

ビッグデータ分析技術ワークショップ
～大規模グラフマイニング技術と応用～

- 日時：2017年3月5日 (日) 14:00～19:30
- 場所：高山市民文化会館 3F 講堂

LOD の応用とグラフマイニングを用いた 分析手法の研究

若原 俊彦, 榎 俊孝

福岡工業大学大学院工学研究科

目次

1. はじめに

- 1.1. オープンデータ
- 1.2. 研究概要

2. Linked Data

- 2.1. Linked Dataの概要
- 2.2. LODの課題
- 2.3. RDF語彙

3. 潜在的リンクの推定

- 3.1. ラベル伝搬アルゴリズム
- 3.2. 評価実験

4. I-Scover LOD

- 4.1. I-Scoverの概要
- 4.2. SPARQL検索
- 4.3. 論文検索支援システム

5. おわりに

目次

1. はじめに

- 1.1. オープンデータ
- 1.2. 研究概要

2. Linked Data

- 2.1. Linked Dataの概要
- 2.2. LODの課題
- 2.3. RDF語彙

3. 潜在的リンクの推定

- 3.1. ラベル伝搬アルゴリズム
- 3.2. 評価実験

4. I-Scover LOD

- 4.1. I-Scoverの概要
- 4.2. SPARQL検索
- 4.3. 論文検索支援システム

5. おわりに

1.1. オープンデータ

□オープンデータとは

ウェブ上に公開された二次利用が可能なデータ。

人口
統計

AED
設置
場所

避難
施設

観光

文献

など。

□明日の日本を支える観光ビジョン

- 2007年：観光立国推進基本法
- 2012年：電子行政オープンデータ戦略

➡ 2016年：明日の日本を支える観光ビジョン

オープンデータによる**広域的**かつ**持続的**な観光事業の展開

1.2. 研究概要

□研究テーマ

LODの知識構造化手法と観光オントロジーの構築に向けた自治体CMS構成法の研究

□サブテーマ

- LODの潜在的リンクを推定するラベル伝搬アルゴリズム
- LODの自動構築を考慮した自治体CMS
- 観光オントロジーのための観光語彙基盤
- LODを用いた音声エージェントシステム
- 学術文献LODのグラフマイニングによる研究動向分析

1.2. 研究概要

□たのしんぐらプロジェクト

福岡県糟屋郡新宮町の観光振興のため，福岡工業大学情報工学部と福岡県糟屋郡新宮町産業振興課のメンバーから構成.



LOD Challenge 2016

アプリケーション部門優秀賞
LOD for 地方創生賞

応募名：観光語彙基盤を用いた新宮町LODの作成と意味検索システムの開発

LODを知識ベースとして利用するための設計や応用

高妻神社の特徴と行き方を教えてください。



[検索例を見る](#)



概要

相島の中央の船原という小高い丘にある小さなお社で正式名は高妻権現社、通称「権現様」と呼ばれていました。島の人たちは、この神社を高神（位の高い神）として信仰しています。祭神の彦火火出見尊（ひこほほでみのみこと）は気性の荒い神様で、気に入らないことがあるとひどい神罰が下るといわれ、かつては参拝にもいくつかの禁忌が設けられていました。また、権現森全域が神山とされ、樹木はすべて神木であり伐採すると神罰が下るとい言い伝えも残されており、権現森はいまだに原始林の様態を保っているようです。

アクセス

高妻神社は、相島渡船待合所から海沿いを北へ向かい相島分校付近に石段があり登った先に

人気記事

- > SPARQL
1,916 views | posted on 2016年11月17日
- > しんぐうコンシェル...
1,874 views | posted on 2016年11月17日
- > 丸山食堂
1,867 views | posted on 2016年9月23日
- > 相島について
1,663 views | posted on 2016年9月20日
- > お地藏様
1,659 views | posted on 2016年10月24日
- > 新宮可宮渡船「しんぐう」...
1,628 views | posted on 2016年9月6日
- > ひとまるの里
1,534 views | posted on 2016年11月10日
- > 若宮神社
1,532 views | posted on 2016年9月6日
- > 沖田中央公園
1,413 views | posted on 2016年10月6日
- > 相島観光案内所
1,390 views | posted on 2016年10月6日

最新記事

- > 立花山日曜市
- > ひとまるの里
- > コミュニティバス「マリックス」
- > 相島きずな館
- > お地藏様

丸山食堂の近くにある神社は？



[検索例を見る](#)

該当スポット

1. 岩宮神社
2. 恵比須神社
3. 若宮神社
4. 金比羅神社
5. 高妻神社

目次

1. はじめに

- 1.1. オープンデータ
- 1.2. 研究概要

2. Linked Data

- 2.1. Linked Dataの概要
- 2.2. LODの課題
- 2.3. RDF語彙

3. 潜在的リンクの推定

- 3.1. ラベル伝搬アルゴリズム
- 3.2. 評価実験

4. I-Scover LOD

- 4.1. I-Scoverの概要
- 4.2. SPARQL検索
- 4.3. 論文検索支援システム

5. おわりに

2.1. Linked Dataの概要

オープンデータの公開レベル

| 公開レベル | 条件 | データの例 |
|-------|--------------------|--------------------|
| ★ | オープンライセンスで公開. | PDF, JPG, MP3, ZIP |
| ★★ | 構造化データ | XLS, XLSX |
| ★★★ | オープンフォーマット | CSV, TXT, XML |
| ★★★★ | URIで意味付けされた物事 | RDF |
| ★★★★★ | 外部リンクを含むLinked RDF | LOD |

(参考) “5-star Open Data”, <http://5stardata.info>, (March 2017)

2.1. Linked Dataの概要

□ Linked Dataとは

- RDFに基づいて主語, 述語, 目的語の3要素で表現されたラベル付き有向グラフ.
- ウェブ上にオープンライセンスで公開されたLinked DataをLinked Open Data (LOD) という.



<http://www.tanoshingu.org/>新宮町

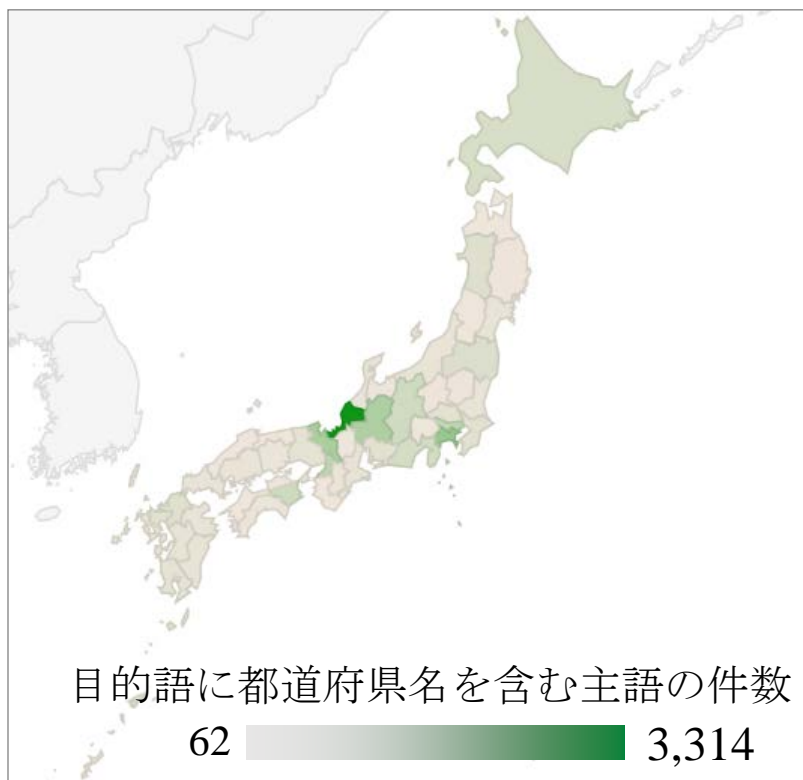
<http://www.tourism.property/>#観光地

<http://www.tanoshingu.org/>相島積石塚郡

RDFモデル

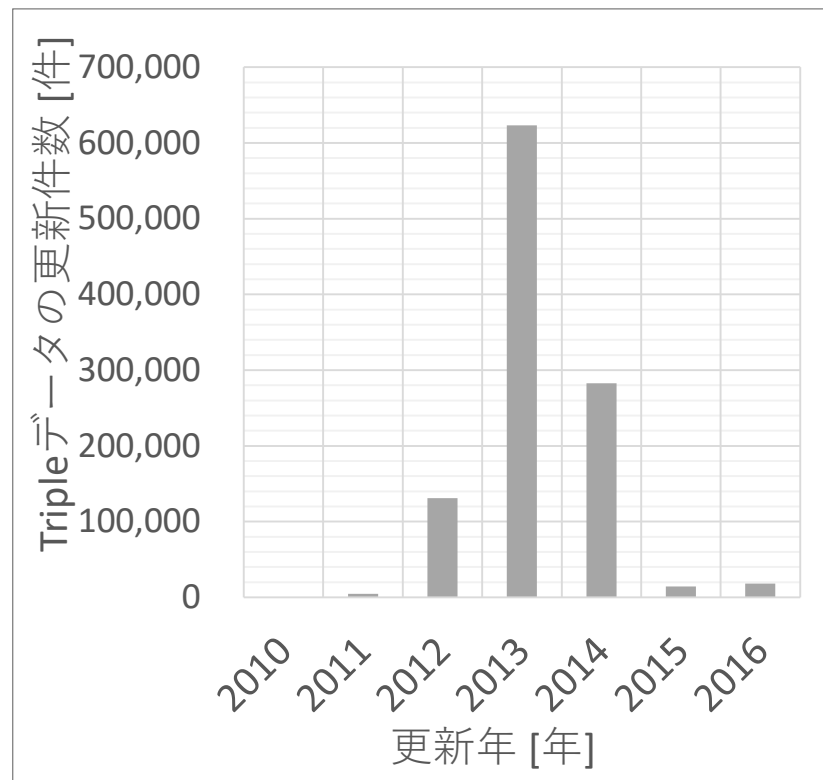
2.2. LODの課題

□LODの公開に積極的な地域が限定的



推定される地域別LODの公開状況

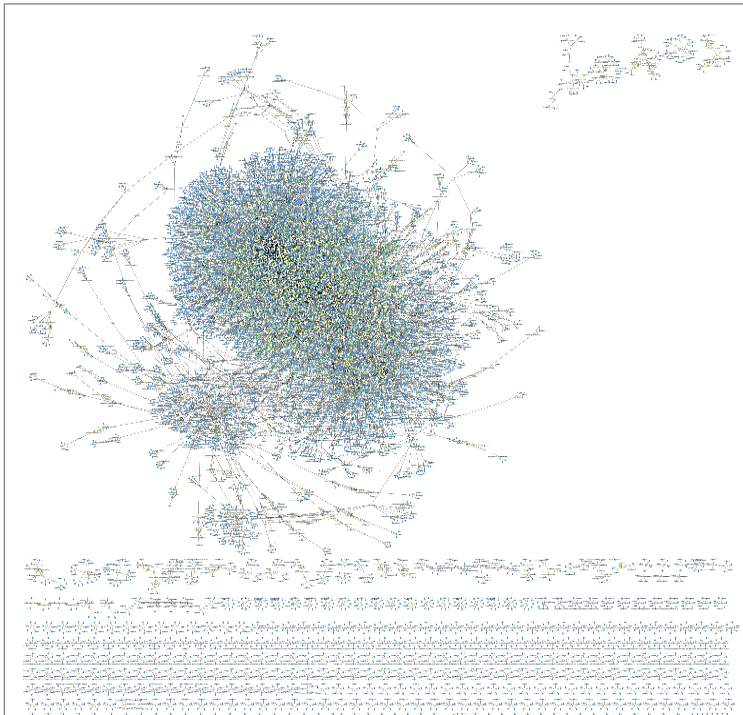
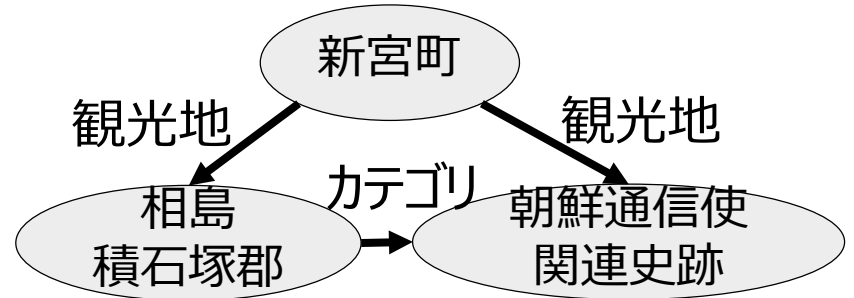
(参考) 榎 俊孝, 若原俊彦, 若橋和生, 小舘亮之, 小林 透, 曾根原 登, “LODの汎用化を図るメタデータの設定手法”, 信学技報, vol. 116, no. 488, LOIS2016-82, pp. 111-116, March 2017.



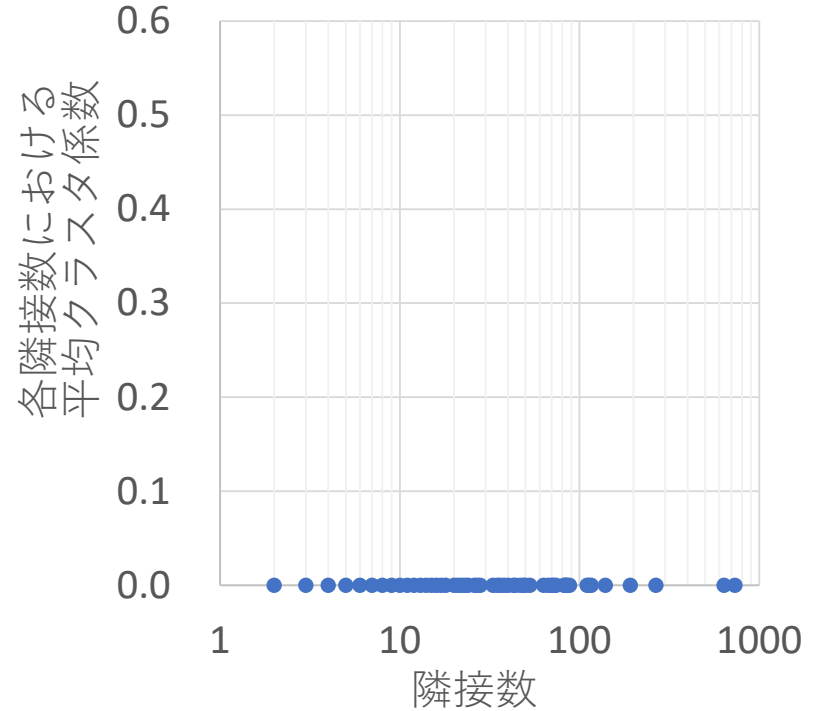
地域関連LODの更新状況

2.2. LODの課題

LODのグラフ密度が低い



2014年に更新された
地域関連LODのグラフ構造



2014年に更新された
地域関連LODのクラスタ係数の分布

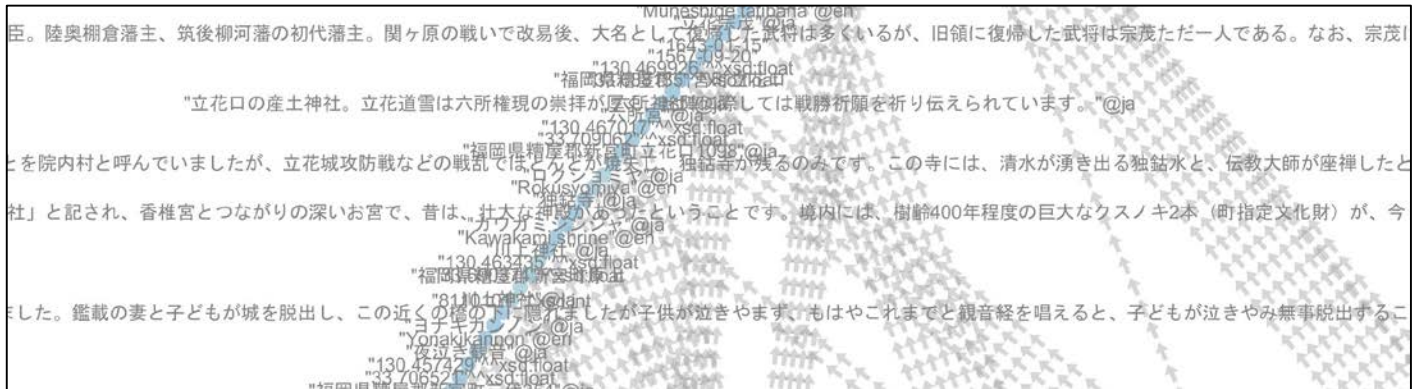
(参考) 榎 俊孝, 若原俊彦, 若橋和生, 小舘亮之, 小林 透, 曾根原 登, “LODの汎用化を図るメタデータの設定手法”, 信学技報, vol. 116, no. 488, LOIS2016-82, pp. 111-116, March 2017.

2.2. LODの課題

データ型の例とLinkData.orgにおける使用率

| 型名 | データ型 | 例 | 使用率 |
|------|--------------|---|-------|
| 文字列型 | xsd:string | “沖縄県”@ja | 60.3% |
| 整数型 | xsd:int | “9070004”^^xsd:int | 14.1% |
| 実数型 | xsd:float | “130.423”^^xsd:float | 10.3% |
| 日付型 | xsd:date | “2017-03-02”^^xsd:date | 0.0% |
| 日時型 | xsd:dateTime | “2017-03-02T16:50”^^xsd:dateTime | 0.0% |
| 年型 | xsd:gYear | “2017”^^xsd:gYear | 0.0% |
| URI型 | xsd:anyURI | < http://www.tanoshingu.org/相島 > | 0.3% |

3.2. ID管理によるリンク化

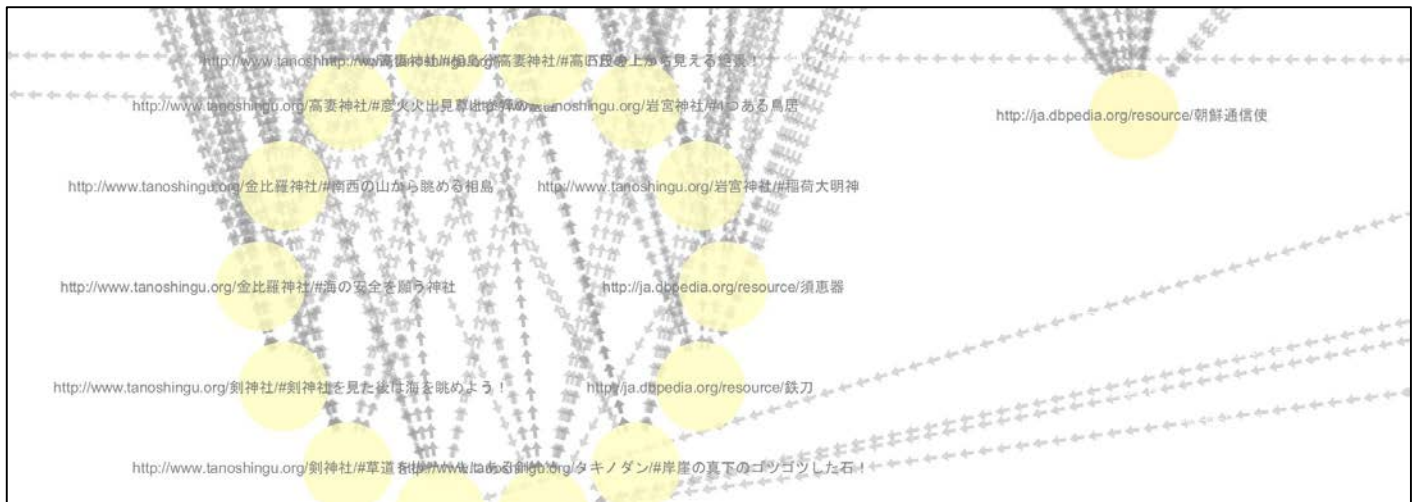


臣。陸奥棚倉藩主、筑後柳河藩の初代藩主。関ヶ原の戦いで改易後、大名として復帰した武将は多くいるが、旧領に復帰した武将は宗茂ただ一人である。なお、宗茂は立花口の産土神社。立花道雪は六所権現の崇拝が厚く、新神祇院の神として戦勝祈願を祈り伝えられています。@ja

とを院内村と呼んでいましたが、立花城攻防戦などの戦乱では、口とが壊滅し、独結寺が残るのみです。この寺には、清水が湧き出る独結水と、伝教大師が座禅したと社」と記され、香椎宮とつながりの深いお宮で、昔は、大きな神宮があったということです。境内には、樹齢400年程度の巨大なクスノキ2本（町指定文化財）が、今

した。鑑載の妻と子どもが城を脱出し、この近くの橋の下に隠れていたが子供が泣きやまず、もはやこれまでと親音経を唱え、子どもが泣きやみ無事脱出するこ

文字列型 (xsd:string) : 入リンクのみ



URI型 (xsd:anyURI) : 入リンク + 出リンク

2.3. RDF語彙

標準化されたRDF語彙の例

| Schemaの名前空間 | Prefix | Propertyの例 |
|---|---------|-----------------------|
| http://www.w3.org/2000/01/rdf-schema# | rdfs | comment, label. |
| http://www.w3.org/2002/07/owl# | owl | Annotation, sameAs. |
| http://purl.org/dc/elements/1.1/ | dcterms | title, description. |
| http://www.w3.org/2001/XMLSchema# | xsd | int, float, etc. |
| http://www.w3.org/2004/02/skos/core# | skos | broader, narrower. |
| http://xmlns.com/foaf/0.1/ | foaf | Organization, Person. |
| http://www.w3.org/2003/01/geo/wgs84_pos# | geo | lat, long. |
| http://imi.ipa.go.jp/ns/core/rdf# | ic | 名称, 人数, 期間. |

2.3. RDF語彙

IPAの共通語彙基盤に定義されている述語の例（位置）

| 識別子 | 項目名 | 値型 | 回数 |
|-----------|--------|------------|------|
| ic:座標参照系 | 座標参照系 | ic:ID型 | 0..1 |
| ic:緯度経度書式 | 緯度経度書式 | xsd:string | 0..1 |
| ic:緯度 | 緯度 | xsd:string | 0..1 |
| ic:経度 | 経度 | xsd:string | 0..1 |
| ic:座標データ | 座標データ | xsd:string | 0..1 |

IPAの共通語彙基盤に定義されている述語の例（年月日）

| 識別子 | 項目名 | 値型 | 回数 |
|----------|-------|-------------|------|
| ic:標準型日付 | 標準型日付 | xsd:date | 0..1 |
| ic:年号 | 年号 | xsd:string | 0..1 |
| ic:年 | 年 | xsd:integer | 0..1 |
| ic:月 | 月 | xsd:integer | 0..1 |
| ic:日 | 日 | xsd:integer | 0..1 |

(参考) 情報処理推進機構 共通語彙基盤整備事業, “共通語彙基盤”, <http://imi.go.jp/ns/core/Core232.html>, (March 2017)

目次

1. はじめに

- 1.1. オープンデータ
- 1.2. 研究概要

2. Linked Data

- 2.1. Linked Dataの概要
- 2.2. LODの課題
- 2.3. RDF語彙

3. 潜在的リンクの推定

- 3.1. ラベル伝搬アルゴリズム
- 3.2. 評価実験

4. I-Scover LOD

- 4.1. I-Scoverの概要
- 4.2. SPARQL検索
- 4.3. 論文検索支援システム

5. おわりに

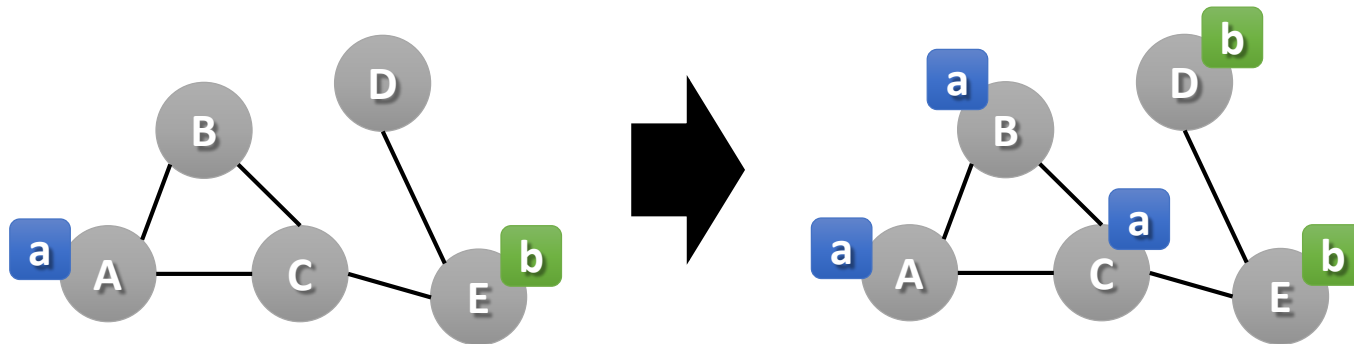
3.1. ラベル伝搬アルゴリズム

□目的

LODの潜在的なリンクを推定し, LODを知識ベース化.

□ラベル伝搬アルゴリズム

「グラフ上の隣接ノードは同じクラスに属する」という仮定の下でノードにラベルを付与する半教師あり学習.

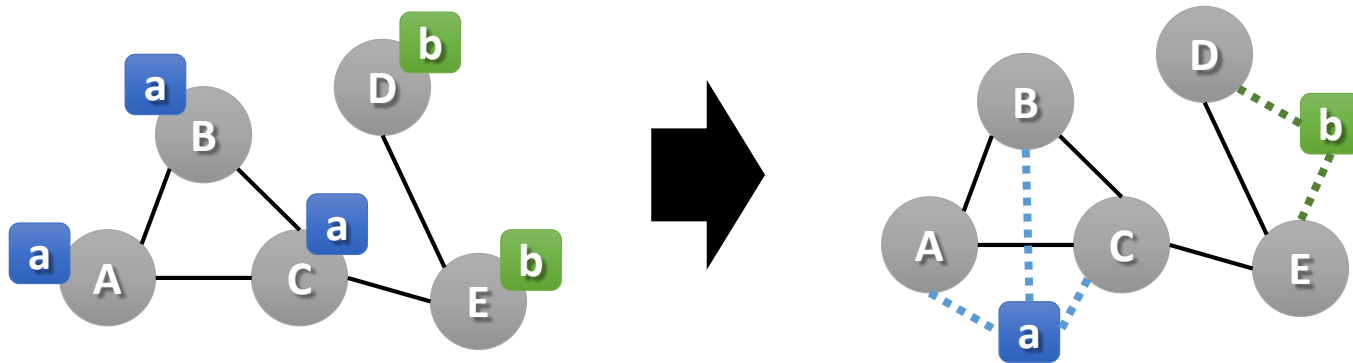


ラベル伝搬アルゴリズムの概略図

3.1. ラベル伝搬アルゴリズム

□ LODの潜在的リンクの推定

- 付与されたラベルを新しいノードとして定義.
- 教師データのプロパティに基づいてデータ型を決定.
- 教師データのプロパティに基づいてリンク重みを決定.



ラベル伝搬アルゴリズムの概略図

3.1. ラベル伝搬アルゴリズム

□LODの潜在的リンクの推定

- 連立方程式を解かず，ラベル単位でリンクを推定（高速）。
- 並立処理，マルチラベル。
- ラベル推定値を意味距離（0.0 ~ 1.0）として評価。

ラベル更新式

$$\begin{aligned} & - \text{deg}_t > 1 \\ & v_k = \varepsilon \frac{w_{t,k} \cdot v_t}{\sqrt{\text{deg}_t - 1}} \end{aligned}$$

$$- \text{deg}_t = 0$$

$$v_k = \varepsilon \cdot w_{t,k} \cdot v_t$$

伝搬定数

$$\varepsilon = \frac{n - 1}{n}$$

$w_{i,j}$: エッジ重み

v_t : ノード t のラベル推定値

v_k : ノード k のラベル推定値

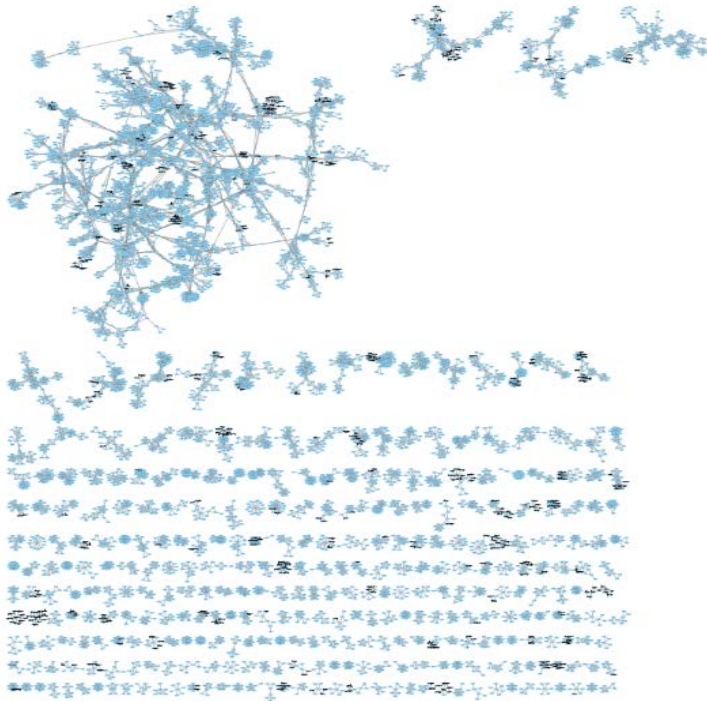
deg_t : ノード t の次数

n : ノード数

(詳細は省略)

3.2. 評価実験

電子情報通信学会の文献検索システムI-DiscoverのLinked Dataを用いて提案手法の処理時間とラベル正解率を評価する.

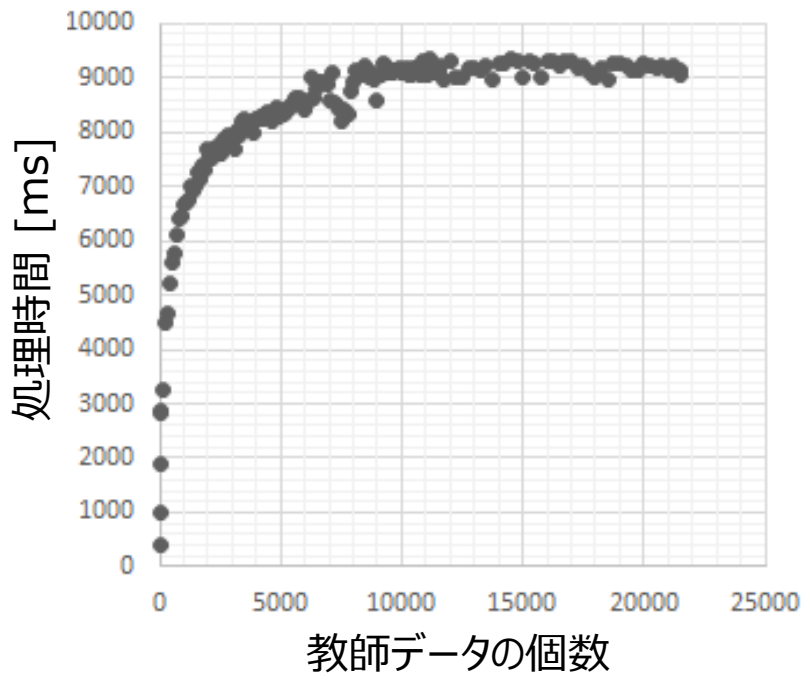


2012年IEICE投稿論文の著者グラフ構造

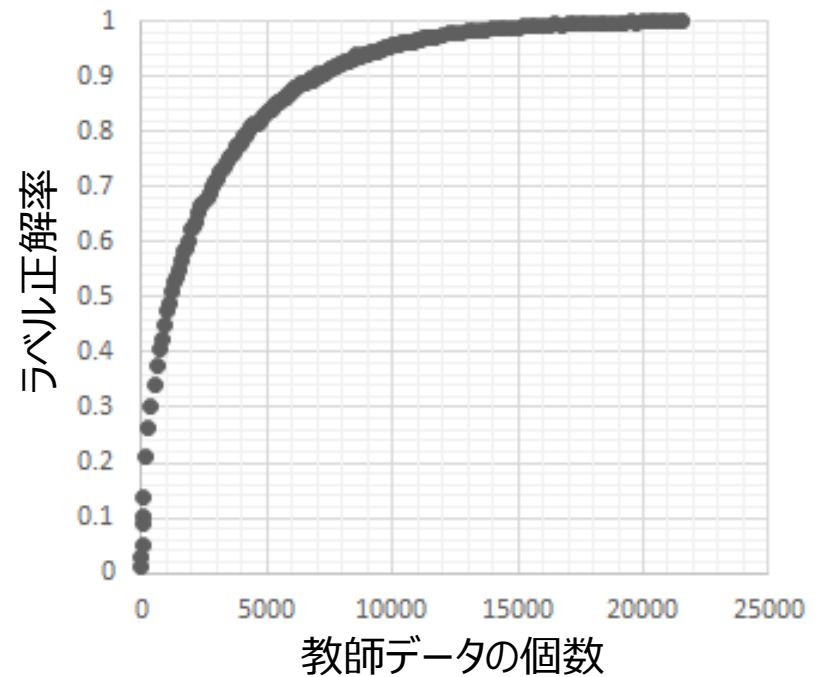
- ノード数 : 11,580
- エッジ数 : 28,992
- ラベル数: 21,522
ラベル = 研究会名.

下限値を0.1とし, 教師データの付与数を変化させ, ランダムにラベルをノードに与えて評価.

3.2. 評価実験



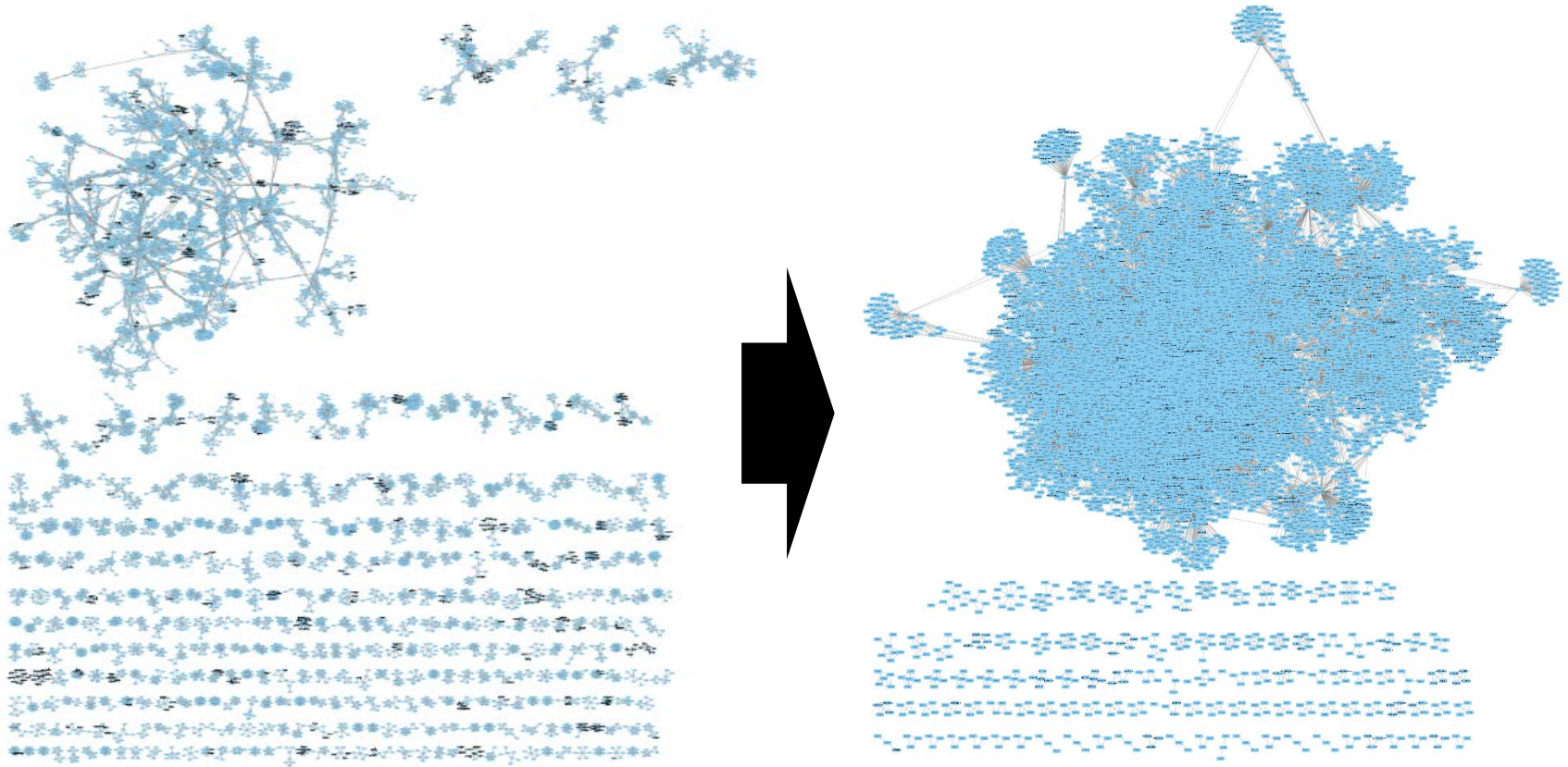
提案手法の処理時間



提案手法のラベル正解率

教師データ数が21,522個中4300個（約20.0%）でラベル正解率は80%を超える。

3.2. 評価実験



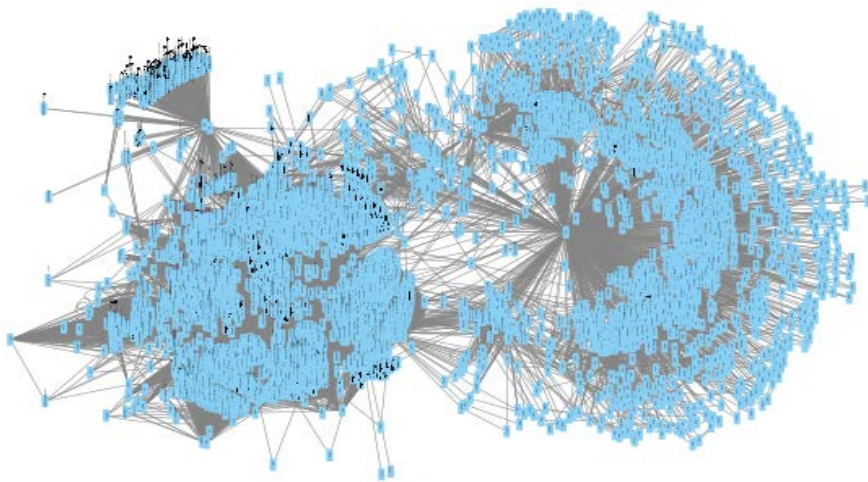
2012年IEICE投稿論文の著者グラフの構造

提案手法により拡張された
著者グラフの構造

教師データ数が4,300個, ラベル推定値が0.9以上のラベルを結合.

3.2. 評価実験

2016年に、LinkData.orgに登録されたLODからランダムに20,000件のTripleデータを取得.

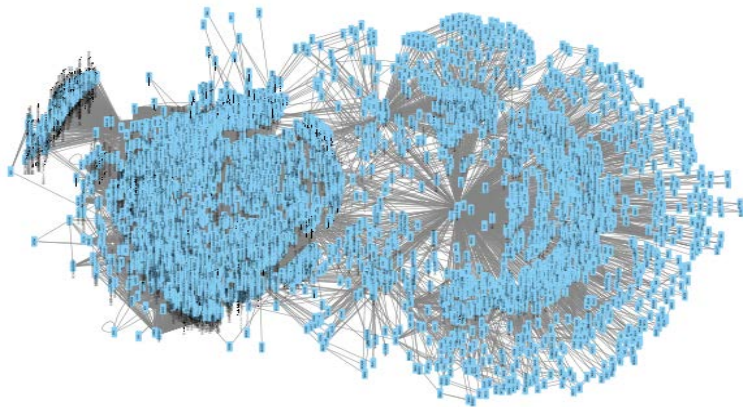


- ノード数 : 4,072
- エッジ数 : 16,029
- ラベル種類数: 15
ラベル = カテゴリの語彙.

下限値を0.1とし,
教師データのラベルを
50ノードに与えて評価.

LinkData.orgのLODから作成したグラフ

3.2. 評価実験



| | | |
|-----|-------------|--------------|
| 推定前 | 4,072 nodes | 16,029 edges |
|-----|-------------|--------------|

| | | |
|-----|-------------|--------------|
| 推定後 | 4,087 nodes | 20,192 edges |
|-----|-------------|--------------|

推定されたラベルの例

| ノード名 | ラベル名 | ラベル推定値 (意味距離) |
|------------|------|------------------|
| AED | 教育 | 0.999 |
| | 福祉 | 0.543 |
| | 病院 | 0.489 |
| 公園 | スポーツ | 0.843 |
| | 学校 | 1.000 |
| 学校 | 教育 | 0.999 |
| | 避難施設 | 0.588 |
| | 福祉 | 0.529 |

目次

1. はじめに

- 1.1. オープンデータ
- 1.2. 研究概要

2. Linked Data

- 2.1. Linked Dataの概要
- 2.2. LODの課題
- 2.3. RDF語彙

3. 潜在的リンクの推定

- 3.1. ラベル伝搬アルゴリズム
- 3.2. 評価実験

4. I-Scover LOD

- 4.1. I-Scoverの概要
- 4.2. SPARQL検索
- 4.3. 論文検索支援システム

5. おわりに

4.1. I-Discoverの概要

□I-Discoverとは

- 電子情報通信学会が運営する文献検索システム.
- 文献メタデータをLinked Dataの形式で管理.
- 2016年12月26日から第2期システムの仮運用を開始.



I-Discoverのトップ画面

<http://i-scover-api.ieice.org/iscover/api/sparql>

SPARQLにより自由に
文献データを取り扱う
ことが可能に.

キーワードによる検索



ライフログ

キーワード

詳細情報

JPN/ENG

| | |
|-------------|--|
| 名前 | ライフログ |
| 解説 | 人間の生活・行動・体験 (Life) などを、画像・音声・テキスト・位置情報などのデジタルデータとして記録したものであり、記録する際に付加されるメタデータなどの情報も含む。 |
| 表示回数 | 21 |
| 付与された文献等の件数 | 217 |
| 備考 | 監修：ライフインテリジェンスとオフィス情報システム(LOIS) 研究専門委員会 |

更新日時 2016-12-09 08:52

出力形式 (UTF-8) : CSV XML BibTex [ダウンロード](#)

関連メタデータ

メタデータ種別: [選択](#)

このキーワードを持つ文献に付与されている他のキーワード (両キーワードがともに付与されている文献数 (関連度) の大きい順)

 **キーワード** 全823件 [一覧表示](#)

ENG [lifelog](#)

表示回数: 2
付与された文献等の件数: 75
関連度: 70

ENG [life-log](#)

表示回数: 1
付与された文献等の件数: 27
関連度: 24

JPN [スマートフォン](#)

表示回数: 3
付与された文献等の件数: 378
関連度: 13

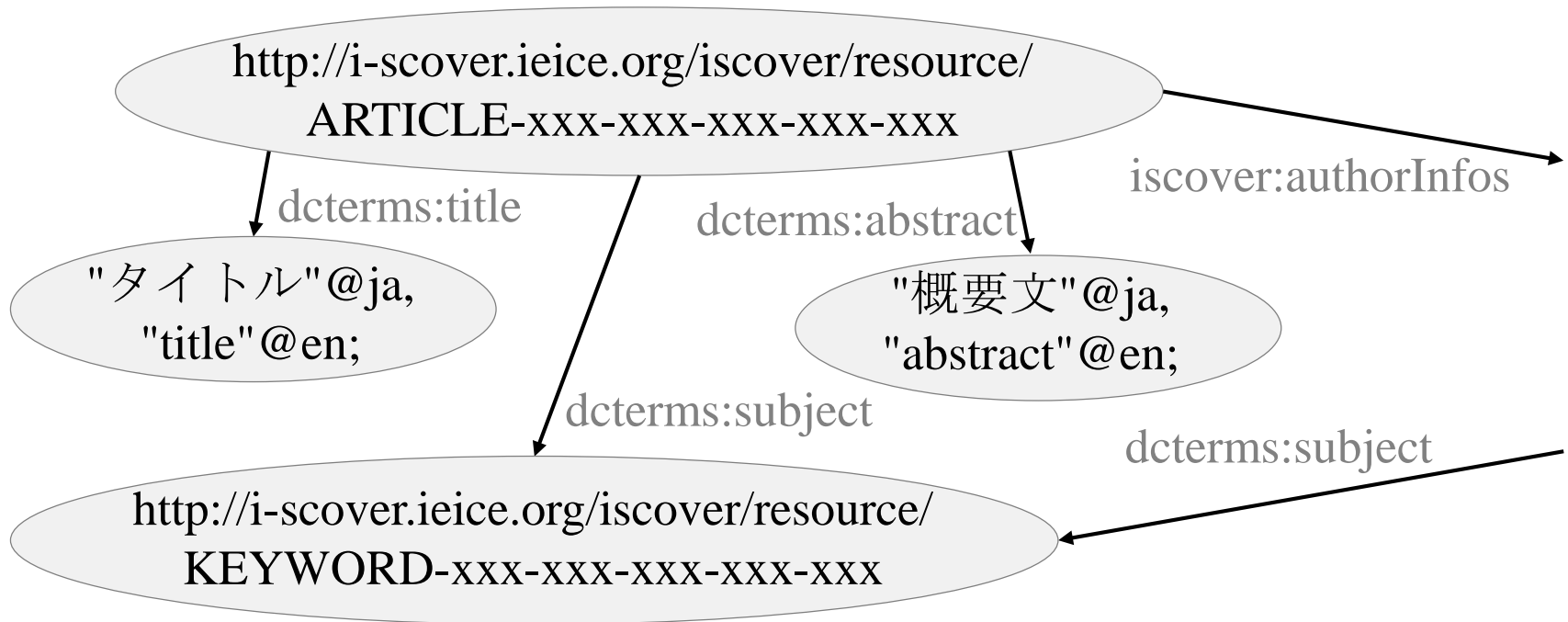
関連文献

このキーワードが付与されている文献

 **文献** 全217件 [一覧表示](#)

(参考) I-Scover, <http://i-scover.ieice.org/iscover/page/KEYWORD-006107127303>, (March 2017)

論文メタデータの構成



I-Scover LODにおけるデータ構造の一例

4.2. SPARQL検索

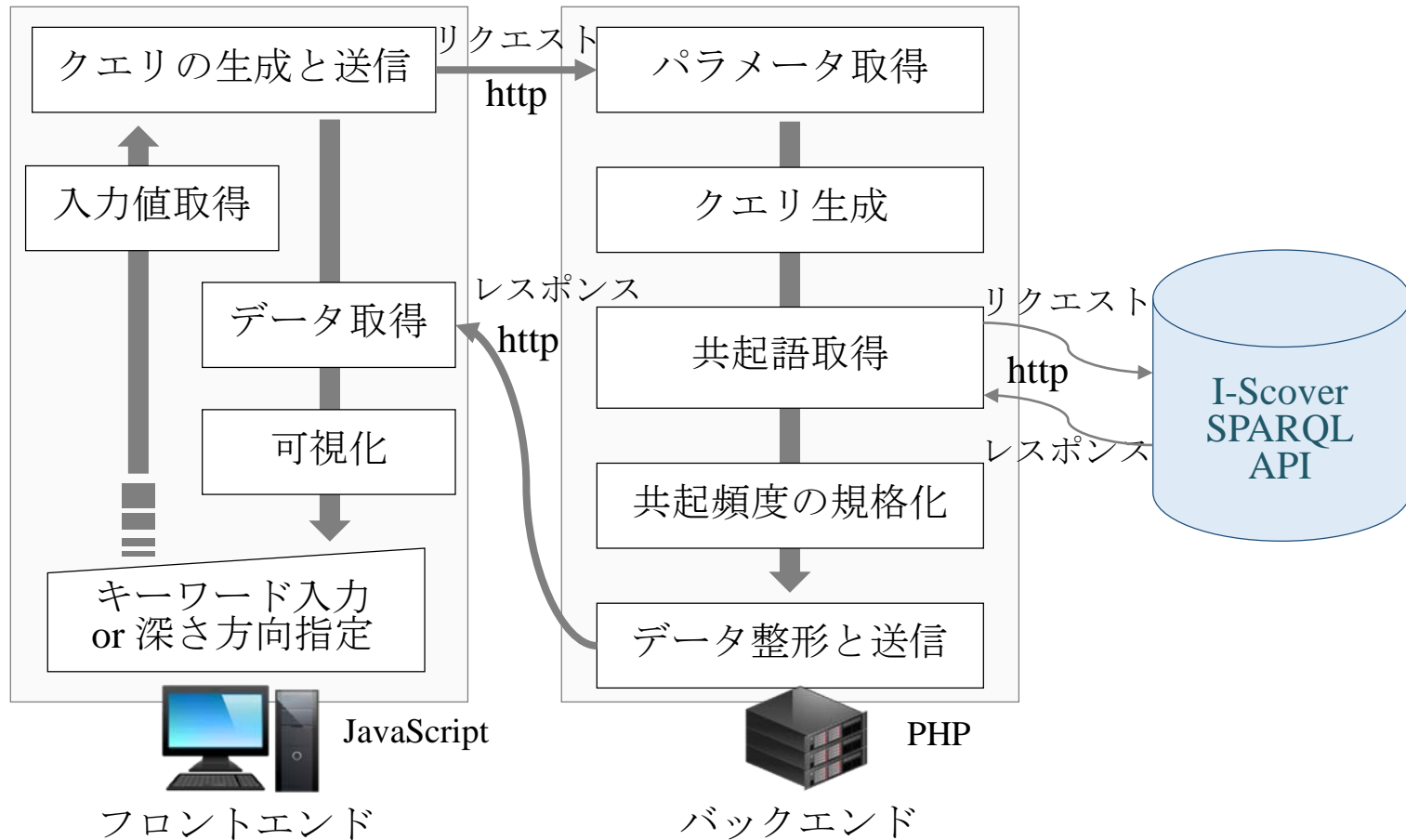
```
select ?term ?score where {  
  {  
    select ?termIRI (count(?termIRI) * ?all) as ?score {  
      {  
        select ?termIRI count(?termIRI) as ?all where {  
          ?articleIRI  
          dcterms:subject ?termIRI;  
          a iscover:Article.  
        }  
      }  
      ?articleIRI  
      dcterms:subject ?termIRI;  
      iscover:articleNumber ?number;  
      a iscover:Article.  
      filter(regex(?number, "LOIS"))  
    }  
  }  
  ?termIRI  
  rdfs:label ?term;  
  a iscover:Term.  
  filter(lang(?term) = "ja")  
}  
order by desc(?score)  
limit 20
```

LOIS研究会の研究トレンドを
調査するクエリの例

| "term", | "score" |
|--------------------|---------|
| "センサネットワーク (SN) ", | 22341 |
| "ライフログ", | 15190 |
| "FCAPS", | 10064 |
| "スマートフォン", | 9072 |
| "セキュリティ", | 7840 |
| "音声認識", | 6876 |
| "画像処理", | 6543 |
| "GPS", | 5238 |
| "認証", | 4169 |
| "可視化", | 4015 |
| "位置情報", | 3995 |
| "携帯電話", | 3896 |
| "クラスタリング", | 3684 |
| "無線LAN", | 3558 |
| "クラウドコンピューティング", | 3000 |
| "データマイニング", | 2817 |
| "機械学習", | 2632 |
| "ニューラルネットワーク", | 2214 |
| "パターン認識", | 2058 |
| "電子透かし", | 1980 |

クエリの実行結果

4.3. 論文検索支援システム



NTT-SICとの共同提案 (I-Scover利活用コンテスト)

4.3. 論文検索支援システム

The screenshot shows a search interface with a search bar at the top containing the text "グラフ". Below the search bar, there are three search results displayed in a list format. To the right of the search results, there is a graph visualization with a central node labeled "グラフ" and five surrounding nodes: "ラベリング", "平面グラフ", "列挙", "アルゴリズム", and "描画".

Search results:

- A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs**
Karypis, George, Kumar, Vipin
fill-reducing orderings, finite element computations, graph partitioning, parallel computations
1998
- An Improved Spectral Graph Partitioning Algorithm for Mapping Parallel Computations**
Hendrickson, Bruce, Leland, Robert
eigenvector, graph partitioning, graph spectrum, load balancing, parallel computations
1995
- Programming in JoCaml**
Mandel, Louis, Maranget, Luc
concurrency, distributed programming, parallel computations, functional programming
2007

Graph visualization:

```
graph TD; G(グラフ) --- L(ラベリング); G --- PG(平面グラフ); G --- R(列挙); G --- A(アルゴリズム); G --- H(描画);
```

入力インタフェースの概観

目次

1. はじめに

- 1.1. オープンデータ
- 1.2. 研究概要

2. Linked Data

- 2.1. Linked Dataの概要
- 2.2. LODの課題
- 2.3. RDF語彙

3. 潜在的リンクの推定

- 3.1. ラベル伝搬アルゴリズム
- 3.2. 評価実験

4. I-Scover LOD

- 4.1. I-Scoverの概要
- 4.2. SPARQL検索
- 4.3. 論文検索支援システム

5. おわりに

5. おわりに

- 従来は、リソースを文字列型で定義することが多かったため横断的なリンク構造になっていなかった。
- 観光分野の語彙をURI (xsd:anyURI) で定義することにより出リンクが可能となり、リンク構造が改善された。
- ノード間のリンク構造を改善するため、新しいラベル伝搬アルゴリズムを導入し、潜在的リンクを推定した。
- I-Discover SPARQL APIを用いて共起語を抽出してグラフを作成し、可視化による論文検索支援システムを実現した。

ご清聴ありがとうございました。