ブログ空間の情報伝播特性を用いた情報源の多面的ランキング

清水　豪士

富山県立大学　情報基盤工学講座
t715038@st.pu-toyama.ac.jp

April 30, 2021

はじめに

背景

- ブログの登場により Web 上の情報公開コストが大幅に低下し，評判情報を含めて多種多様な情報が得られるようになった.
- しかし，情報量が膨大で，重複している情報や無益・有害な情報が多く含まれており，ユーザが有効に活用できているとはいいがたい.

目的

- ユーザの情報探索の目的に合わせた有用な情報源の推薦を実現するために，情報伝播ネットワーク構造の情報源の影響範囲である部分ネットワークの 3 種類の情報伝播特性の使い分けによる多面的なランキング手法を提案.

ブログの情報源の多面的ランキング

ユーザの情報探索指向に応じた多面的ランキング

- 多くのブロガが情報源から入手した情報をブログエントリを書いてより多くの人たちに連鎖的に情報を伝えた結果として情報伝播ネットワークが生成される.
- 情報拡散，情報集約および情報転送の基本操作に関する情報伝播特性を定量化すれば，複数の観点から情報源の重要度を知ることができ，さらに関連している３種類の情報伝播特性を使い分ければ，注目されている情報，資料性の高い情報，口コミで伝わりやすい情報など，ユーザの情報探索の目的に合わせた情報源の推薦を実現できる.

情報伝播ネットワークの抽出

ブログエントリの検索と収集

- ブログ空間には多種多様な情報が伝播しているが，特定の情報の流れを詳しく解析するためには，特定のトピックに関するブログエントリに限定する必要がある．
- そこで，Yahoo!または Technorati Japan のブログ検索を用いて，そのトピックを表す検索語が含まれるブログエントリを取得する．

本文の特定とハイパーリンクの抽出

- ブログのエントリには内容とは直接の関係がない多量のハイパーリンクが含まれるために，本文部分を特定して，その中に含まれるハイパーリンクだけを抽出する．

情報伝播ネットワークの抽出

ノードとエッジの生成時刻の特定

- 抽出したネットワーク構造の分析に加えて，時間的変化とそれにともなう特性の変化も分析できるように，ノードとエッジの生成時刻を特定する．
- ノードの場合
  収集済みのブログの場合はその生成時刻を使用し，それでなければそのノードを一番最初にリンクした時刻を生成時刻とする．
- エッジの場合
  リンク下のノード生成時刻をエッジの生成時刻とする．

情報伝播ネットワークの抽出

情報伝播ネットワークの作成

- 抽出した全ての URL に対して異なるサーバからの被リンク数を計算し，指定された閾値 T 以上の URL だけを特に注目されている第１次情報源とみなし，そこからハイパーリンクを逆向きにたどり伝播経路を特定した.
- 実際には，生成時刻の古いブログエントリから順番に，リンク先 URL がすでにノードとして登録されていた場合，その URL からブログエントリへの有向エッジを作成し，未登録のリンク先 URL が閾値以上の被リンク数の持つ場合，ノードとして登録し，リンク先からブログエントリへの有向エッジを作成する.
- この結果，情報源とブログエントリをノードとし，その間のハイパーリンクとは逆方向の有向エッジを持つ情報伝播ネットワークが得られる.

情報伝播ネットワークの抽出

因果関係に矛盾するエッジ・ノードの除去

- 情報伝播には方向性があり，古いブログエントリから新しいブログエントリに情報が伝播する．
- すなわち古いブログエントリを新しいブログエントリがリンクするという因果関係がある．
- この因果関係に従わない双方向リンクや古いブログエントリから新しいブログエントリへの逆方向リンクが存在する場合には，エッジとノードの両方を除去する．

情報伝播特性の定量化とランキング

情報源の影響範囲の特定と部分ネットワークの分割

- 情報伝播ネットワークは均一ではなく，各情報源が周囲に与える影響の違いに応じて各部が異なる構造を持つ．
- 各情報源ノードを始点とし，有向エッジをたどって到達できる範囲までを，その情報源の影響範囲の部分ネットワークとして分割する．

情報伝播ネットワークを $G = (V, E)$(V はノード集合，E はエッジ集合)，V に含まれる n 個のノードを v_i，E に含まれるノード v_i からノード v_j への有向エッジを e_{ij} とする．情報源であるノード v_k からたどることができる部分ネットワーク $G_k = (V_k, E_k)$ のノード集合 V_k とエッジ集合 E_k を次のように定義する．

$$V_k = \{v_i | v_i = v_k \vee directed\_path(v_k, v_i)\},$$
$$E_k = \{e_{ij} | v_i \in V_k \vee v_j \in V_k\}.$$

ここで $directed\_path(v_k, v_i)$ は，ノード v_k からノード v_i の間に有向パスが存在することを示す．

情報伝播特性の定量化とランキング

情報伝播ネットワークの基本構造

- 情報伝播ネットワークおよびその部分ネットワークを構成する基本単位として，２エッジグラフに着目する．

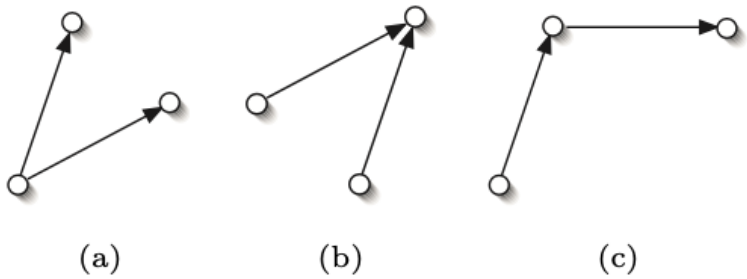

図 1: ３種類の２エッジ連結部分グラフ

- (a),(b),(c) は情報伝播ネットワークにおける基本構造を示しており，それぞれ情報拡散構造，情報集約構造，情報転送構造という．

情報伝播特性の定量化とランキング

情報源であるノード v_k からたどれる部分ネットワーク $G_k = (V_k, E_k)$ に含まれる各基本構造の数である情報拡散構造数 $N_s(G_k)$, 情報集約構造数 $N_g(G_k)$, 情報転送構造 $N_t(G_k)$ は，２エッジ部分グラフの接続点であるノード v_i の入次数 $d_{in}(v_i)$ と出次数 $d_{out}(v_i)$ から，次のように求められる．

$$N_s(G_k) = \sum^{v_i \in V_k} \frac{d_{out}(v_i) \times (d_{out}(v_i) - 1)}{2},$$

$$N_g(G_k) = \sum^{v_i \in V_k} \frac{d_{in}(v_i) \times (d_{in}(v_i) - 1)}{2},$$

$$N_t(G_k) = \sum^{v_i \in V_k} d_{in}(v_i) \times d_{out}(v_i).$$

情報伝播特性の定量化

- 情報拡散構造数は v_i を始点とする２本の出エッジの組み合わせ数
- 情報集約構造数は v_i を終点とする２本の入エッジの組み合わせ数
- 情報転送構造数は v_i を中間点とする入エッジと出エッジの組み合わせ数

情報伝播特性の定量化とランキング

異なる部分ネットワークを互いに比較できるように，部分ネットワーク G_k のノード数 $|V_k|$ で正規化し，情報拡散度 $P_s(G_k)$, 情報集約度 $P_g(G_k)$, 情報転送度 $P_t(G_k)$ を定義する．

$$P_s(G_k) = \frac{N_s(G_k)}{|V_k|},$$
$$P_g(G_k) = \frac{N_g(G_k)}{|V_k|},$$
$$P_t(G_k) = \frac{N_t(G_k)}{|V_k|}.$$

多面的ランキング

- 情報伝播ネットワークの各情報源の特性の定量化指標である情報拡散度，情報集約度，情報転送度をランキングに使用する．
- ユーザが情報探索の目的に合わせてランキングに用いる定量化指標を切り替えることで，情報源の多面的ランキングを実現する．

評価

データセット

- 評価には図２に示している５個のデータセットを用いる．

No.	検索語	期間	エントリ	リンク	ノード	エッジ	情報源
1	iPhone	08/7/10〜17	10,801	25,836	646	772	45
2	毎日新聞	08/8/11〜25	3,863	13,857	350	508	20
3	Google Chrome	08/9/1〜9	1,926	6,764	649	772	27
4	Doblog	08/9/28〜09/5/1	1,019	4,288	211	250	13
5	地デジカ	09/4/28〜5/12	1,179	6,025	418	583	22

図 2: 評価に用いるデータセット

- 情報源の異なるサーバからの被リンク数の閾値は，各データセット共通で $T = 10$ とする．
- 図２の各欄は収集に使用した検索語，ブログエントリの生成期間，ブログエントリ数，総ハイパーリンク数，情報伝播ネットワークのノード数，エッジ数，情報源数を示す．

評価

ランキングの相関の分析

- 各ランキング結果の類似度合いをスピアマンの順位相関係数 ρ を用いて調べる

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i}{N(N^2-1)}$$

- N は要素数，d_i は順位差
- ３種類の情報伝播特性を比較するために，既存のリンク解析によるランキングによく用いられる出自数も用意した.

評価

ランキングの相関の分析

- 各ランキング手法の組のすべてのデータセットに対する順位相関係数の平均値を図 3 に示す.
- 括弧内の値は分散

	出次数		情報拡散度		情報集約度	
情報拡散度	0.966	(0.00129)	—		—	
情報集約度	−0.00672	(0.196)	−0.0669	(0.166)	—	
情報転送度	0.619	(0.0581)	0.503	(0.0641)	0.281	(0.104)

図 3: 順位相関係数の平均値と分散

- 出次数と情報拡散度の間には強い相関がある.
- 出次数と情報転送度，情報拡散度と情報転送度の間位には中程度の相関がある.
- 情報拡散度，情報集約度，情報転送度のランキング結果が異なっていることを示している.

評価

情報源の抽出精度と分類結果

- データセットからの情報源の抽出精度と抽出された情報源の種類によるランキング結果の違いを調べるために，オフィシャルサイト，ニュース，CGM，その他の適合情報および不適合情報源の５種類に分類.

	1	2	3	4	5
オフィシャルサイト	11	3	11	1	4
ニュース	22	5	14	7	10
CGM	4	8	2	3	5
その他（適合）	0	1	0	0	0
その他（不適合）	8	3	0	2	3
合計	45	20	27	13	22
抽出精度	0.82	0.85	1.0	0.85	0.86

図 4: 情報源の抽出精度と分類結果

- 情報源抽出精度の平均は 0.88 である.
- 分類内容を調べると，ブログ空間から注目されている情報源はオフィシャルサイトとニュースの割合が多い.

評価

ランキング手法の特性分析

- 各ランキング手法の特性を分析するために，Web 検索や QA システムの結果の質の評価に用いられる MRR と MAP を用いる．
- 検索後に対するランキング結果に対しての３種類の情報源をそれぞれ正解とみなした場合の３つのデータ評価ようデータを作成し，それぞれの性能を比較することでランキング手法の違いを分析する．

MRR(Mean Reciprocal Rank)

- 各課題の適合文書が最初に見つかった順位の逆数を全課題に対して平均した値

MAP(Mean Average Precision)

- 課題ごとに各適合文書が見つかった順位における精度の平均を求め，さらに全課題に対して平均した値

評価

ランキング手法の特性分析

- 図２に示した５つのデータセットに対して，情報源の種類がオフィシャルサイト，ニュース，CGM である場合に正解と見なした場合の出次数，情報拡散度，情報集約度，情報転送度の４種類の MRR と MAP の結果を図５と図６に示す.

	出次数	情報拡散度	情報集約度	情報転送度
オフィシャルサイト	1	1	0.384	0.5
ニュース	0.9	0.9	0.653	1
CGM	0.356	0.356	0.814	0.86

図 5: MRR の結果

	出次数	情報拡散度	情報集約度	情報転送度
オフィシャルサイト	0.654	0.654	0.0894	0.314
ニュース	0.473	0.473	0.234	0.669
CGM	0.144	0.144	0.350	0.417

図 6: MAP の結果

評価

ランキング手法の特性分析

- MRR は，出次数と情報拡散度はほぼ同じ傾向を示すが，情報集約度，情報転送度ともにかなり異なる傾向を示す．
- 上位になりやすい情報源の種類は，出次数と情報拡散度ではオフィシャルサイト，情報集約度では GCM ，情報転送度ではニュースと，顕著に分かれている．
- MAP も，MRR と似た傾向を示す．
- 出自数と情報拡散度ではオフィシャルサイト，情報集約度では GCM，情報転送度ではニュースが最上位に限らず全体的に上位にランキングされる傾向がある．

評価

ランキングの順位差の分析

- どの種類の情報の場合に情報伝播先でも情報拡散構造が生まれやすいかを調べるために，各データセットでオフィシャルサイト，ニュース，CGM のそれぞれで，順位差がある文書で，どの程度の順位差があるかを調べる．
- 情報源の種類ごとに，出次数と情報拡散度の順位の差が生じた場合の順位差の平均をとり，これを ARD と呼び．

	1	2	3	4	5	平均
オフィシャルサイト	−6.6	−3	0.3	0	0	−1.85
ニュース	1.46	0	−0.142	−0.2	0	−0.224
CGM	2	0.333	0	−0.5	0	0.367

図 7: ARD の結果

- ランキング上位の結果が同じでも，ランキング下位においては情報拡散度は出次数よりもオフィシャルサイトの順位が低くなる傾向があり，CGM の順位が高くなる傾向があることがわかる．

6. おわりに

まとめ

- ある検索語でブログを検索した結果から，ブロガが注目している情報源を起点とする情報伝播ネットワークを抽出し，各情報源からたどれる部分ネットワークの３種類の情報伝播特性を使い分けることで，ブログの情報源を多面的にランキングする手法を提案した.
- ランキングに使用する情報伝播特性によって，オフィシャルサイト，ニュース，CGM のような種類が異なる情報源が高く評価されることを示した.

今後の課題

- 抽出次の閾値 T を下げたときにスパムブログが多く抽出され，特に情報集約度に無視できない影響を与えるため，機械生成されるスパムブログが持つリンク構造の特徴を用いたスパムブログ排除機能の実装が必要