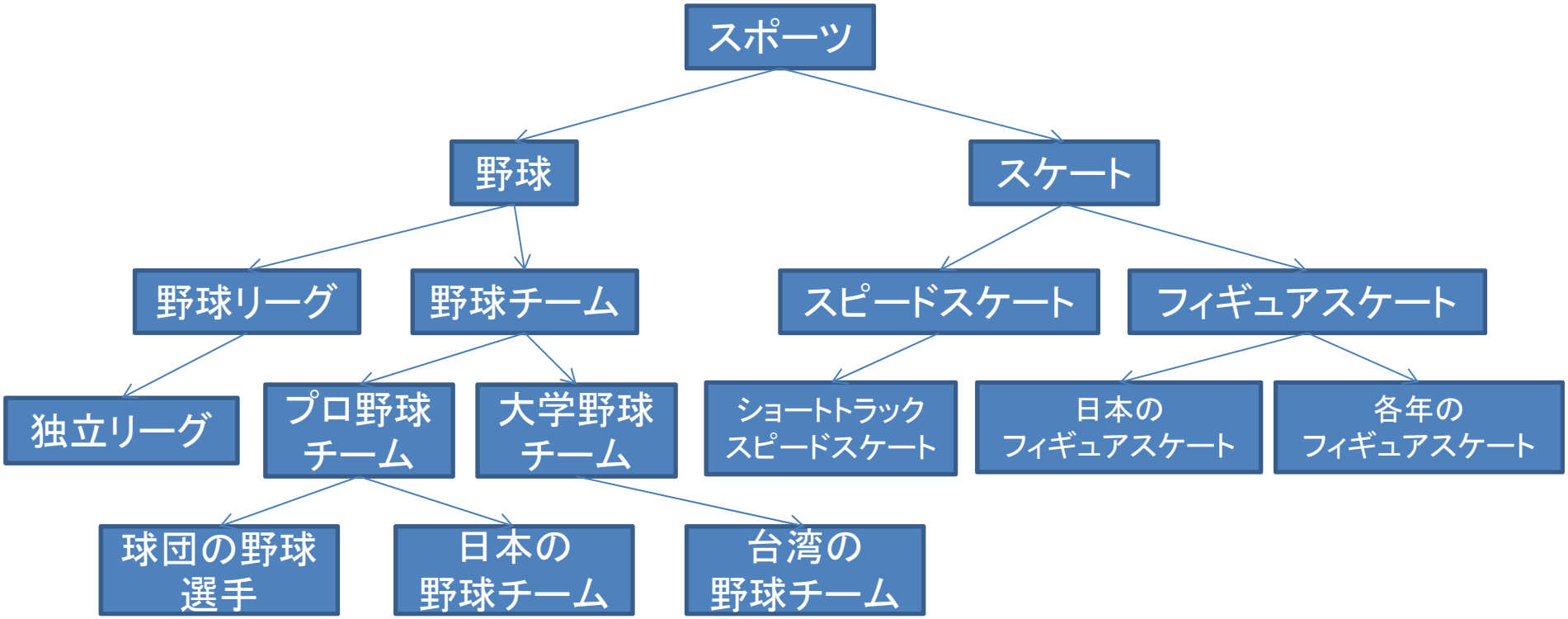
共起情報とオントロジーを併用した動画タグの階層化手法の提案

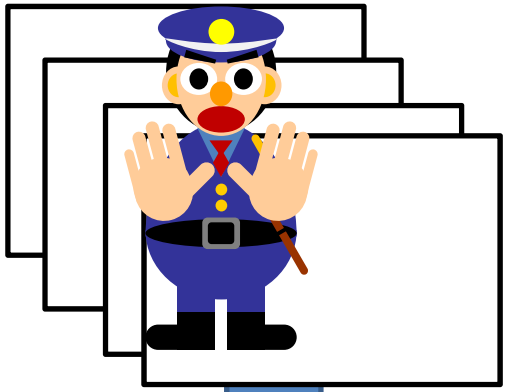
松永大希

オントロジーとは



簡潔に表現すると,
「言葉の階層構造とネットワーク」

研究の目的



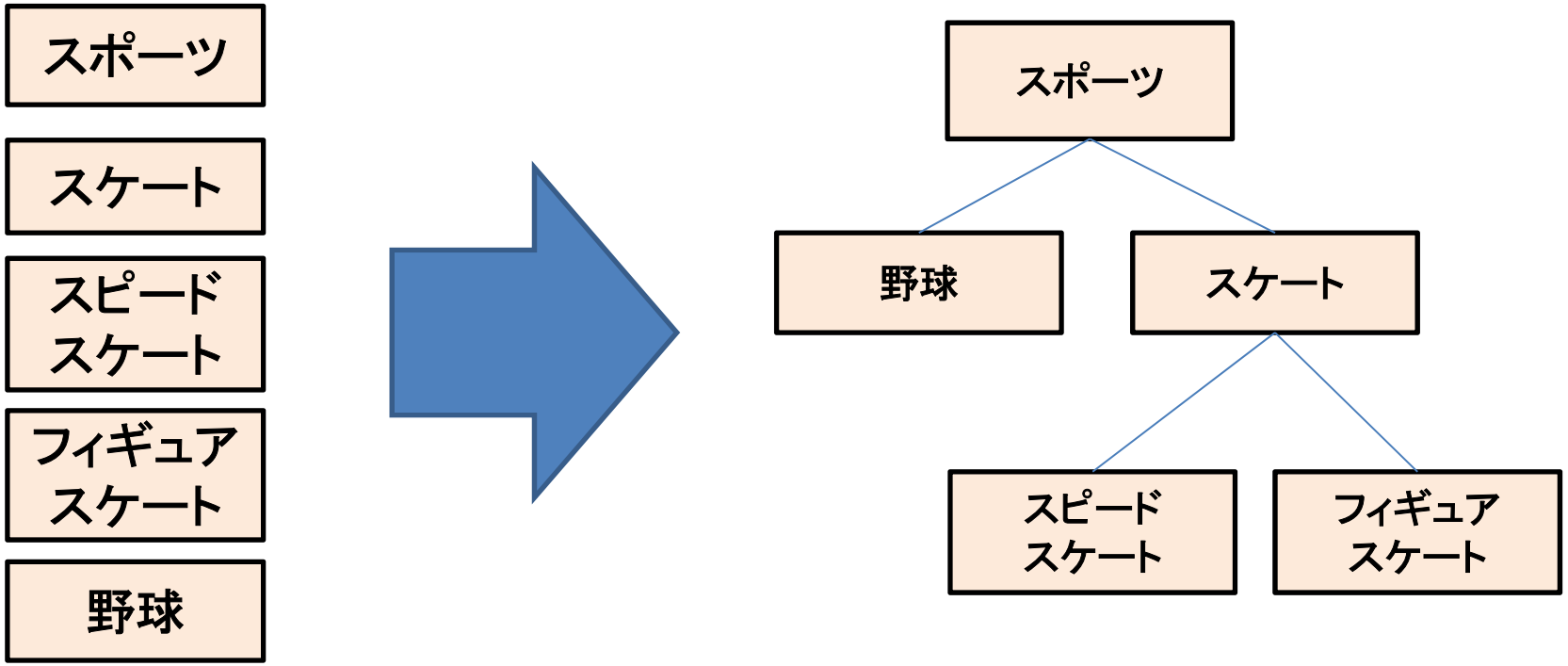
検索

好みの動画を見つけるには
高度な検索システムが必要…

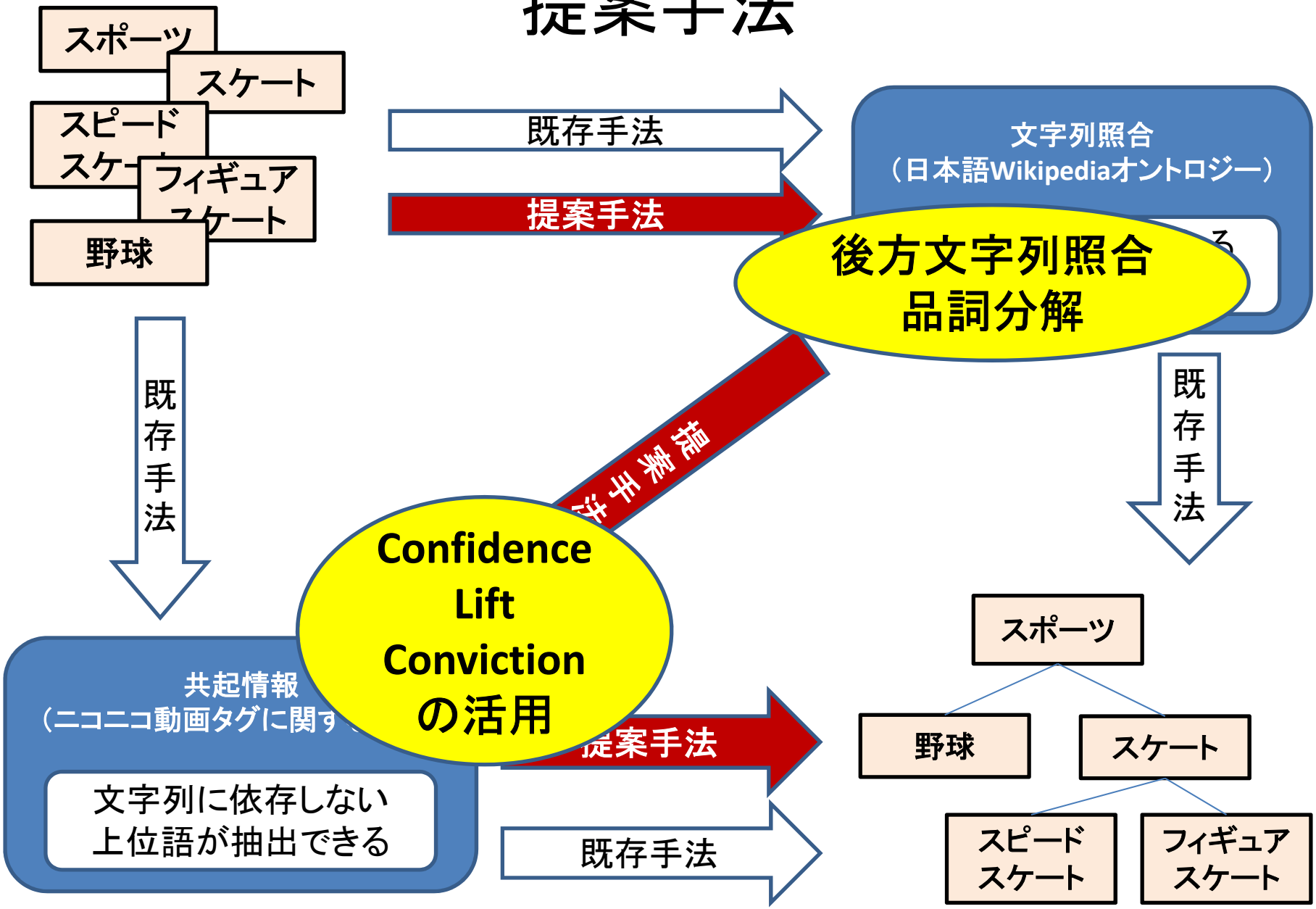
オントロジーを作ろう！

Ref: 村上直至，伊東栄典” 動画投稿サイトで付与された動画タグの階層化”(2010)

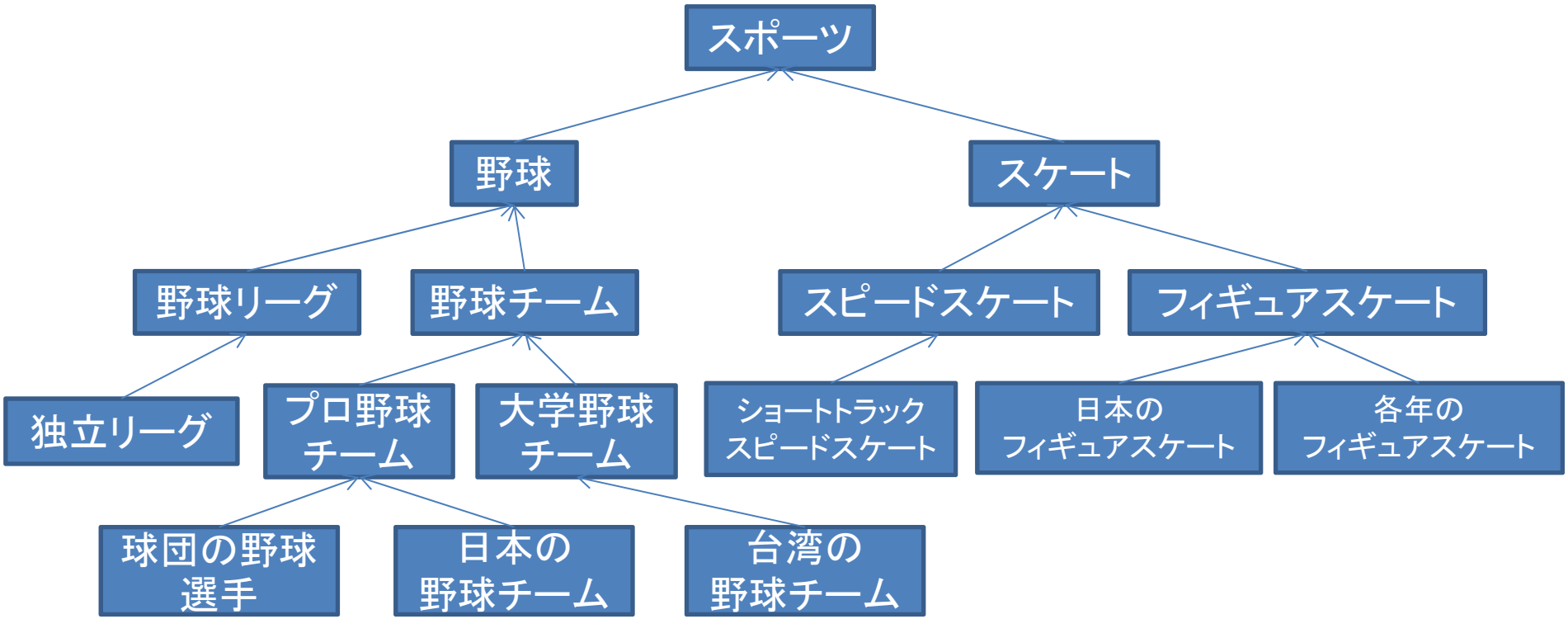
オントロジーを作るとは…



提案手法

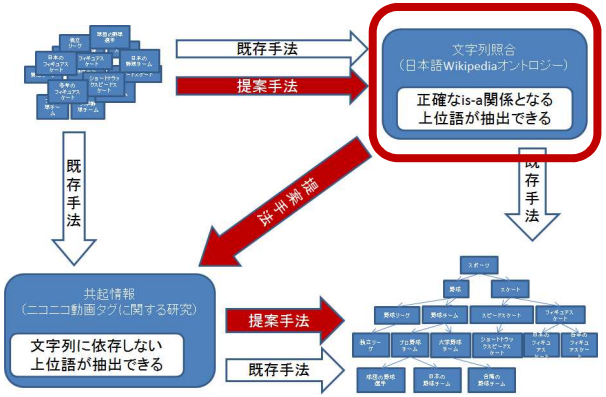

スポーツ
スケート
スピード
スケート
フィギュア
スケート
野球
既存手法
提案手法
文字列照合
（日本語Wikipediaオントロジー）
後方文字列照合
品詞分解
既存手法
既存手法
提案手法
Confidence
Lift
Conviction
の活用
共起情報
（ニコニコ動画タグに関す
文字列に依存しない
上位語が抽出できる
提案手法
既存手法
スポーツ
野球
スケート
スピード
スケート
フィギュア
スケート

上位語を抽出する理由



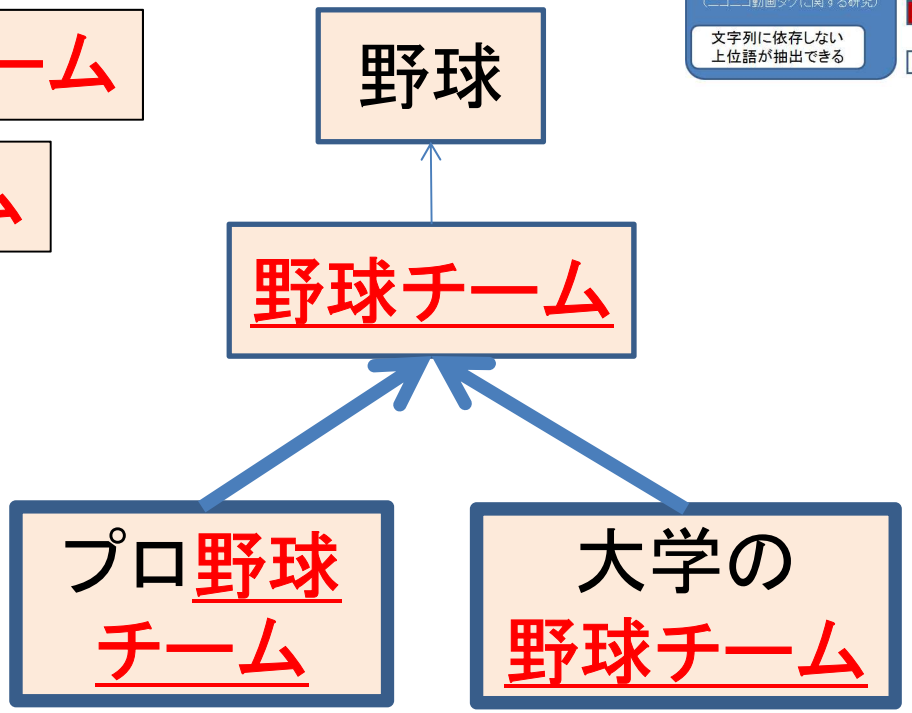
上位語を辿ると，オントロジーができる

後方文字列照合



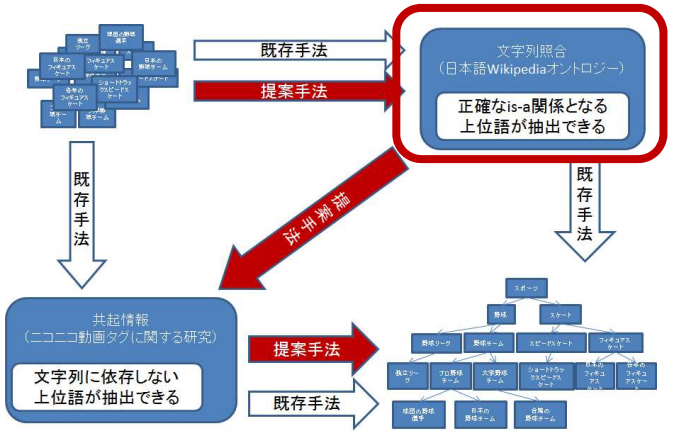
大学の**野球チーム**

プロ**野球チーム**



親カテゴリ名と子カテゴリ名を比較し，子カテゴリ名が
"任意の文字列+親カテゴリ名"となっているものを抽出する

形態素解析



[上位語を求めたいタグ]

野球選手 **名** で 歌っ て み た

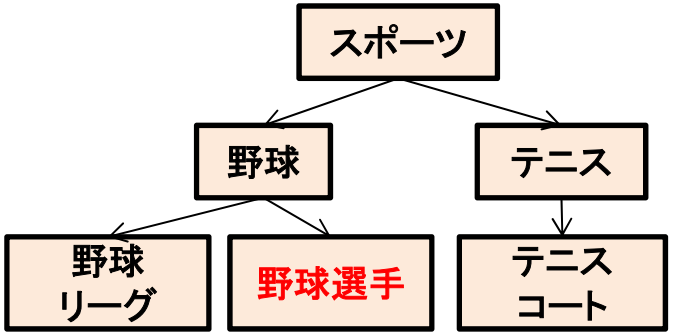
[名詞集合]

野球選手

名

各タグとの編集距離を求める

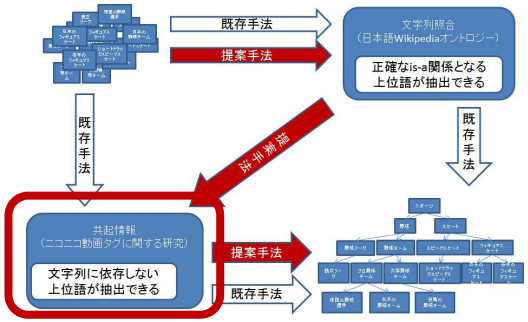
各タグとの編集距離を求める



編集距離
最小

[上位語] **野球選手**

共起頻度計算



□ ∈{フットサル, 選手, ゴール, サッカー, 卓球}

□ ∈{フットサル, サッカー, 練習, スキー, 柔道}

□ ∈{フットサル, ゴルフ, サッカー, ゴール, 水泳}

□ ∈{フットサル, サッカー, ボール, スケート, マラソン}

□ ∈{フットサル, ゴール, コート, 柔道, 選手}

共起回数は？		フットサルの上位概念候補
フットサル と サッカー	4	(1)サッカー
フットサル と ゴール	3	(2)ゴール
フットサル と 選手	2	(3)選手

制約

- Confidence
  - $confidence(\textbf{野球} \rightarrow スポーツ) = \frac{P(\textbf{野球} \rightarrow スポーツ)}{P(\textbf{野球})}$
- Lift
  - $lift(\textbf{野球} \rightarrow スポーツ) = \frac{P(\textbf{野球} \rightarrow スポーツ)}{P(\textbf{野球}) \cdot P(スポーツ)}$
- Conviction
  - $conviction(\textbf{野球} \rightarrow スポーツ) = \frac{1 - support(\textbf{スポーツ})}{1 - confidence(\textbf{野球} \rightarrow スポーツ)}$

実験

- データセット
  - 国立情報学研究所提供のニコニコデータセット
    - スポーツ関連の動画データを抽出して用いる
- 性能評価実験
  - 文字列照合のみ
  - 共起頻度のみ
  - 提案手法（文字列照合＋共起頻度）

- 対象としたタグ
  - スポーツ関連タグ
  - 上位語を含めたwikipedia登録語

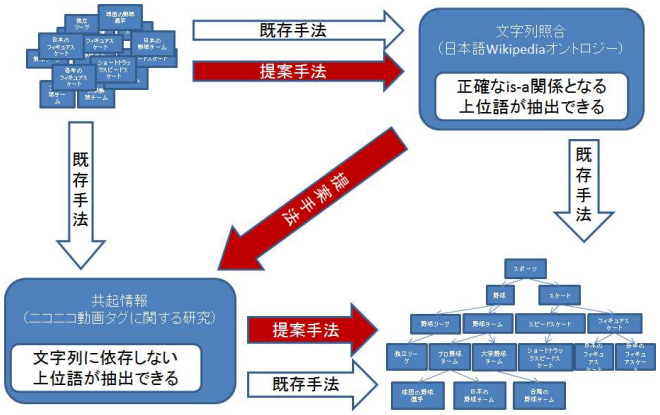
- 上位語の正誤判定
  - wikipediaオントロジーに従う

	データセット
スポーツ関連タグの種類数	134,763
1動画あたりの平均タグ数	6.4
上位語wikipedia登録タグ数（実験対象タグ）	927(0.7%)

評価実験結果

文字列照合

	top1
正解タグ数	197
再現率	21.3%



共起頻度（1）と提案手法（2）

	正解タグ数	再現率（1）	正確性（1）	再現率（2）
制限なし	67	7.23%		24.60%
lift	71	7.66%	16.86%	24.70%
conf	0	0.00%		21.25%
conv	12	1.29%	13.48%	22.44%
lift+cf	0	0.00%		21.25%
lift+cv	12	1.29%	13.48%	22.44%
cf+cv	0	0.00%		21.25%
lift+cf+cv	0	0.00%		21.25%

まとめ

- 結果
  - 実験では既存の共起情報のみを用いた手法より、精度が向上したものの、実用的なオントロジーとは言えない結果にとどまり、多くの課題が残った。
- 今後の課題
  - 日本語wordnetと日本語wikipediaオントロジーの統合
  - ニコニコ大辞典の活用
  - 下位語の特定によるオントロジー構築
    - 上位語下位語の特定による、中間概念の特定