3F3-1

大規模エンティティ関係グラフにおけるランキング

Ranking on Large-Scale Entity-Relationship Graphs

森純一郎
Junichiro Mori

東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

In this research, we aim to develop a method for ranking on large-scale entity-relationship graphs. Most of existing methods rely on solely IR-based approaches to rank entities. However, link-based approach may be useful for ranking entities. In particular, several link-based features together with document-based features can be considered on entity-relation graphs. To find the best combination of the features in ranking on entity-relation graphs, we examine several weighting methods for ranking entities on graphs. Our method contributes to entity search by providing the general ranking framework on entity-relation graphs.

1. はじめに

ウェブはテキストとテキストのハイパーリンクにより構成される膨大なネットワークであるが、近年のマイクロブログやソーシャルネットワークといったソーシャルメディアの爆発的な普及により、従来のテキストのネットワークの上に人と人あるいは人とモノ・コトのつながりが紡ぎだす膨大なネットワークがウェブに表出してきている。これらのネットワークは人や組織や場所などのエンティティとそれらの関係による構成される膨大なグラフデータを生成している。

エンティティとそれらの関係性がデータとして扱えるようになり、従来のウェブのリンク関係を用いた情報検索に対して、エンティティの関係性に着目した情報検索が近年提案されている。このようなエンティティを介した情報検索においては、データとしてのエンティティ関係グラフから任意の検索要求に対して適切なエンティティあるいはその関係を抽出しランキングづけすることが必要となる。特に、twitter, facebook などのソーシャルメディア、dbpedia や freebase などの linkeddata のような大規模なエンティティ関係グラフの情報源からエンティティを検索する際は、どのように膨大なエンティティからランキングにより所望のエンティティを検索するかは大きな課題である。

エンティティ関係グラフは、ノードとしてのエンティティとノードとノードを結ぶリンクとしての関係から構成される。従来のウェブグラフにおいては、HITS や PageRank のようなリンク構造に着目した手法がウェブページのランキング有効であることが示されている。エンティティ関係グラフにおけるエンティティランキングにおいても同様の手法を適用することが考えられるが、ウェブグラフとエンティティ関係グラフでは異なる点がある。多くは非構造化データであるウェブページ同士がハイパーリンクという単一の関係で結ばれたウェブグラフと比較すると、エンティティ関係グラフではノードやリンクがさまざまなメタデータが保持していること。例えば、ソーシャルネットワークであればノードである人は構造化された属性情報と他の人との複数の関係性情報を保持している。そのため、これらの特徴を加味したランキング手法を設計する必要がある。

連絡先: 森純一郎，東京大学大学院工学系研究科，東京都文京区弥生 2-11-16 工学部 9 号館 119 号室，03-5841-1161，jmori@ipr-ctr.t.u-tokyo.ac.jp

本研究では、ウェブにおけるエンティティ情報の検索を目的とし、ウェブドキュメント検索ランキングに現在用いられている PageRank の大規模エンティティ関係グラフにおけるランキングへの適用について提案する。提案手法においては、エンティティ関係グラフの特徴に基づき、PageRank によるランキングに寄与する複数の重み付けの比較検討を行う。実験においては特に人物エンティティのソーシャルネットワークを対象に、大規模なエンティティ関係グラフに提案手法を適用し考察を行う。

2. 大規模エンティティ関係グラフにおけるランキング

2.1 PageRank のエンティティ関係グラフへの適用

PageRank [Brin 07] はウェブのリンク構造に基づくウェブページの重要度計算手法であり、現在ウェブ検索におけるウェブページランキング指標の計算手法として広く用いられている。PageRank によるウェブページの重要度は以下の式で表される:

$$\pi_t = \alpha P^T \pi_{t-1} + (1-\alpha) r.$$

π_t は各ウェブページの重要度に対応するスコアベクトルである。P はページ間リンクに基づくページ間の遷移確率を表す遷移行列である。r は任意のページへのジャンプを表すテレポートベクトルである。α はリンク遷移とテレポートをバランスさせるダンプリング係数である。PageRank の計算はマルコフ連載であり、その値は収束することが示されている。実用においては繰り返しにより $\pi_t - \pi_{t-1}$ が十分小さい値となれば計算を停止する。

PageRank において、ページをエンティティ、リンクを関係と考えると上記の式は、エンティティ関係グラフにおけるエンティティの重要度の計算に適用できる。しかしながら、ウェブグラフに対してエンティティ関係グラフではエンティティや関係がメタデータとなる様々な情報を包含しており、単一のリンク構造にのみに基いて計算される PageRank においてエンティティや関係の情報をどのように反映するかを考慮する必要がある。ここでは特に、PageRank の計算において重要な以下の重み付けについて検討する。

スコアベクトルの初期重み

ウェブグラフにおいては初期のスコアベクトル π_1 は一般にページ数で単一の値に標準化される。エンティティ関係グラフにおいても同様の初期化が考えられるが、初期値として各エンティティに固有の重要度を考慮することが考えられる。任意のエンティティe_i の重要度を $\phi(e_i)$ とすると、初期のスコアベクトルは以下のように表される:

$$\pi_1 = \frac{A}{\Sigma_i \phi(e_i)}.$$

ここで、A は $A = (\phi(e_1), \phi(e_2), ..., \phi(e_n))$ なるベクトルである。

遷移行列の重み

遷移行列の遷移確率は一般にすべてのリンクの重みを均一に扱い計算されるが、リンクの重みを考慮することが考えられる。エンティティ関係グラフにおいては、リンクの重みは関係の強さに対応する。任意のエンティティe_i,e_j の関係の強さを $R(i,j)$ とすると e_i から e_j へ遷移確率は以下のように計算される:

$$P_{ij} = \frac{R(i,j)}{\Sigma_{k=1}^{n} R(i,k)}.$$

テレポートベクトルの重み

テレポートベクトル r は一般に任意のページへテレポートするランダムサーファーモデルが用いられるが、検索クエリーにより関連したページへのテレポートを考慮した PageRank も提案されている [Chakrabati 07]。エンティティ関係グラフにおいても検索クエリーとエンティティの関連度を考慮したテレポートが考えられる。検索クエリー q とエンティティe の関連度は条件確率 $p(e|q) \simeq p(q|e)p(e)$ によって計算される。ここで $p(q|e)$ はエンティティの情報から言語モデルにより計算される。これによりテレポートベクトル r は以下のように計算される。

$$r = \frac{T}{\Sigma_i p(e_i|q)}.$$

ここで、T は $T = (p(e_1|q), p(e_2|q), ..., p(e_n)|q)$ なるベクトルである。

以下の実験では、エンティティ関係グラフにおける PageRank のこれらの重み付けについて実際のデータを用いて評価を行う。

2.2 実験:大規模ソーシャルグラフにおける人物ランキング

本実験においては大規模なエンティティ関係グラフとして、人物のソーシャルネットワークを対象とする。人物のソーシャルネットワークは、人検索サイト *SPYSEE* から取得した。*SPYSEE* では、人物情報と関係情報をウェブから自動抽出している。本実験においては 100 万の人物エンティティとそれらの関係をからエンティティ関係グラフを構築した。関係は人物間の共起によって重み付けされている。関係は複数の種類が考えられるが、ここではすべての関係を一つの重み付けで表すことにする。

エンティティ関係グラフに対する PageRank の計算において以下の重み付けの比較を行う。

- スコアベクトルの初期重み (F1):人物のアクセス頻度重みあり (Y)/なし (N)
- 遷移行列の重み (F2):人物間の関係重みあり (Y)/なし (N)
- テレポートベクトルの重み (F3):クエリーと人物の関連度重みあり (Y)/なし (N)

表 1: 重み付け条件に対する NDCG

F1	F2	F3	NDCG@5	NDCG@10
N	N	N	.3604	.3840
Y	N	N	.3604	.3840
N	Y	N	.0050	.0083
N	N	Y	.3818	.3928
Y	Y	N	.0050	.0083
N	Y	Y	.1566	.1813
Y	N	Y	.3818	.3928
Y	Y	Y	.1566	.1813

正解として 50 の検索クエリーとそれぞれの検索クエリーに対するランキングの正解データを人手により作成した。正解データはランキングに対して複数人が合意したものを使用した。それぞれの重み付け条件を変更し、検索クエリーに対するランキングを計算し、正解データに基づき NDCG (Normalized Discounted Cumulative Gain) により評価した。任意のクエリーに対するランキング上位 k の DCG は以下のように計算される:

$$DCG@k = score_1 + \Sigma_{i=2}^{k} \frac{score_i}{log_2 i}.$$

ここで、$score_k$ はクエリーに対する正解ランク k_{true} の逆数 $1/k_{true}$ とした。NDCG は DCG を正解ランキングの DCG によって正規化したものである。

なお、PageRank の計算においてはダンプリング係数は 0.85 とした。

2.3 考察

表 1 は、各重み付け条件に対する NDCG の値を示している。ランキングに対するスコアベクトルの初期重みの影響はほとんどない。一方、遷移確率において関係の重み考慮した場合は、NDCG が著しく減少する。これは、関係重みの分布の偏りにより、特的のエンティティに遷移が集中してしまうことに起因すると考えられる。また、複数の関係を一つの重みに集約していることも原因の一つと考えられる。エンティティ関係の重み付け関数を適切に設計することが必要である。テレポートベクトルの重みは、ランキングに貢献している。より適切なランキングのためにはクエリーとエンティティの関連度の推定を適切に行うことが考えられる。

3. おわりに

本研究では、ウェブにおけるエンティティ情報の検索を目的とし、PageRank の大規模エンティティ関係グラフにおけるランキングへの適用における重み付けについて、人物エンティティの大規模ソーシャルネットワークを対象とした実験を通して考察をおこなった。今後は学習に基づくランキング精度の向上を行う。

参考文献

[Chakrabati 07] Chakrabarti, S.: Dynamic personalized pagerank in entity-relation graphs, Proceedings of the 15th international conference on World Wide Web, 2007.

[Brin 07] Brin, S. and Page, L.; The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks and Isdn Systems, Vol.30, 1998.