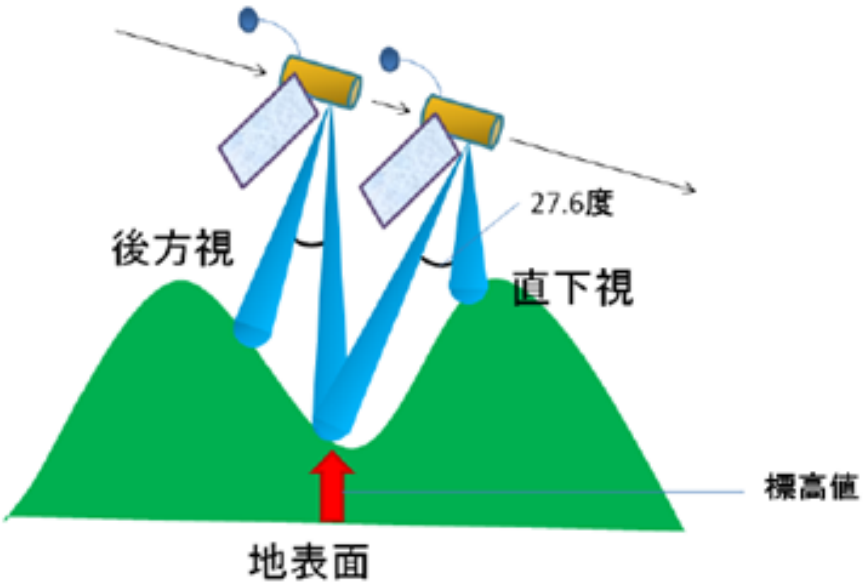
- ASTER衛星画像からのイベント検出
- 気象文献データベースからの専門語抽出

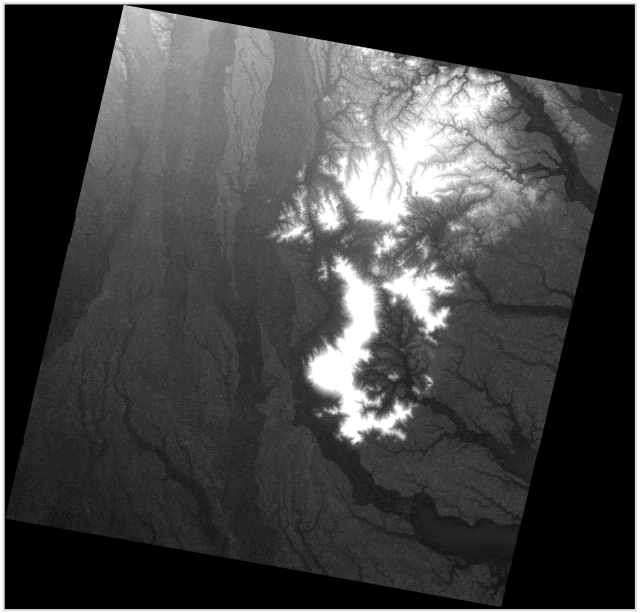
時系列ASTER衛星画像と Webコンテンツを用いた 建造物生成イベントの検出

高木 崇, 川島英之, 天笠俊之, 北川博之

背景と目的

- GEO Gridのアーカイブする衛星観測データ
  - ASTER DEM (Digital Elevation Model)
    - 後方視センサと直下視センサからステレオ画像を生成
    - 水平空間解像度：15m, 垂直空間解像度：1m
    - 16ビット，1チャンネル
    - 観測期間：2000年－2009年





背景と目的

- 衛星画像の利用
  - 時系列衛星画像から各種事象の検出
    - 地震による被害地検出[Ⅰ]
    - 火山モニタリング[Ⅱ]
    - 新築建物の検出[Ⅲ]

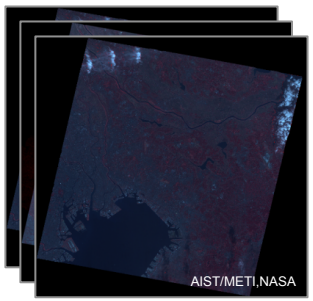
時系列衛星画像から検出される事象：
イベント

標高値変化からの建造物生成イベントの検出

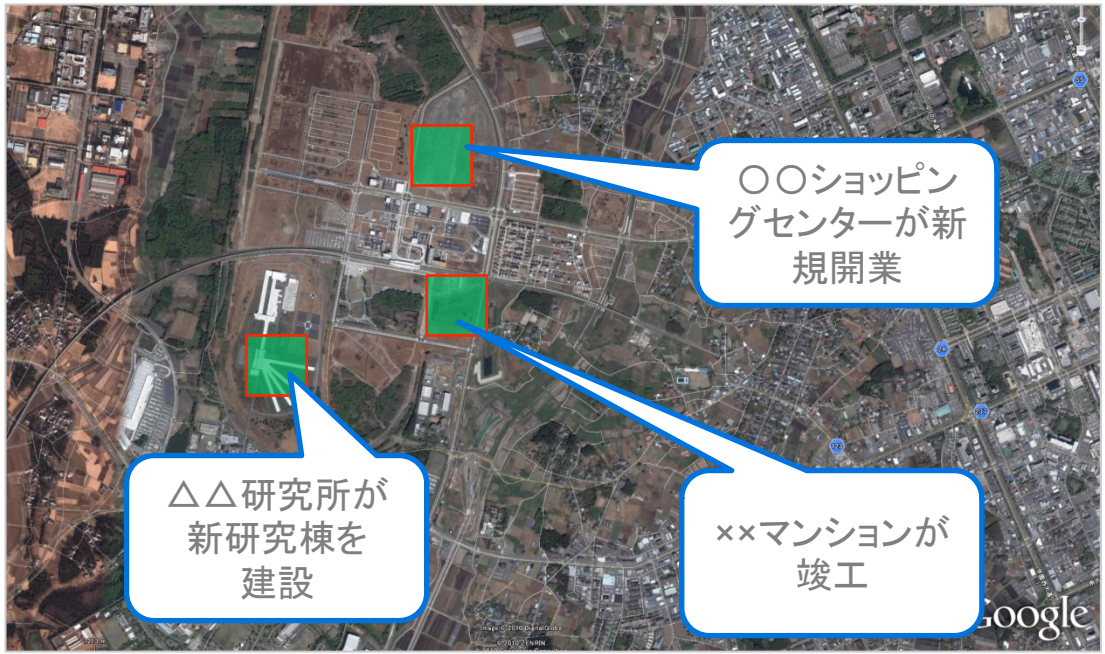
I. 松岡，堀江，大倉：「人工衛星SARによる被害地域検出手法の2004年新潟県中越沖地震への適用と高度化」
II. 浦井，福井：「ASTERによるグローバル火山観測計画と火山画像データベース」
III. 三浦，翠川：「建物GISデータの更新を目的とした高分解能衛星画像からの建物の自動検出手法」

背景と目的

- 衛星画像を用いたイベント関連Webコンテンツの提示
  - 衛星画像から建造物生成イベントを検出
  - 建造物生成イベントに関連するWebコンテンツの収集
  - キーワードベースでは検索困難であるイベント情報の提示
  - 衛星画像解析により検出されたイベントよりも詳細な情報の提示

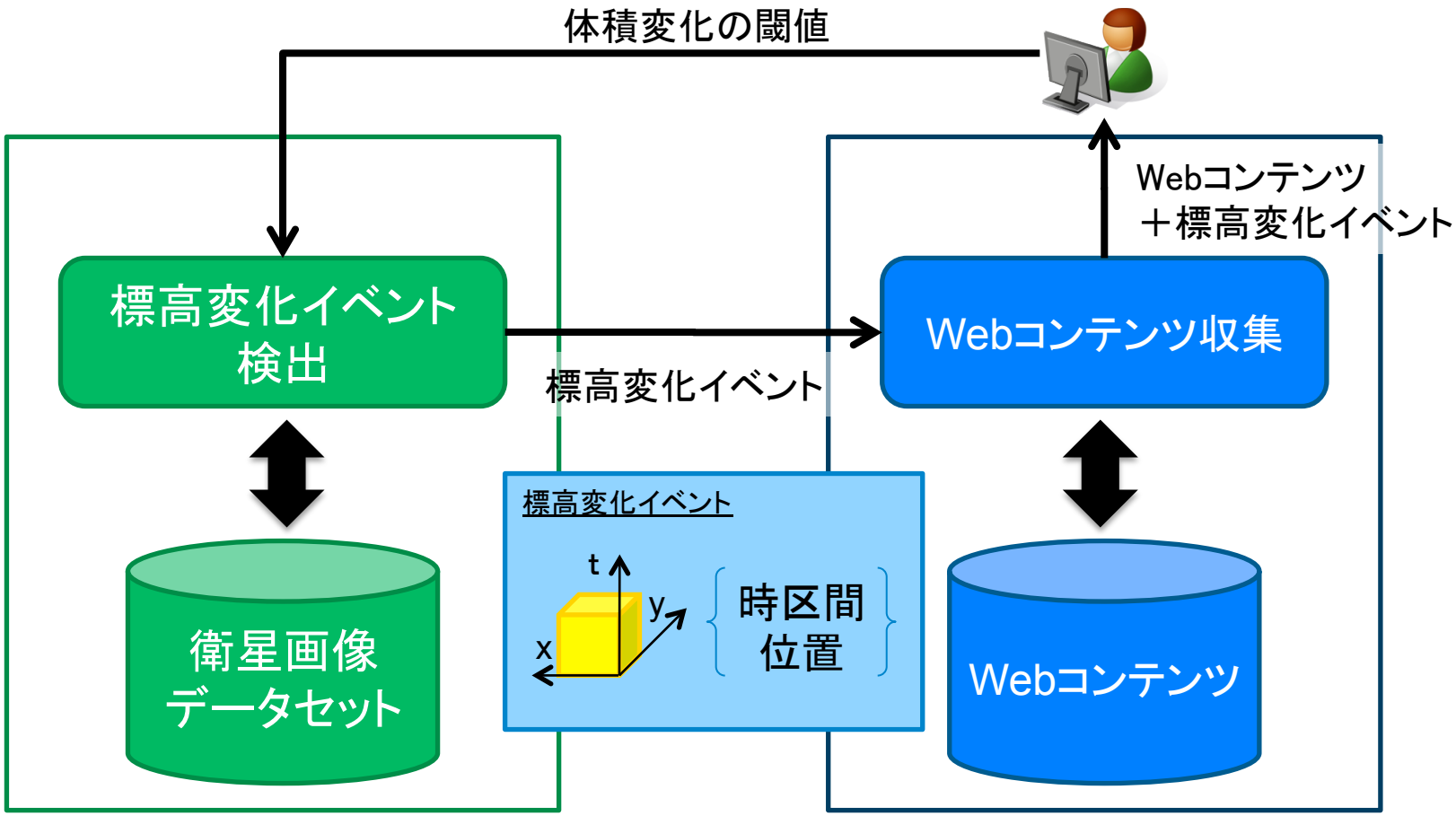


衛星画像



提案システム

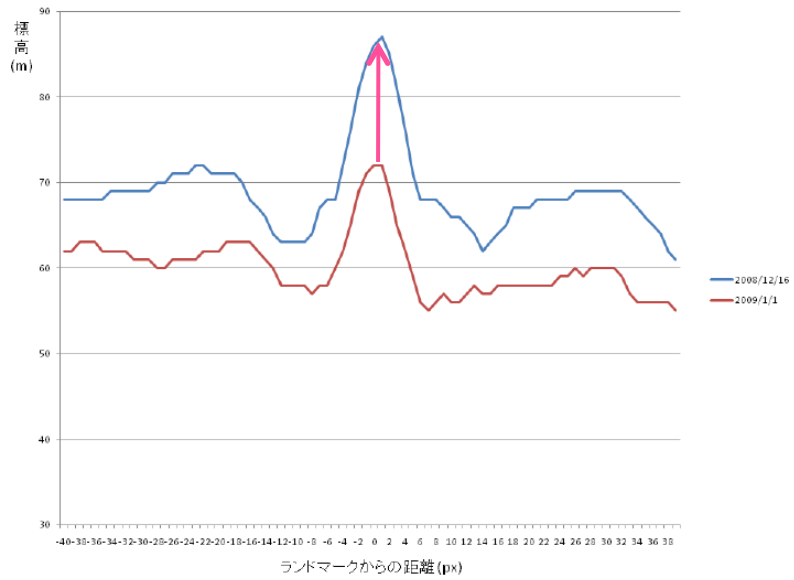
- 衛星画像とWebコンテンツを用いた建造物生成イベントの検出



提案システム　衛星画像の補正

- ASTER DEMの有する誤差
  - 相対的なオフセットが存在（児玉らの研究[*]）
  - 垂直方向：7.8*m*
  - 水平X方向：1.18*px*
  - 水平Y方向：0.67*px*

水平方向・垂直方向のオフセット除去



ASTER DEMの有する誤差

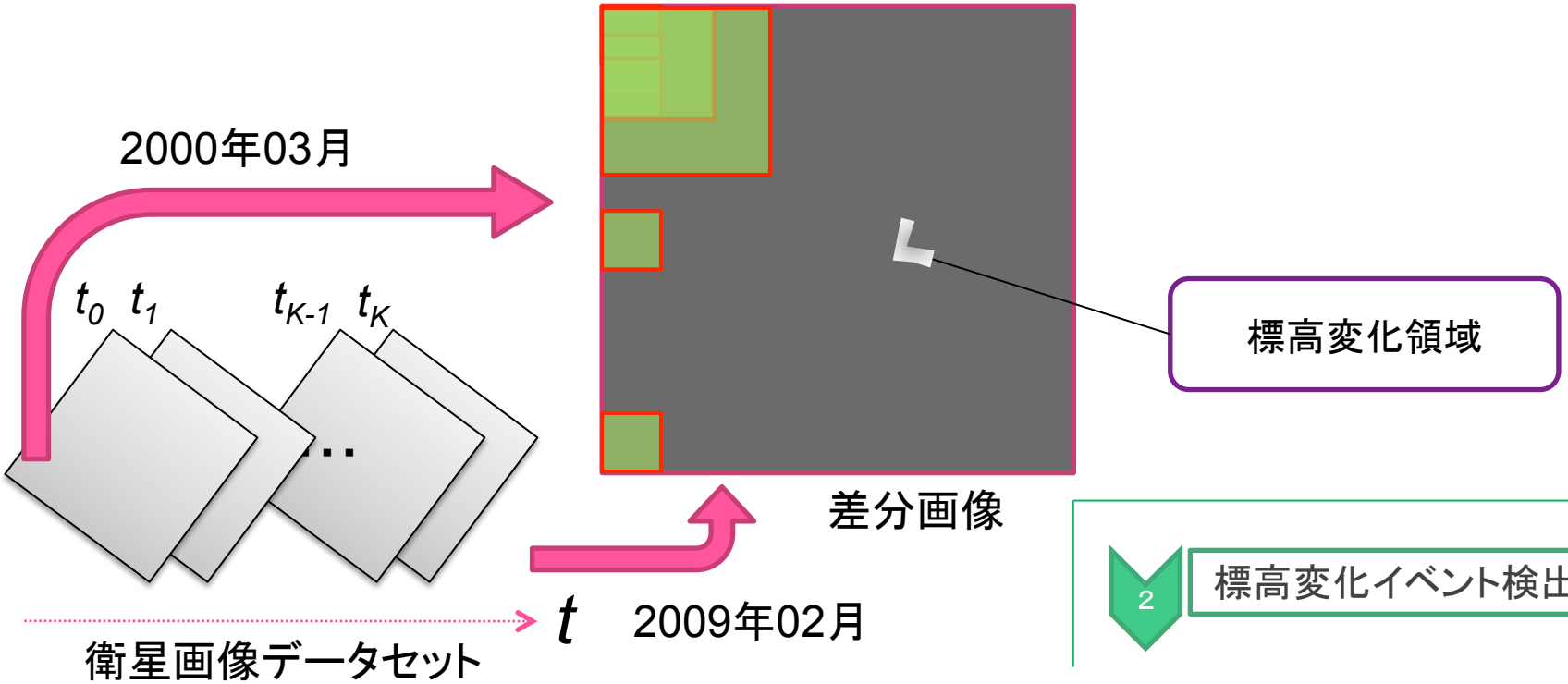
1

- 補正

[*] 児玉信介，在岡麻衣，三尾有年，中村良介，岩男弘毅．ASTER DEM の位置精度検証．第43 回日本リモートセンシング学会学術講演会(2007).

提案システム　標高変化イベント検出

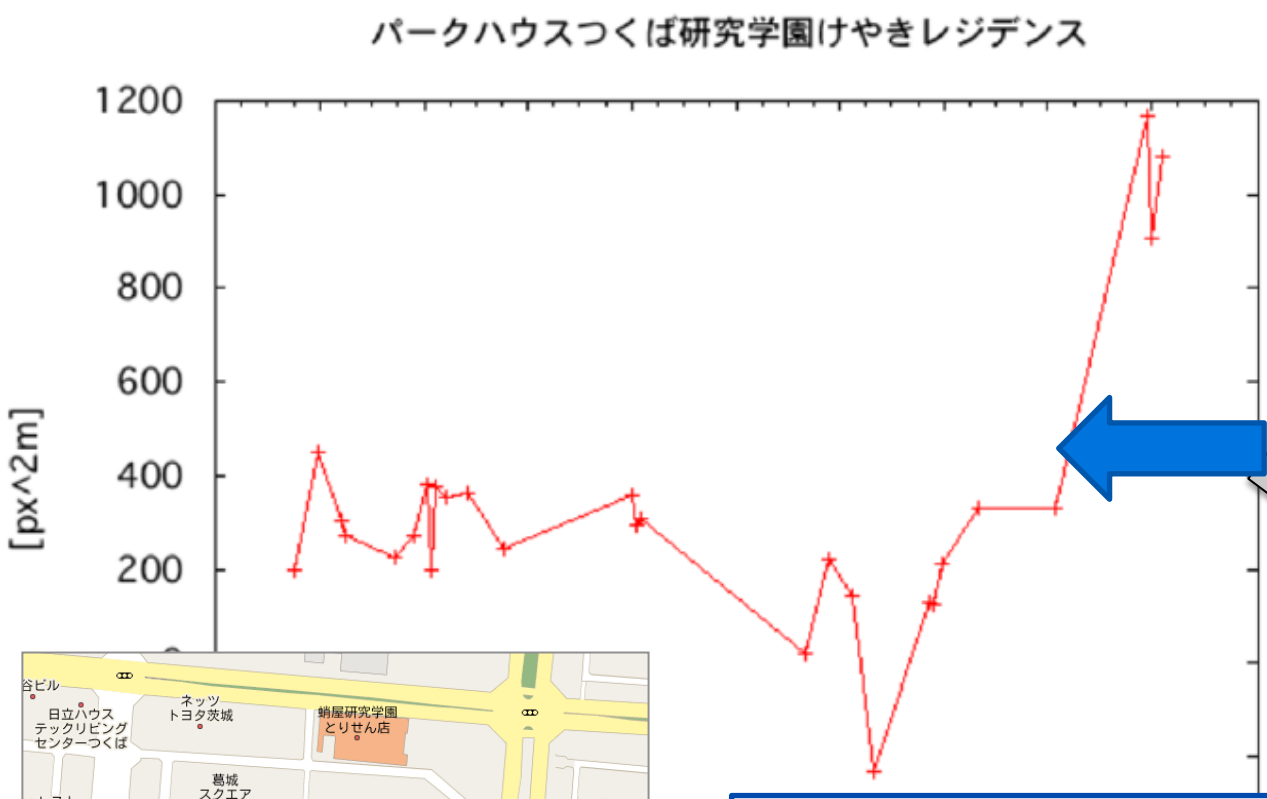
- 標高変化領域検出
  - 衛星画像データセット内の最新・最古画像から差分画像を生成
  - 探索ウィンドウが差分画像内をスウィープ
    - 探索ウィンドウ：正方矩形領域
  - 体積変化が閾値を超える探索ウィンドウ位置を検出
  - 探索ウィンドウ幅を最小~最大幅で変化させながら，上記を繰り返す



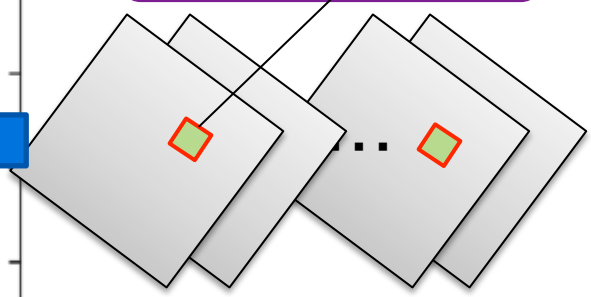
2 標高変化イベント検出

提案システム　標高変化イベント検出

標高変化時間の検出



標高変化領域



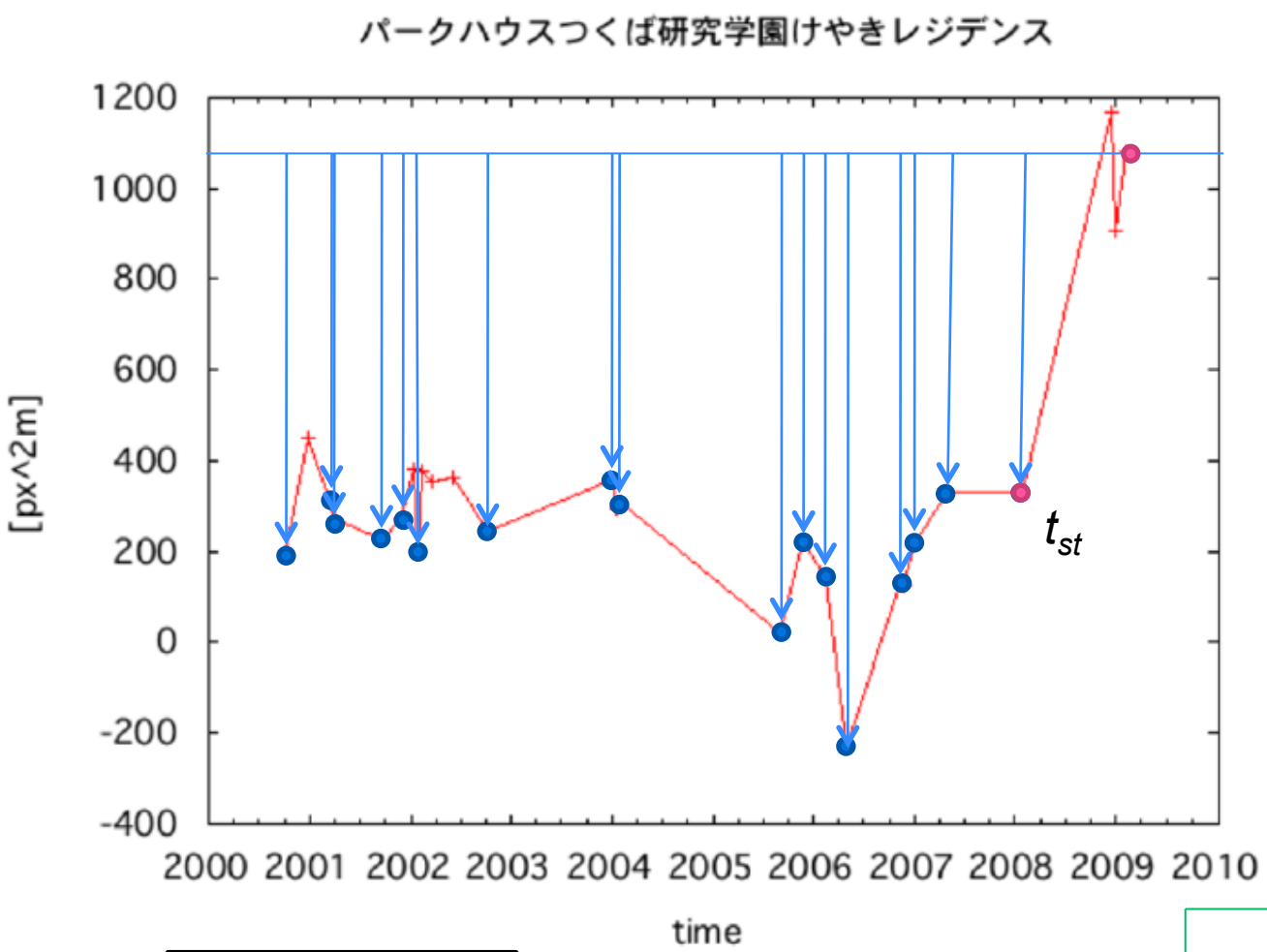
t

衛星画像データセット

標高変化領域内の標高値総和を
データセット内すべての画像で計算し
時系列にプロット

2 標高変化イベント検出

提案システム　標高変化イベント検出



閾値：800px^{2}m

標高変化時間の検出

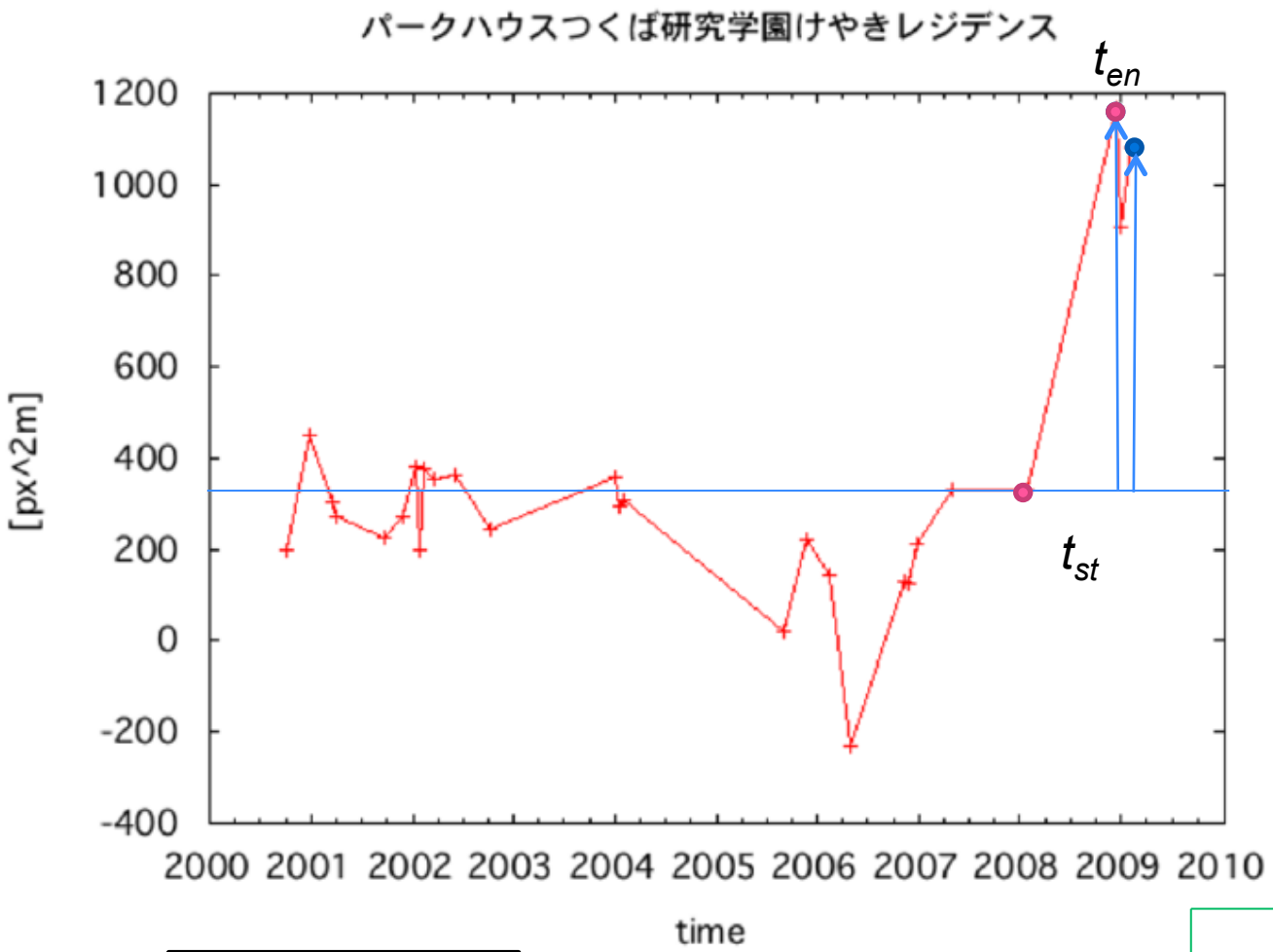
t_{st}の算出

①標高変化領域において最新画像との標高値総和の差が閾値を超える画像を検出
②検出した画像のうちタイムスタンプの最も遅いものを検出

標高変化時間=(t_{st}, t_{en})

2 標高変化イベント検出

提案システム　標高変化イベント検出



閾値：800px^{2}m

標高変化時間の検出

t_{en}の算出

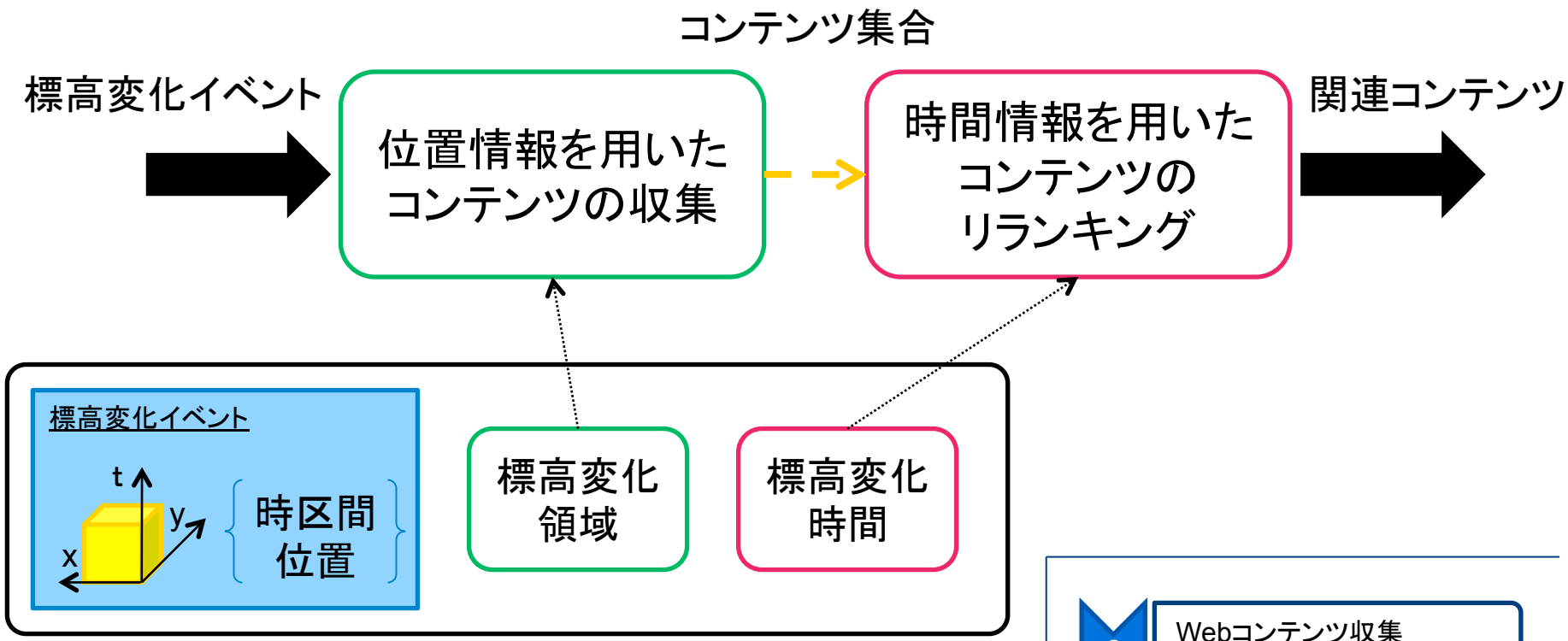
①標高変化領域においてタイムスタンプt_{st}を持つ画像との標高値総和の差が閾値を超える画像を検出
②検出した画像のうちタイムスタンプの最も早いものを検出

標高変化時間=(t_{st}, t_{en})

2 標高変化イベント検出

提案システム　関連Webコンテンツの収集

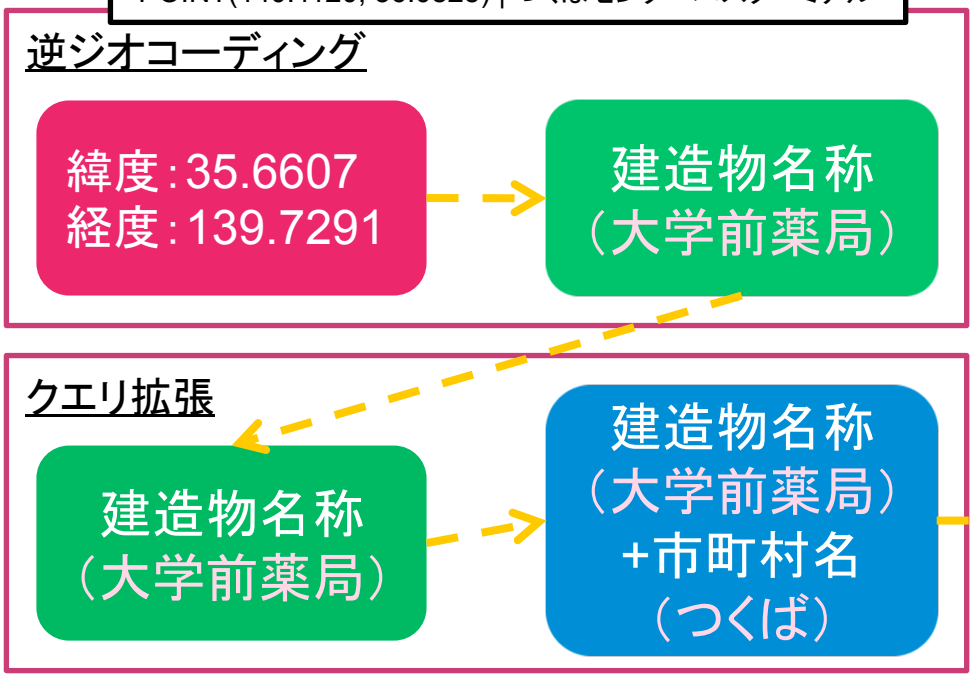
- 関連Webコンテンツの収集方法
  - 標高変化イベントと時空間距離の近いコンテンツの収集



提案システム　関連Webコンテンツの収集

- 位 … テンツの収集
  - … の作成

位置座標	建造物名称
POINT(140.1168, 36.0847)	吾妻一丁目住宅
POINT(140.1166, 36.0840)	時の公園
POINT(140.1169, 36.0829)	吾妻一丁目住宅
POINT(140.1165, 36.0838)	時の公園
POINT(140.1091, 36.0797)	吾妻一丁目住宅
POINT(140.1091, 36.0794)	吾妻一丁目住宅
POINT(140.1105, 36.0812)	南1駐車場管理センター
POINT(140.1103, 36.0817)	ダイワロイネットホテルつくば
POINT(140.1106, 36.0800)	OSIつくば吾妻2
POINT(140.1120, 36.0825)	つくばセンターバスターミナル



—“大学前薬局”[つくば]
—“大学前薬局”[新潟]
—“大学前薬局”[金沢]

複数の建造物が存在

3 Webコンテンツ収集

評価実験

- 標高変化イベント検出処理の検出精度評価
- Webコンテンツ収集処理の適合率評価
- システム評価

実験データ

ASTER DEM	
観測期間	2000年から2009年
観測地域	つくば市
データ件数	27
垂直方向解像度	1m
水平方向解像度	15m

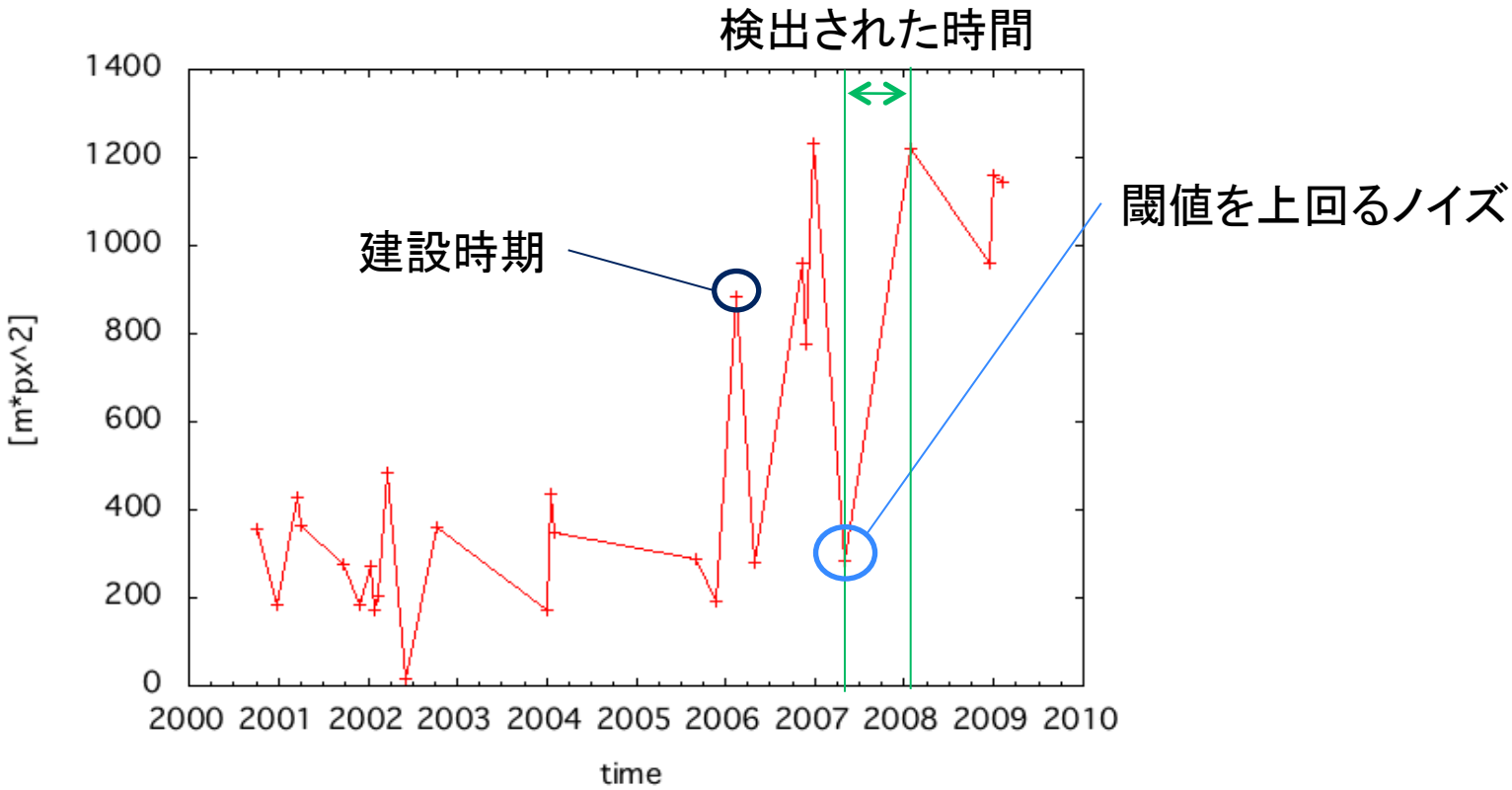
評価実験　標高変化イベント検出

建造物名	標高変化時間 開始時刻	標高変化時間 終了時刻	建設時期	建設事象検出
つくばスカイホテル	2007-05	2008-01	1982-01	失敗
アドヴァンスつくば	2006-12	2008-12	2008-02	成功
イーアスつくば	2007-05	2009-01	2008-10	成功
サーパスつくば研究学園	2007-05	2009-02	2007-09	成功
パークハウスつくば研究学園けやきレジデンス	2007-05	2008-12	2009-01	失敗
パークハウスつくば研究学園さくらレジデンス	2007-05	2008-12	2009-01	失敗
ファースト・ラフィーネ・学園	2006-05	2007-05	2006-06	成功
ホテルベストランド	2006-05	2007-05	2007-03	成功
山新グランステージつくば	2007-05	2008-01	2006-05	失敗
日本板硝子（株）技術研究所筑波研究センター	2006-05	2006-11	1983-10	失敗
研究学園駅前岡田ビル	2007-05	2009-02	2007-03	失敗

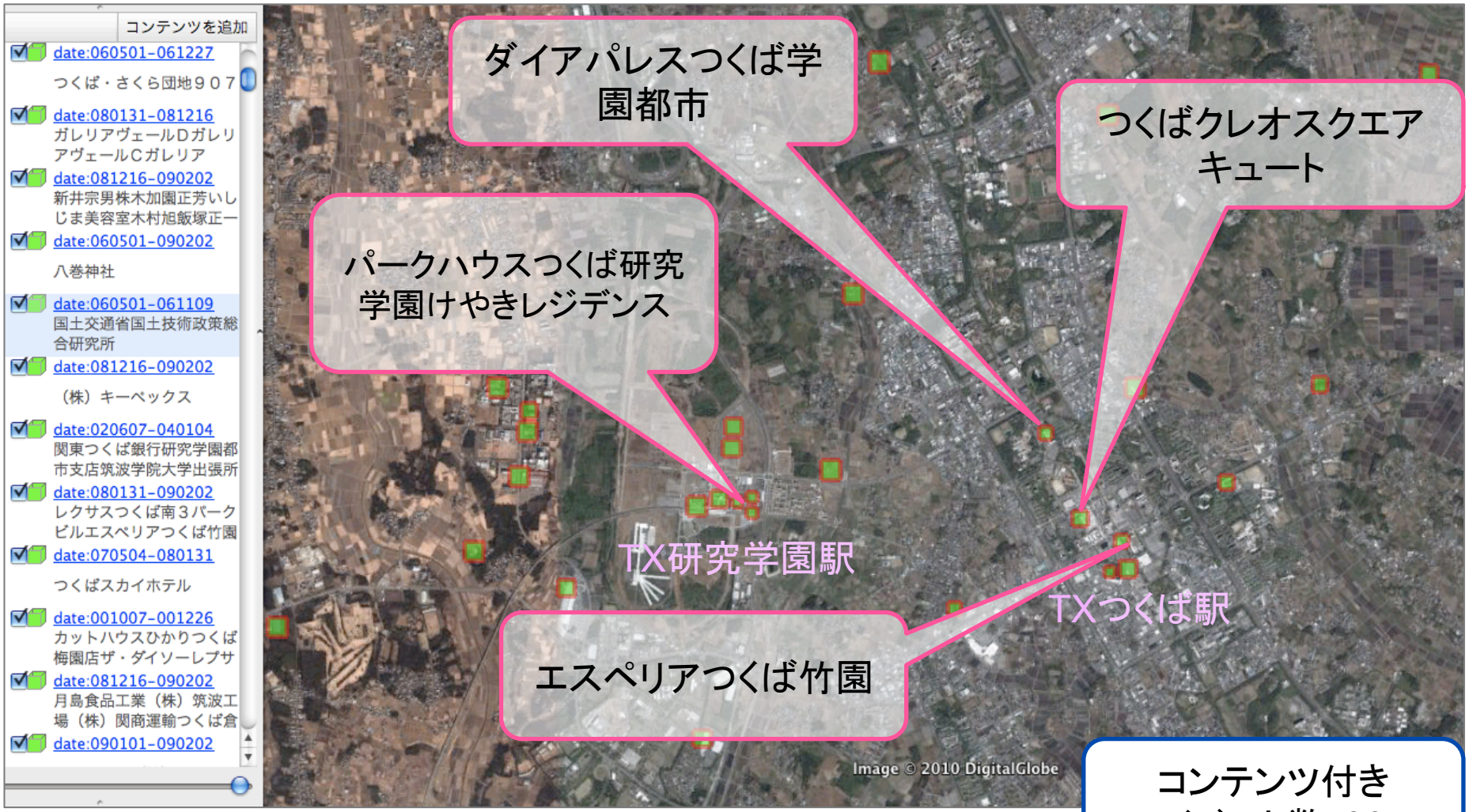
- 標高変化時間を考慮しない場合（建設時期が2000年-2009年）
  - 適合率：82%（検出成功：9件、検出失敗：2件）
- 標高変化時間を考慮する場合
  - 適合率：45%（検出成功：5件、検出失敗：6件）

評価実験　標高変化イベント検出

- 山新グランステージ
  - 建設時期：2006-05



評価実験 システム評価



コンテンツ付き
イベント数：30

- 体積変化の閾値：$800px^2m=180000m^3$

まとめと今後の課題

- まとめ
  - 衛星画像を用いた有用なWebコンテンツの収集
  - 衛星画像とWebコンテンツから建造物生成イベントの検出
  - 衛星画像からの標高変化イベント検出
    - 82%の適合率で標高変化領域を検出
    - 45%の適合率で標高変化領域・時間を検出
  - 標高変化イベントを用いた関連Webコンテンツの収集
    - 平均61%の適合率で関連Webコンテンツを収集
- 今後の課題
  - 衛星画像補正の精度向上
    - 森林領域によるフィルタリング
  - Webコンテンツ収集の精度向上
    - HTML構造に着目した記事分割

単語の専門性に着目した
気象学論文からの専門語抽出

宇井敬一朗，天笠俊之，北川博之

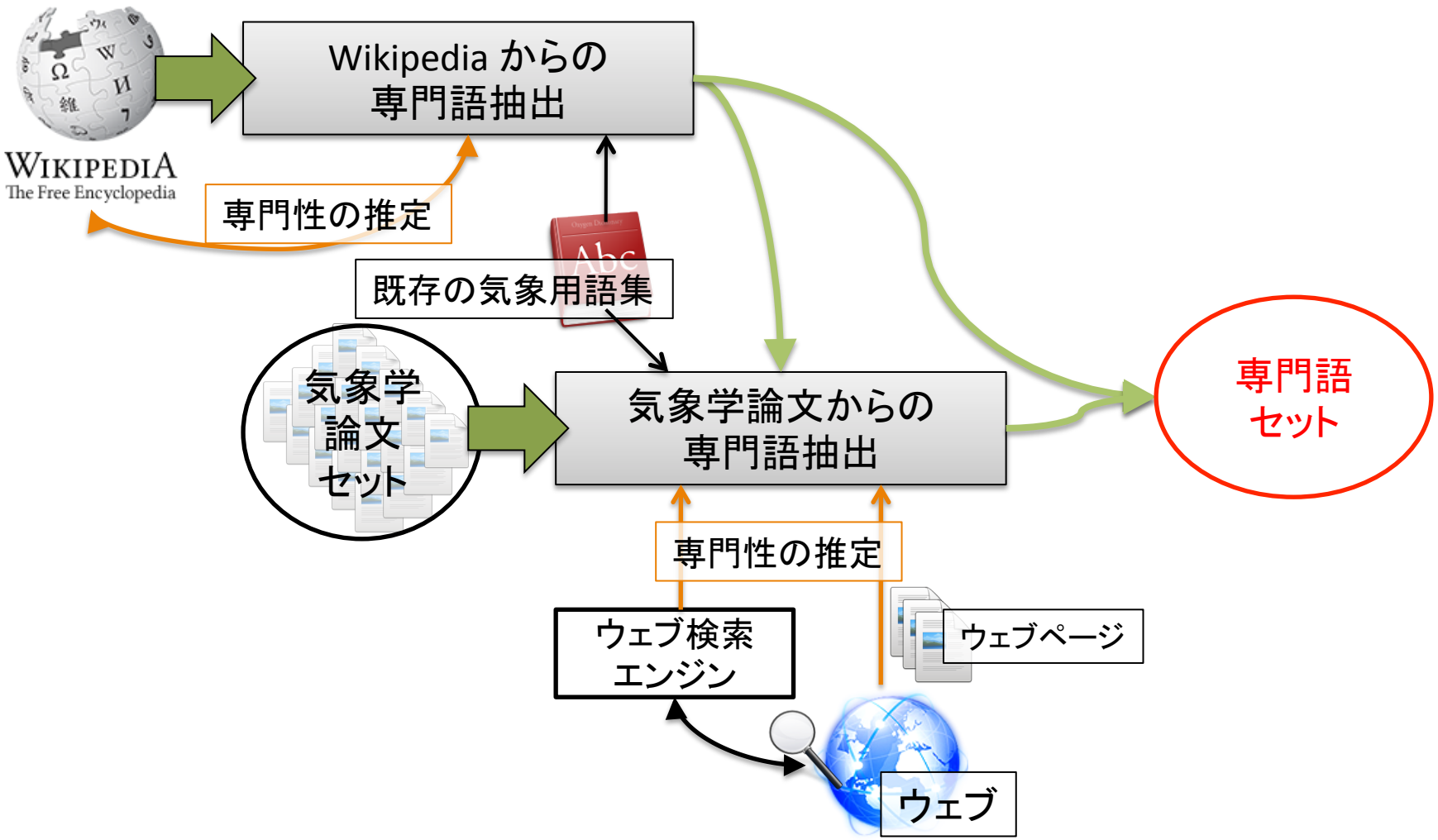
サイエンスにおける文書情報の利活用

- 科学データ
  - 構造データ
    - 観測データ，実験データ，　…
  - 非構造データ
    - 論文，レポート，議事録，WWW, …
  - 例
    - 分子生物学分野
      - Medline, PubMed
      - 遺伝子の関連情報を文献情報から取得
- 文書情報の利活用の重要
  - 文書に埋もれた知識の発見
  - ドメインオントロジーの構築
  - …

気象学を対象にした専門語抽出

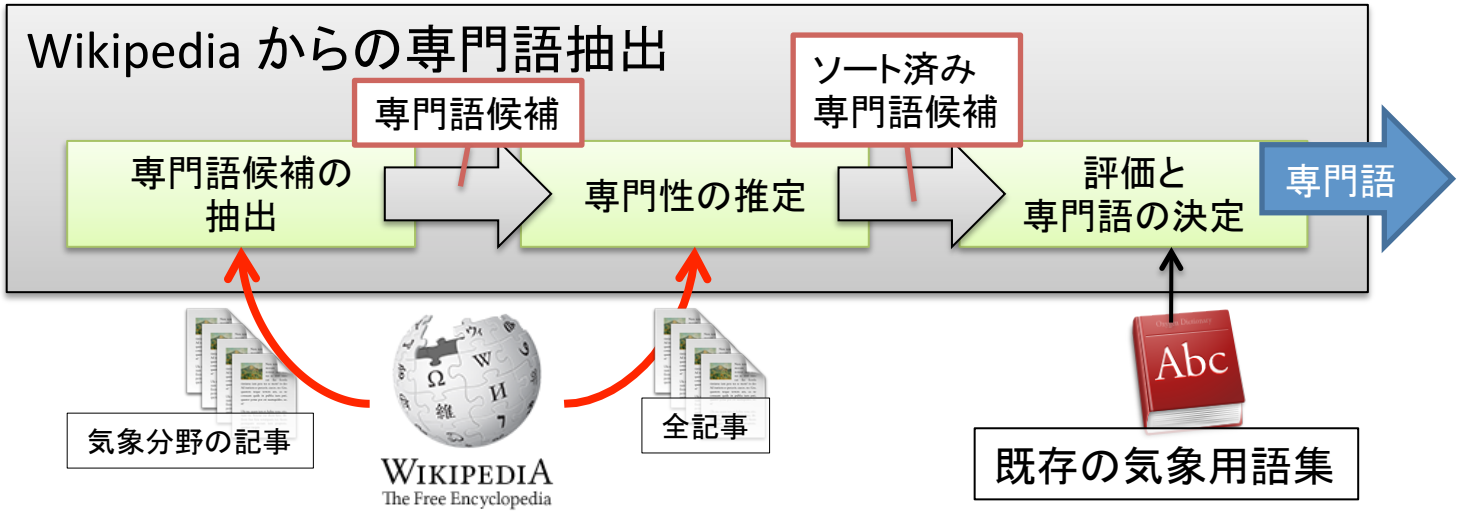
- 他分野の文書セットを用意する必要を除去
  - Wikipedia
    - 気象分野，他分野の両方を持つ
  - 気象学論文
    - 気象分野のみ
  - ウェブ
    - 気象分野，他分野の両方を持つ
- 新たな専門性の推定手法
  - 気象学と他分野に関する文書における出現頻度
  - ウェブ検索エンジンを利用

提案手法: 概略


WIKIPEDIA
The Free Encyclopedia
Wikipedia からの
専門語抽出
専門性の推定
既存の気象用語集
気象学
論文
セット
気象学論文からの
専門語抽出
専門語
セット
専門性の推定
ウェブ検索
エンジン
ウェブページ
ウェブ

Wikipedia からの専門語抽出

- 三つのフェーズ
  - 専門語候補の抽出
  - 専門性の推定
  - 評価と専門語の決定



Wikipedia からの専門語抽出
- 専門性の推定 -

- ベクトルで表現
  - カテゴリベクトル V_{Cat} を生成
  - 専門語候補 t をベクトル V_t で表現

記事 d1 は，気象学に関する記事**ではない**

記事 d2 は，気象学に関する記事**である**

	d1	d2	d3	…	dN
V_{Cat} = [	0	1	1	…	0]
V_t = [	0	10	3	…	7]

専門語候補 t は，記事 d1 において**出現しない**

専門語候補 t は，記事 d2 において **10 回出現**

Wikipedia からの専門語抽出
- 専門性の推定 -

- ベクトル空間モデルを用いた専門性推定
  - V_t: 専門語候補 t のベクトル
    - V'_t: ベクトル V_t を正規化したベクトル
  - V_{cat}: ドメインカテゴリベクトル
  - df_t: 専門語候補 t が出現する記事の数
- 推定したスコアによる降順ソート

$$\mathrm{score}_{cat}(t) = \frac{V_{cat} \cdot V'_t}{\log(df_t)}$$

$$V'_t = \frac{V_t}{|V_t|}$$

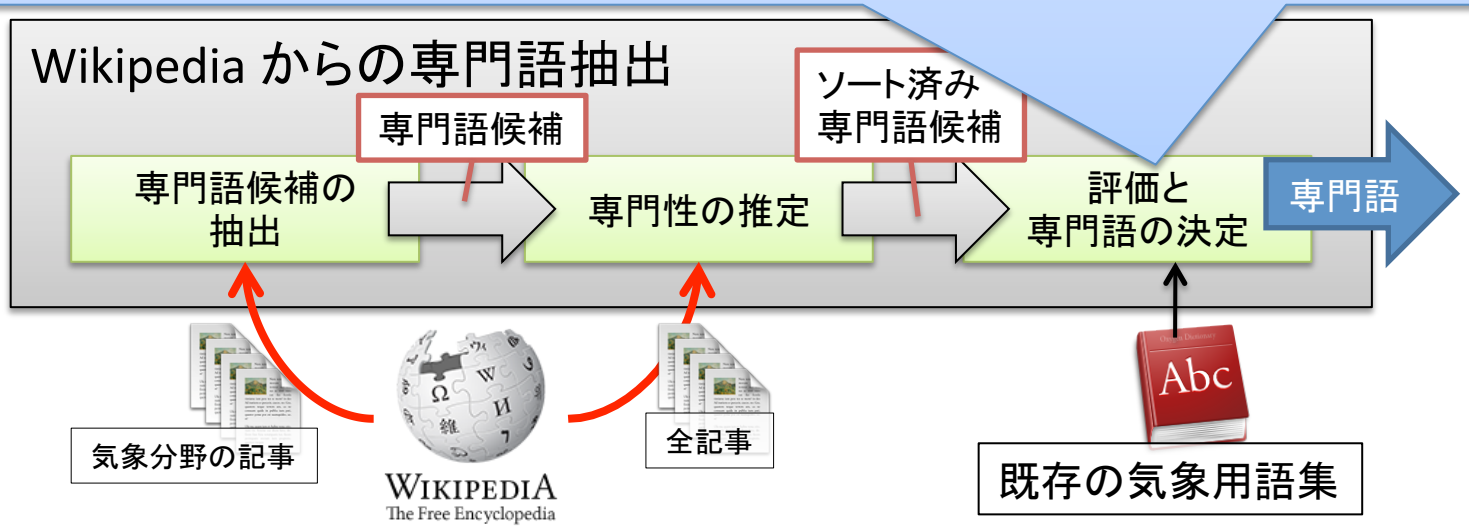
ソート

専門語候補	score
hazard information	2.93×10^{-2}
NOAA WEATHER OFFICE	2.92×10^{-2}
forecast office	2.59×10^{-2}
…	…
serious error	3.97×10^{-4}
…	…

降順

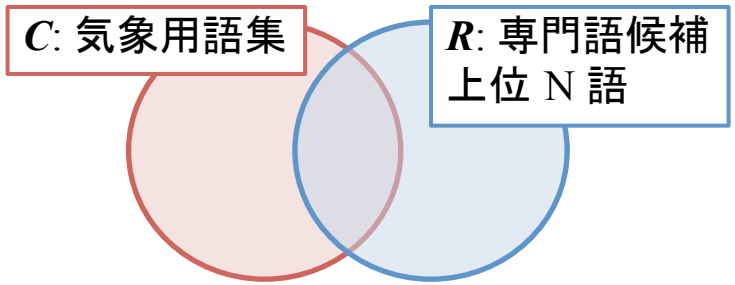
Wikipedia からの専門語抽出
- 評価と専門語の決定 -

- 評価
  - 既存の気象用語集を正解とする
- 専門語の決定
  - 評価の結果から，上位何語を専門語とするか決定する



Wikipedia からの専門語抽出
- 評価と専門語の決定 -

- 評価
  - Recall, Precision, F-measure を算出
  - 評価対象がどれだけ正解しているか
    - 正解セット ***C***: 既存の気象用語集
    - 候補セット ***R***: 専門語候補上位 N 語

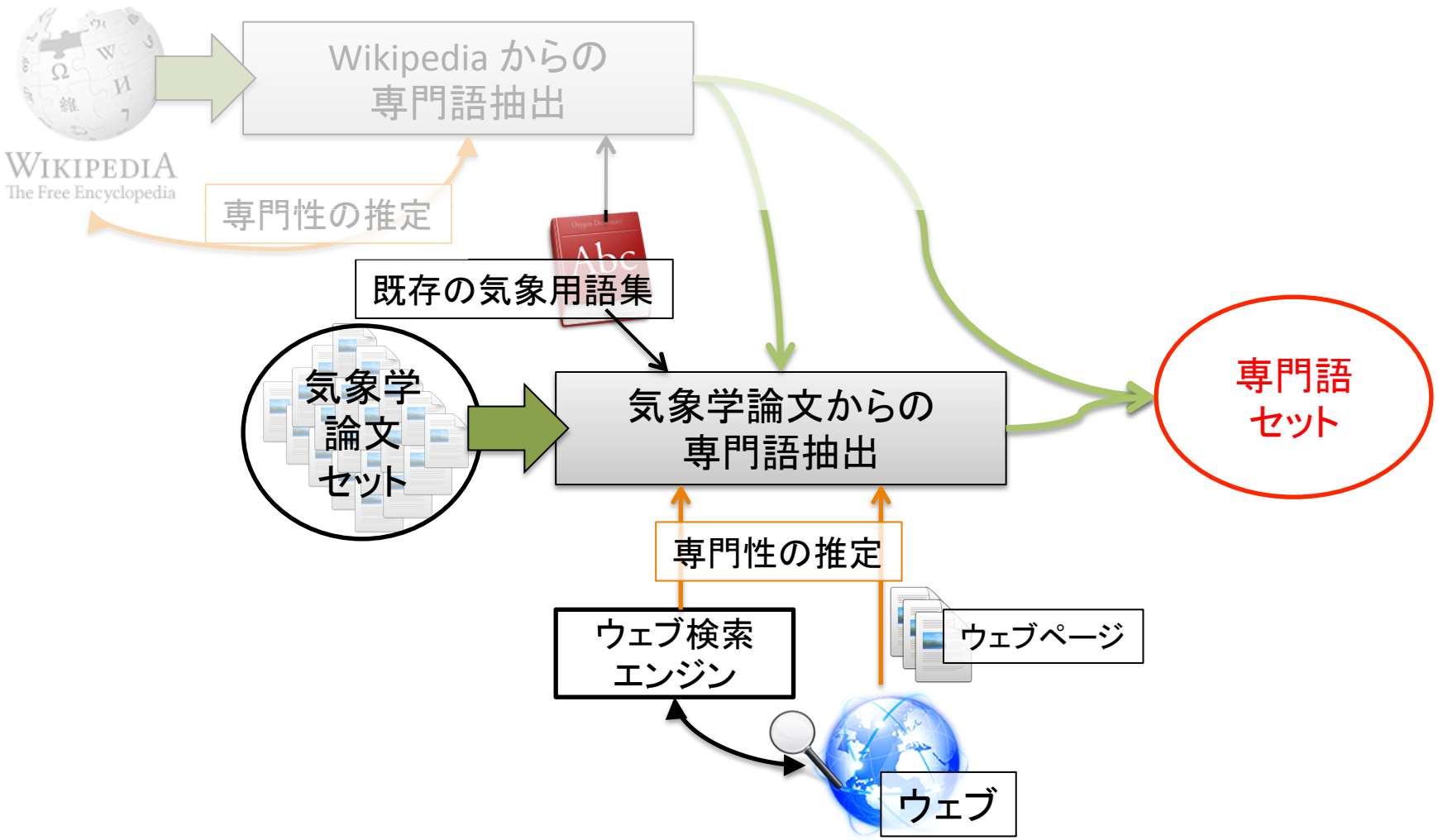


$$\text{Recall} = \frac{\boldsymbol{C} \cap \boldsymbol{R}}{\boldsymbol{C}}, \text{Precision} = \frac{\boldsymbol{C} \cap \boldsymbol{R}}{\boldsymbol{R}}$$

$$\text{F-measure} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

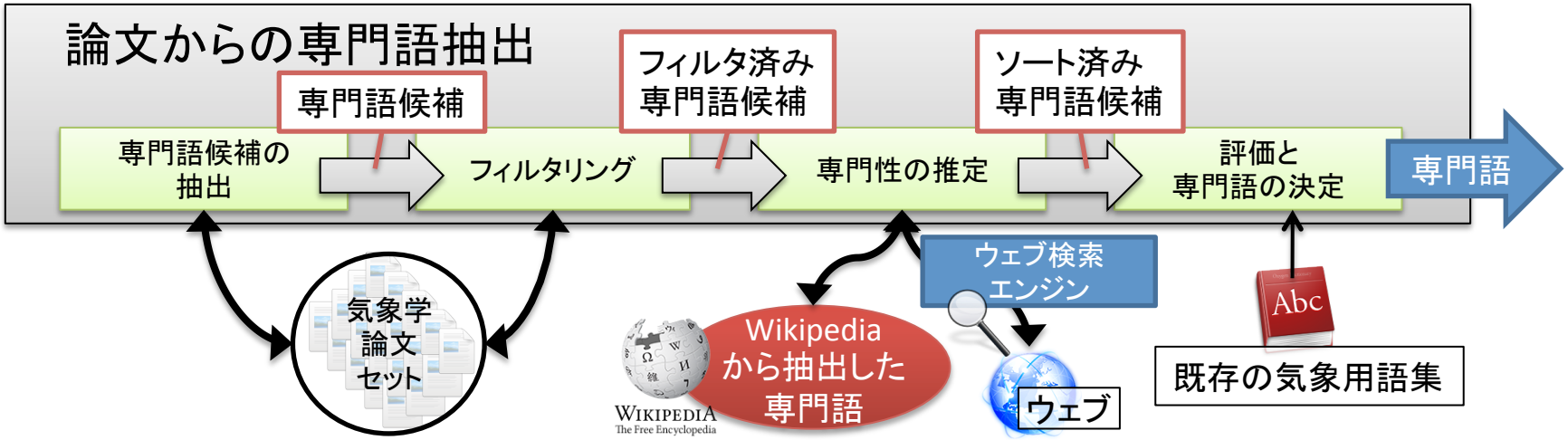
- 専門語の決定
  - Recall, Precision, F-measure から，上位何語を専門語とするか決定する

提案手法: 概略


WIKIPEDIA
The Free Encyclopedia
Wikipedia からの
専門語抽出
専門性の推定
既存の気象用語集
気象学
論文
セット
気象学論文からの
専門語抽出
専門語
セット
専門性の推定
ウェブ検索
エンジン
ウェブページ
ウェブ

気象学論文からの専門語抽出

- 四つのフェーズ
  - 専門語候補の抽出
  - フィルタリング
  - 専門性の推定
  - 専門語の決定

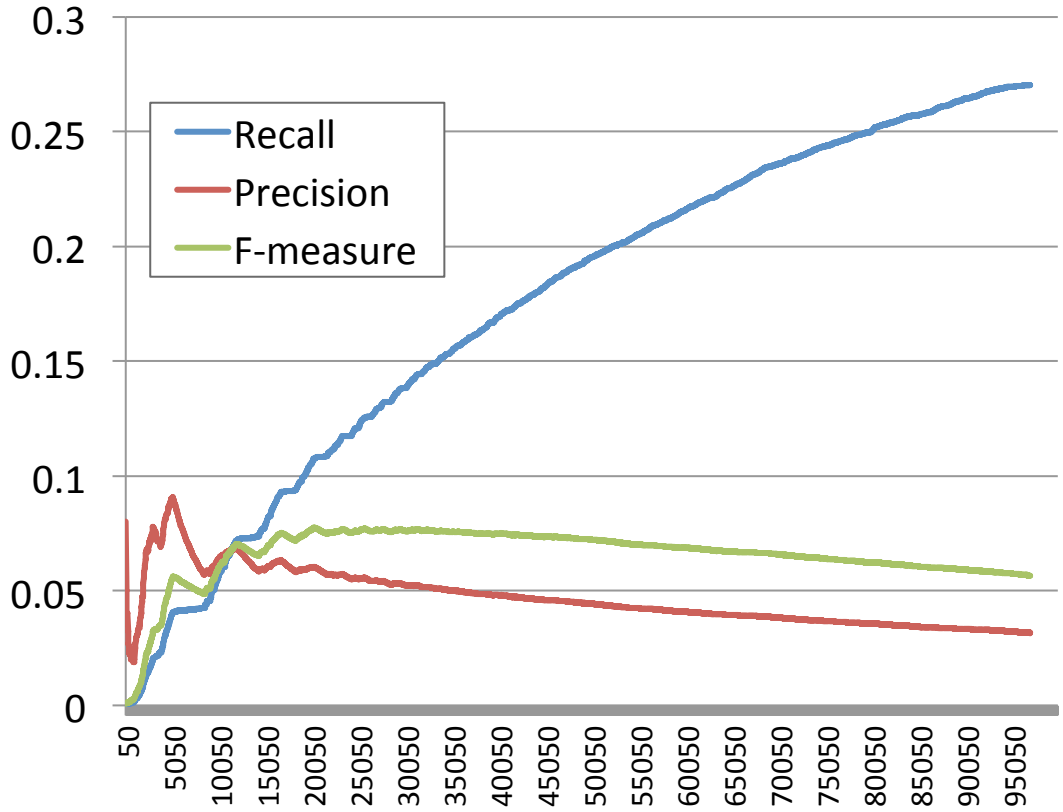


実験
Wikipedia からの専門語抽出

- 英語 Wikipedia のダンプデータ
  - 全記事数: 1,000万 (24 GB)
  - 気象学に関する記事数: 2,199

- 処理時間
  - 専門語候補の抽出
    - 24 分
  - 専門性の推定
    - 1 時間 31 分

F-measure が最大となる
上位 20,200 語
を専門語として抽出

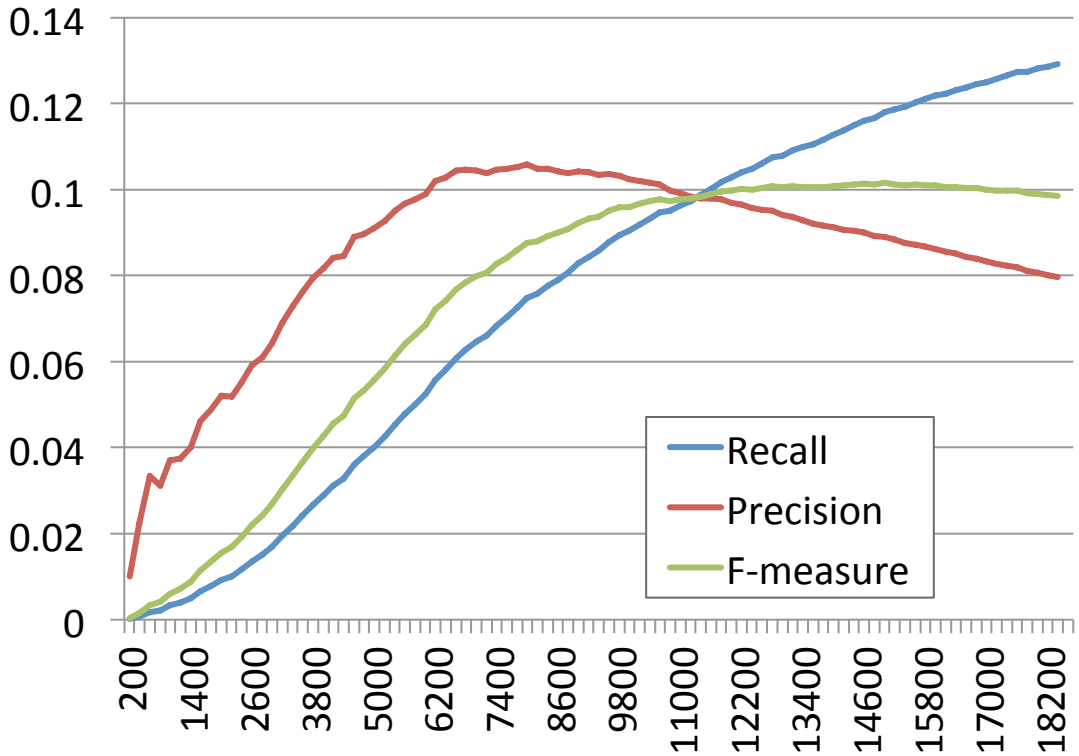


実験
気象学論文からの専門語抽出

- 気象学論文
  - 日本気象学会: 2 誌，3,067 論文（日本語論文は除去）
  - American Meteorology Society: 15 誌，21,396 論文
  - 計 24,463 の気象学論文

- 処理時間
  - 専門語候補の抽出
    - 25 時間
  - フィルタリング
    - 2 時間 32 分
  - 専門性の推定
    - 14 時間

Precision が最大となる
上位 7,400 語
を専門語として抽出



実験
専門家による評価

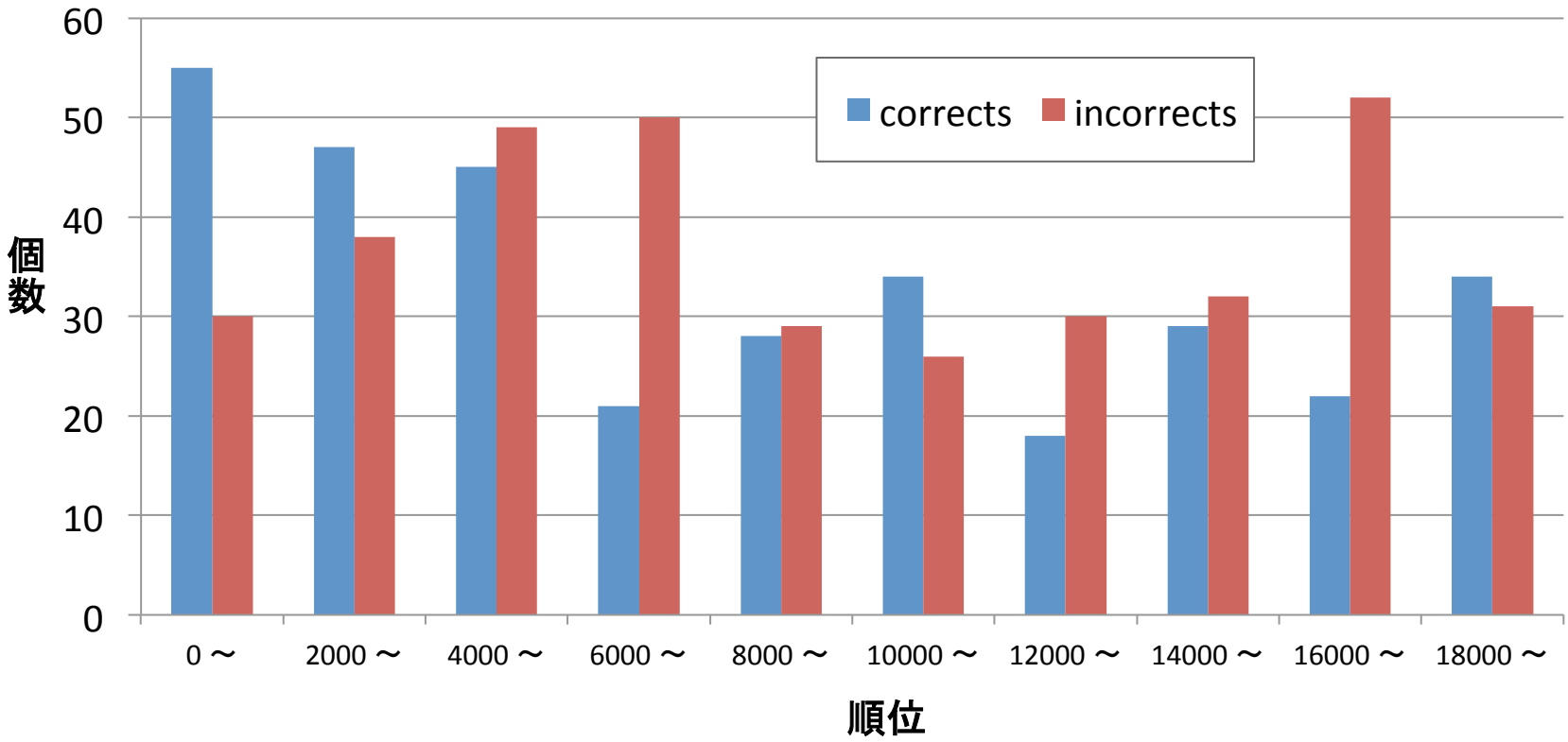
- 評価方法
  - 気象学・気候学を専門としている学生
    - 博士3名，修士4名
  - 専門語として抽出した200語
    - それぞれ異なった語を評価
    - 内訳
      - 100語: Wikipedia から抽出した専門語 20,050 語から無作為抽出
      - 100語: 気象学論文から抽出した専門語 72,000 語から無作為抽出

200語中，専門語として判断された語の数

抽出元＼添削者	博士A	博士B	博士C	修士D	修士E	修士F	修士G	平均
Wikipedia	70	45	34	39	59	35	51	47.57
気象学論文	87	54	34	35	77	37	56	54.29
合計	157	99	68	74	136	72	107	101.86

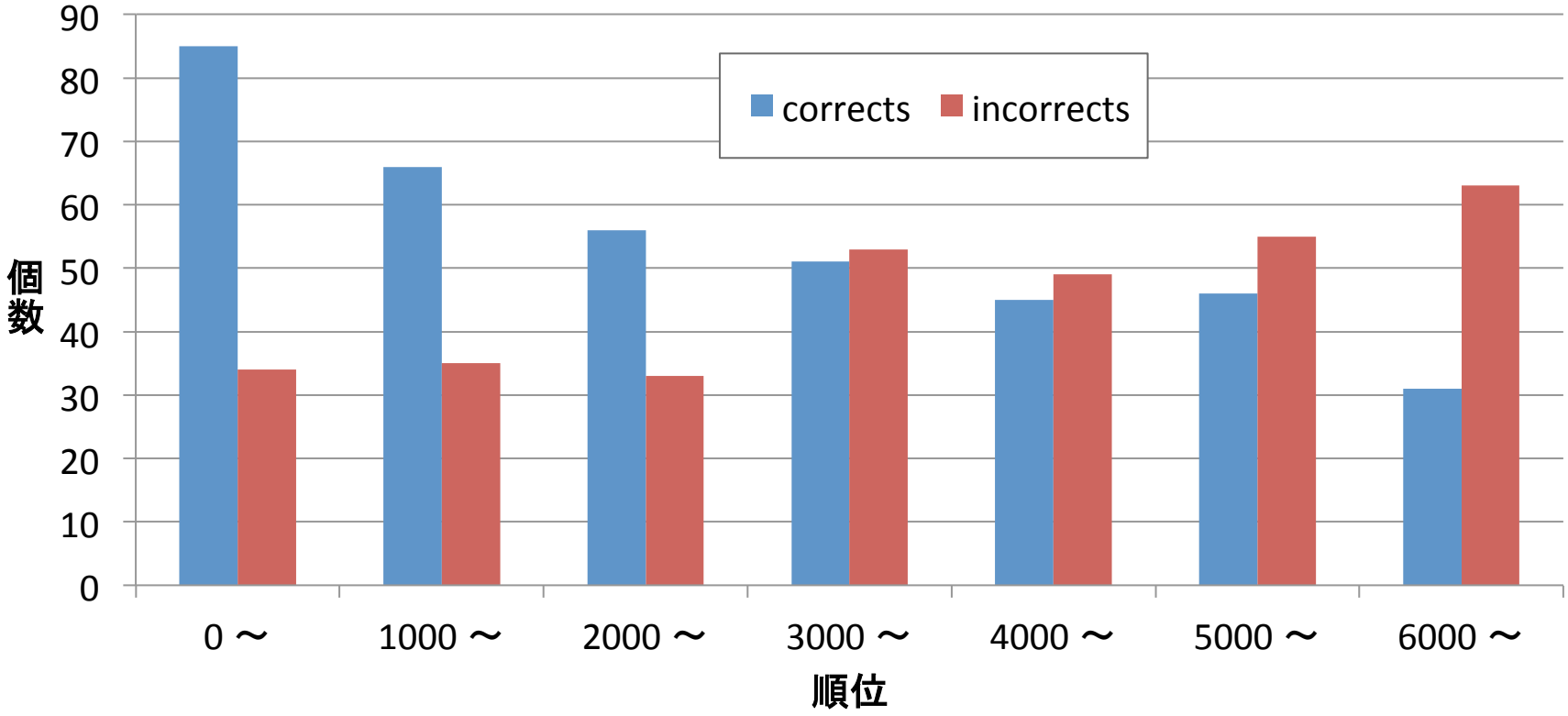
実験
専門家による評価

- 正解・不正解の分布
  - Wikipedia からの専門語抽出



実験
専門家による評価

- 正解・不正解の分布
  - 気象学論文からの専門語抽出



まとめと今後の課題

- 気象学分野を対象に専門語抽出を試みた
  - 新たな抽出手法の提案
    - TermExtractor の考え方をベース
    - Wikipedia からの専門語抽出
    - 気象学論文からの専門語抽出
  - 実験
    - Wikipedia から 22,000，気象学論文から 7,400 を専門語として抽出した
    - 気象学専攻の学生に評価してもらった
      - 約半数が正解であると判断された

- 今後の課題
  - スコア（専門性の推定）の改良
  - ソートした専門語候補全体に対し専門家による評価
    - スコアリングの手法の妥当性を確認
    - 上位 22,000 や 74,000 語で切るのが正しいのか検証