【I3-2】

グラフ構造を利用したエンティティ検索

駒水　孝裕
（名古屋大学）

LOD
(Linked Open Data)

- オープンデータの公開方式
  - Tim Berners-Lee の提唱
  - 5★オープンデータ (https://5stardata.info)
  - URI, HTTP, RDF, SPARQL, 相互接続
- LODの活発化
  - LOD Cloud (https://lod-cloud.net/, 2018年6月)
    - 1,220 データセット
      - 1,000個のトリプルがないと認められない
    - 16,095 個のデータセット間リンク
      - 50個のデータセットとリンクしていないと認められない

RDF
(Resource Description Framework)

- 三つ組でデータを記述する枠組み
  - 三つ組：(主語，述語，目的語)

    - dbr := http://dbpedia.org/resource
    - rdfs := https://www.w3.org/2000/01/rdf-schema#

  - 主語，述語は必ず URI
  - 目的語は URI または リテラル（文字列）
- URI (Uniform Resource Identifier)

エンティティ検索

- LODにおける基本的な検索
  - 入力：キーワードクエリ（単語集合）
  - 出力：エンティティ集合
- 問題：エンティティの文書表現
  - どのリテラルがエンティティに「関連」するか？
    - RDFで記述 → データがグラフ構造
  - ナイーブ：エンティティが主語のトリプルを利用
  - 既存手法
    - 単語重み：BM25, BM25-CA, LM, SDM, PRMS, MLM-all
    - フィールド拡張：MLM-CA, FSDM, BM25F-CA
    - Entity Linking：LM-ELR, SDM-ELR, FSDM-ELR

本研究の貢献

1. 再現率の調査
2. RWRDoc: グラフ上の距離を用いた表現学習
   - グラフ上の「近さ」に応じてリテラルを
     エンティティの表現に組み込む
3. PPRSD: グラフ分析に基づく再ランキング
   - ランキング性能 (NDCG) の向上

再現率の調査

- 対象データ：
  DBpedia-Entity v2
- 再現率@k
  - @10 vs. @1000
    ➔ 58% ダウン
  - @100 vs. @1000
    ➔ 18% ダウン
  - @1000
    ➔ 13% の見落とし

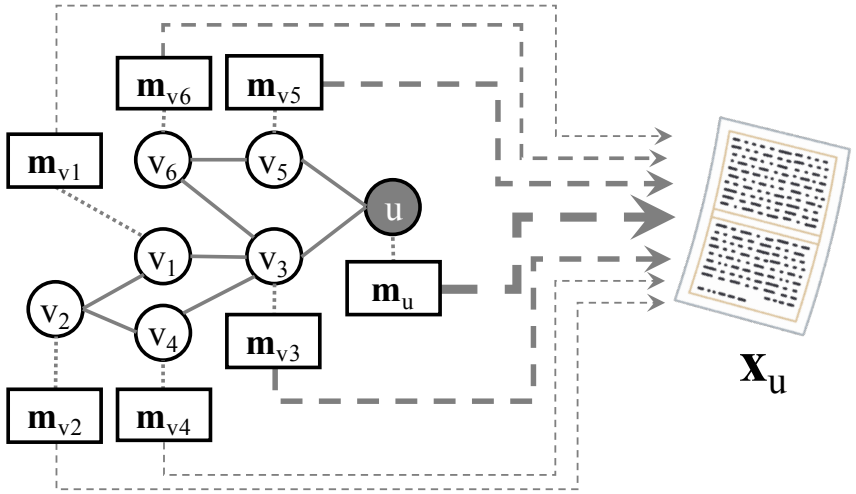
論文中の表1より

Model	Total		
	@10	@100	@1000
BM25	.1823	.5175	.7703
PRMS	.2522	.5919	.8009
MLM-all	.2571	.6136	.8009
LM	.2607	.6413	.8009
SDM	.2659	.6674	.8633
LM-ELR	.2646	.6483	.8006
SDM-ELR	.2739	.6782	.8633
MLM-CA	.2639	.6370	.8329
BM25-CA	.2782	.6727	**.8708**
FSDM	.2812	.6667	.8455
BM25F-CA	.2811	**.6912**	.8653
FSDM-ELR	**.2872**	.6765	.8450
max	.2872	.6912	.8708
gap	.5836	.1796	—

- 近傍のリテラルだけでは不十分 ➔ RWRDoc
- @10，@100にまだまだ改善の余地 ➔ PPRSD

RWRDoc
(RWR-basd Document Learning)

- アイデア
  - エンティティから到達可能なリテラルを利用
  - より「近い」リテラルを信頼
- RWRDoc
  - 初期：各エンティティを
    隣接リテラルの
    TF-IDF で表現
  - 到達可能エンティティを
    「近さ」に応じて重み付け
    - 近さ：RWR
      (Random Walk with Restart)



PPRSD
(Personalized PageRank-based Score Distribution)

- アイデア
  - 既存手法でも Top-1000 は 80+% の再現率
  - より関連のあるものを上位に再ランキング
  - LODのグラフ構造を活用
- PPRSD
  - PPRで計算済みスコアを再分配
    - 単純にPPRのスコアを組み合わせると性能が低下

$$\mathbf{pprsd}_q = (1-d) \cdot \mathbf{pprsd}_q A + d \cdot \mathbf{t}$$

Top-1000 エンティティの
誘導部分グラフの隣接行列

Top-1000 エンティティの
スコアベクトル

実験

- 質問
  1. グラフ上の距離を用いたエンティティの
     文書表現は再現率の向上に有用か?
  2. グラフ分析はエンティティ検索の
     ランキング性能 (NDCG) 向上に有効か?
  3. これらを組み合わせた手法は
     エンティティ検索の性能 (NDCG) 向上に有効か?
- データセット：DBpedia-Entity v2
  - 対象LOD：DBpedia 2015-10
  - クエリセット：SemSearch ES, INEX-LD,
    ListSearch, QALD-2

結果（再現率）

赤：改善
青：ほぼ同じ

論文中の表2より

Model	SemSearch ES			INEX-LD			ListSearch			QALD-2			Total		
	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
BM25	.2563	.6669	.9280	.1730	.4860	.7554	.1093	.4598	.7221	.1891	.4677	.6929	.1823	.5175	.7703
PRMS	.3719	.7499	.9412	.2312	.5339	.7796	.1839	.5476	.7525	.2273	.5428	.7420	.2522	.5919	.8009
MLM-all	.3887	.7705	.9412	.2343	.5527	.7796	.1840	.5655	.7525	.2280	.5706	.7420	.2571	.6136	.8009
LM	.3812	.8236	.9412	.2425	.5807	.7796	.1899	.5772	.7525	.2355	.5910	.7420	.2607	.6413	.8009
SDM	.3884	.8581	.9865	.2409	.6224	.8567	.1987	.6121	.8256	.2398	.5921	.7991	.2659	.6674	.8633
LM-ELR	.3863	.8278	.9412	.2364	.5894	.7796	.1913	.5940	.7536	.2474	.5909	.7401	.2646	.6483	.8006
SDM-ELR	.3898	.8581	**.9865**	.2366	.6307	.8567	.2105	.6180	.8256	.2589	.6172	.7991	.2739	.6782	.8633
MLM-CA	.4096	.7843	.9420	.2249	.5917	.8051	.1861	.5834	.8038	.2377	.5953	.7894	.2639	.6370	.8329
BM25-CA	.3991	.8326	.9766	.2372	.6266	.8603	.2110	.6261	.8431	.2650	.6157	.8164	.2782	.6727	.8708
FSDM	.4459	.8515	.9581	.2390	.6153	.8191	.1980	.5999	.8175	.2466	.6102	.7970	.2812	.6667	.8455
BM25F-CA	.4097	**.8707**	.9704	.2607	.6526	.8544	.2042	.6189	.8325	.2548	.6341	.8157	.2811	.6912	.8653
FSDM-ELR	**.4536**	.8539	.9562	.2477	.6253	.8191	.2022	.6075	.8162	.2507	.6275	.7970	.2872	.6765	.8450
RWRDoc	.4001	.8303	.9801	.2408	.6391	**.8624**	.2177	.5902	**.8613**	.2390	.6433	**.8298**	.2744	.6757	**.8834**
Imp. (%)	-11.79	-4.64	-0.65	-7.71	-2.07	+0.24	+3.18	+5.73	+2.16	-9.81	+1.45	+1.64	-2.51	-2.24	+1.45
RWRDoc*	.4325	.8511	.9801	**.2618**	**.6671**	**.8624**	**.2307**	**.6582**	**.8613**	**.2655**	**.6716**	**.8298**	**.2976**	**.7120**	**.8834**
Imp. (%)	-4.65	-2.25	-0.65	+0.42	+2.22	+0.24	+9.34	+5.13	+2.16	+0.19	+5.91	+1.64	+3.62	+3.01	+1.45

- RWRDocで主に @1000 の再現率が改善
- +PPRSDで @10, @100 の再現率が改善

結果（NDCG）

- PPRSD（図中の*）により，ほぼすべての手法の結果が改善
- 最良はすべてPPRSDで改善したもの
- RWRDoc + PPRSDが全体で最良

論文中の表3より

Model	SemSearch ES		INEX-LD		ListSearch		QALD-2		Total	
	@10	@100	@10	@100	@10	@100	@10	@100	@10	@100
BM25	.2497	.4110	.1828	.3612	.0627	.3302	.2751	.3366	.2558	.3582
BM25*	.2839	.4463	.2903	.3816	.2534	.3543	.2953	.3624	.2812	.3847
Rise (%)	+13.7	+8.59	+58.8	+5.65	+304	+7.30	+7.34	+7.66	+9.93	+7.40
PRMS	.5340	.6108	.3590	.4295	.3684	.4436	.3151	.4026	.3905	.4688
PRMS*	.5388	.6162	.3590	.4295	.3684	.4436	.3151	.4026	.3913	.4698
Rise (%)	+0.90	+0.88	0.00	0.00	0.00	0.00	0.00	0.00	+0.20	+0.21
MLM-all	.5528	.6247	.3752	.4493	.3712	.4577	.3249	.4208	.4021	.4852
MLM-all*	.5578	.6303	.3752	.4493	.3712	.4577	.3249	.4208	.4030	.4863
Rise (%)	+0.90	+0.90	0.00	0.00	0.00	0.00	0.00	0.00	+0.22	+0.23
LM	.5555	.6475	.3999	.4745	.3925	.4723	.3412	.4338	.4182	.5036
LM*	.5606	.6529	.3999	.4745	.3925	.4723	.3413	.4338	.4191	.5046
Rise (%)	+0.92	+0.83	0.00	0.00	0.00	0.00	+0.03	0.00	+0.22	+0.20
SDM	.5535	.6672	.4030	.4911	.3961	.4900	.3390	.4274	.4185	.5143
SDM*	.5564	.6718	.4030	.4912	.3961	.4902	.3394	.4274	.4191	.5152
Rise (%)	+0.52	+0.69	0.00	+0.02	0.00	+0.04	+0.12	0.00	+0.14	+0.17
LM-ELR	.5554	.6469	.4040	.4816	.3992	.4845	.3491	.4383	.4230	.5093
LM-ELR*	.5608	.6518	.4040	.4816	.3992	.4847	.3491	.4383	.4240	.5103
Rise (%)	+0.97	+0.76	0.00	0.00	0.00	+0.04	0.00	0.00	+0.24	+0.20
SDM-ELR	.5548	.6680	.4104	.4988	.4123	.4992	.3446	.4363	.4261	.5211
SDM-ELR*	.5577	.6716	.4105	.4988	.4129	.4999	.3449	.4364	.4271	.5218
Rise (%)	+0.52	+0.54	+0.02	0.00	+0.15	+0.14	+0.09	+0.02	+0.23	+0.13
MLM-CA	.6247	.6854	.4029	.4796	.4021	.4786	.3365	.4301	.4365	.5143
MLM-CA*	.6249	.6895	.4029	.4798	.4020	.4786	.3365	.4301	.4361	.5150
Rise (%)	+0.03	+0.60	0.00	+0.04	-0.02	0.00	0.00	0.00	-0.09	+0.14
BM25-CA	.5858	.6883	.4120	.5050	.4220	.5142	.3566	.4426	.4399	.5329
BM25-CA*	.6040	.7024	.4132	.5048	.4302	.5181	.3607	.4544	.4475	.5404
Rise (%)	+3.11	+2.05	+0.29	-0.04	+1.94	+0.76	+1.15	+2.67	+1.73	+1.41
FSDM	.6521	.7220	.4214	.5043	.4196	.4952	.3401	.4358	.4524	.5342
FSDM*	.6549	.7269	.4214	.5044	.4196	.4951	.3401	.4359	.4527	.5350
Rise (%)	+0.43	+0.68	0.00	+0.02	0.00	-0.02	0.00	+0.02	+0.07	+0.15
BM25F-CA	.6281	.7200	.4394	.5296	.4252	.5106	.3689	.4614	.4605	.5505
BM25F-CA*	.6444	**.7361**	**.4494**	.5336	.4288	.5166	**.3699**	.4672	.4673	.5581
Rise (%)	+2.60	+2.24	+2.28	+0.76	+0.85	+1.18	+0.27	+1.26	+1.48	+1.38
FSDM-ELR	.6563	.7257	.4354	.5134	.4220	.4985	.3468	.4456	.4590	.5408
FSDM-ELR*	**.6572**	.7307	.4354	.5135	.4219	.4985	.3466	.4455	.4587	.5416
Rise (%)	+0.14	+0.69	0.00	+0.02	-0.02	0.00	-0.06	-0.02	-0.07	+0.15
RWRDoc	.5877	.7215	.4189	.5296	.4119	.5845	.3346	.5163	.4348	.5643
RWRDoc*	.6379	.7288	.4413	**.5462**	**.4355**	**.6015**	.3591	**.5623**	**.4684**	**.6097**
Rise (%)	+8.54	+1.01	+5.35	+3.13	+5.73	+2.91	+7.32	+8.91	+7.73	+8.05
Imp. (%)	-2.94	-0.99	-1.80	+2.36	+1.23	+16.1	-2.92	+20.4	+0.24	+9.25

青：悪化
赤：最良

実験まとめ

1. グラフ上の距離を用いたエンティティの文書表現は再現率の向上に有用か? – Yes, but limited
   - @1000 の向上に効果的
   - @10, @100には有効性は見いだせない
2. グラフ分析はエンティティ検索のランキング性能 (NDCG) 向上に有効か? – Yes
   - ほぼすべての手法について向上
   - 簡単すぎる検索 (SemSearch ES) への効果は薄い
3. これらを組み合わせた手法はエンティティ検索の性能 (NDCG) 向上に有効か? - Yes
   - RWRDoc の弱点 (@10, @100) をPPRSDが補った

まとめと今後の課題

- まとめ
  - 対象：LODにおけるエンティティ検索
  - 提案手法
    - RWRDoc: 距離に基づくエンティティの表現学習
    - PPRSD: グラフ分析を利用した再ランキング
  - 結果：9.25%のエンティティ検索の性能向上
- 今後の課題
  - 述語の考慮
  - 潜在的表現学習の導入