セマンティック情報を用いた情報検索システム -意味を考慮した検索システム-

欅 惇志 (東京工業大学)

検索システムの利用は社会活動の一部

- 本日のイベント参加に関連する検索の例
  - ACT-I先端研究フォーラムとは？
  - 過去のフォーラムの様子は？
  - 今回の発表者・課題名は？
  - 日本科学未来館はどこ？/乗り換え方法は？

検索質問 (クエリ) と文書の照合の問題

- 自然言語には曖昧性が存在

ジャガー

動物

ジャガー
ジャガー(Jaguar)という名前は南アメリカインディアン語で「一突きで殺す者」を意味する。

車メーカー

ジャガー (自動車)
ジャガー（Jaguar Cars）は、イギリスの高級車メーカーである。

語の共起を手掛かりに意図推定
(従って 1 語のクエリでは困難)

本課題における解決方法 (のイメージ)

- 関連語による重要度推定

ジャガー　餌

ジャガー
ジャガー(Jaguar)という名前は南アメリカインディアン語で「一突きで殺す者」を意味する。

関連語を含む

○○動物園
ネコ科動物としては珍しく泳ぎが得意な動物である
ジャガーはカピバラ、カメ、ワニなどを捕食します。

- クエリ語の意味 (語義) 特定

ジャガー　餌

語義候補：動物 / 車メーカー
語義候補：動物に与える食べ物 / 人を誘惑する手段

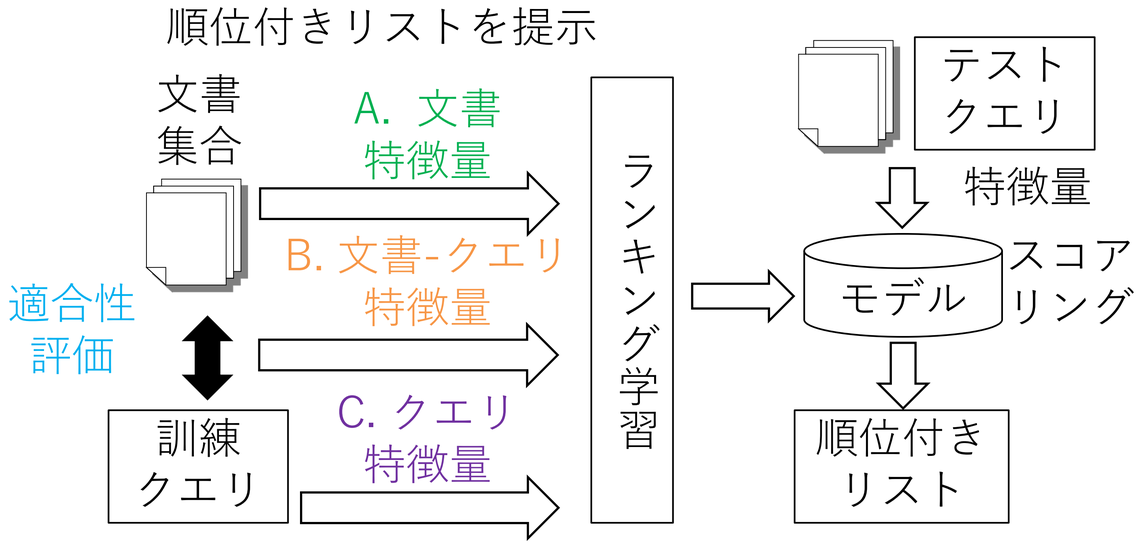
ジャガー (動物) × 餌 (動物に与える食べ物) の確率が最も高い

意味を考慮×ランキング学習×Web 検索

- 意味を考慮した検索
  - 局所的な意味情報：語義曖昧性解消
  - 大域的な意味情報：分散表現利用
- ランキング学習 (learning to rank)
  - (半) 教師あり学習
- Web 検索
  - 未知のクエリが頻繁に発行される
    - 学習データの存在を仮定できない

ランキング学習の概要

- 手順
  1. 文書集合，訓練クエリ，適合性評価からモデル構築
  2. (テスト) クエリに対して文書のスコアリング，順位付きリストを提示



- 主要な特徴量
  - A. 文書長，ページランク，Doc2vec
  - B. TF-IDF，BM25，**語義スコア**，**分散表現スコア**
  - C. クエリ語，クエリ長，Query2vec

	クエリ	文書	評価	クエリ長	文書長	BM25
訓練	1	1000	2	4	500	0.4
訓練	1	1001	0	4	700	0.6
テスト	2	2000	1	3	600	0.3
テスト	2	2001	1	3	400	0.5

語義の曖昧性解消

- 語義の付与
  - ツール: IMS [1]
    - 単語に対して最大 10 種類の語義を付与
      - 周辺語と文法ルール利用
    - 入力 (文): “Water heaters are necessary”
    - 出力 (語義付き語とその確率): water#4|0.63, water#1|0.34, water#2|0.23
  - 文書に対する語義付与
    - 文単位に分割し IMS 適用
  - クエリに対する語義付与
    - 自然言語文法に基づかないため直接 IMS 適用不適切
    - 説明文WSD：クエリの説明文 (description) に IMS 適用
      - 説明文にクエリ語が含まれるため適切な文脈情報取得
      - 一般的な Web 検索では説明文を仮定できない
    - 共起WSD [2]：文法情報を利用せず語義付与
      - クエリ語の共起のみに着目
- 語義スコア
  - クエリと文書の語義を比較
  - 全語義：付与された全ての語義を考慮
  - 単一語義：最も高い確率を付与された語義を考慮

[1] Zhong+: “It Makes Sense: A Wide- Coverage Word Sense Disambiguation System for Free Text”, in Proc. of ACL 2010.
[2] Keyaki+: “Part-of-speech Tagging for Web Search Queries Using a Large-scale Web Corpus”, in Proc. of SAC 2017.

	nDCG@20
説明文WSD	**.341**
共起WSD	.289

分散表現

- 分散表現として表現
  - 文書：doc2vec，クエリ：query2vec
- 分散表現スコア
  - 文書とクエリのベクトル差，距離，cos 類似度

実験結果

- データセット
  - ClueWeb09-b
  - TREC Web track topics
- 4 分割交差検定
  - 訓練 150 クエリ
  - テスト 50 クエリ

*: $p < 0.05$, **: $p < 0.001$

		nDCG@20
Baseline	BM25	.267
Baseline	Common features	.334
語義スコア	全語義	.341*
語義スコア	単一語義	.333
分散表現スコア	ベクトル差	.338
分散表現スコア	ベクトル距離	.335
分散表現スコア	cos類似度	.341
	単一語義+cos類似度	**.349****