

8. グラフデータを分析する

PageRank, TwitterRank, 影響最大化, ...

梅本 和俊

情報通信研究機構／東京大学
umemoto@tkl.iis.u-tokyo.ac.jp

お知らせ

今さらながら…

講義資料をまとめたサイトを作りました

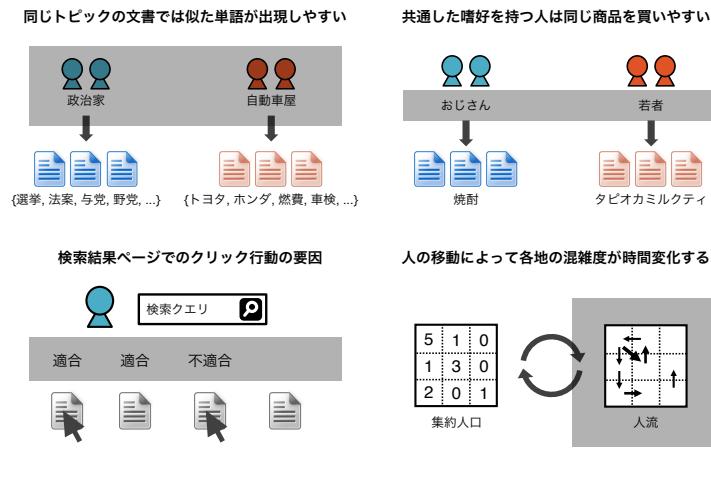
<https://umemotsu.github.io/introduction-to-data-science-2019/>

- Speaker Deckにアップロードしていたスライドを移動
- Q&A用のページを作成
- アンケート・演習リポジトリへのリンクを掲載

おさらい

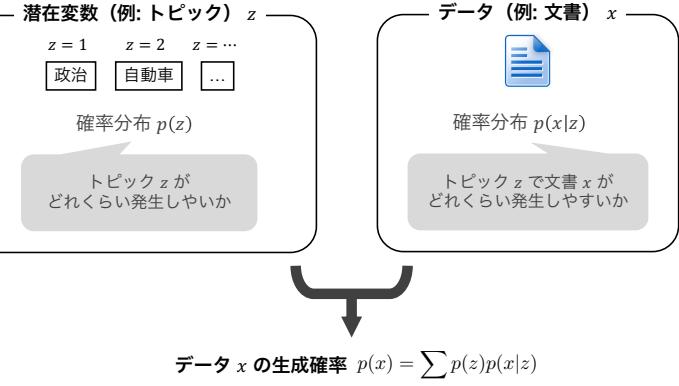
前回 | 潜在変数モデル

閑話休題 | 観測データの背後に潜む隠れた情報¹⁷



潜在変数モデル | ステップ1 (モデル)¹⁸

潜在変数やデータの生成過程の記述に適切な確率分布を決定



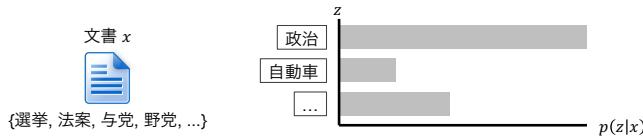
潜在変数モデル | ステップ3 (応用)²⁰

ステップ2 (後述) により $\{p(z)\}_z$ と $\{p(x|z)\}_{z,x}$ が推定できた

ペイズの定理

$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{p(z)p(x|z)}{\sum_{z'} p(z')p(x|z')} = \frac{p(z)p(x|z)}{\sum_{z'} p(z')p(x|z')}$$

(既知 or 未知) データ x の潜在変数の分布は...



応用例

- 同じ話題に関する文書集合をまとめる
- キーワードと（表層的には異なるが）意味的にマッチする文書を検索する
- 似た嗜好を持つユーザが好む商品を推薦する

潜在変数モデル¹⁸

隠れた情報 (潜在変数 z) からデータ x が生成される過程を
条件付き確率 $p(x|z)$ でモデル化

基本的な流れ

1. モデル

潜在変数やデータの生成過程の記述に適切な確率分布を決定

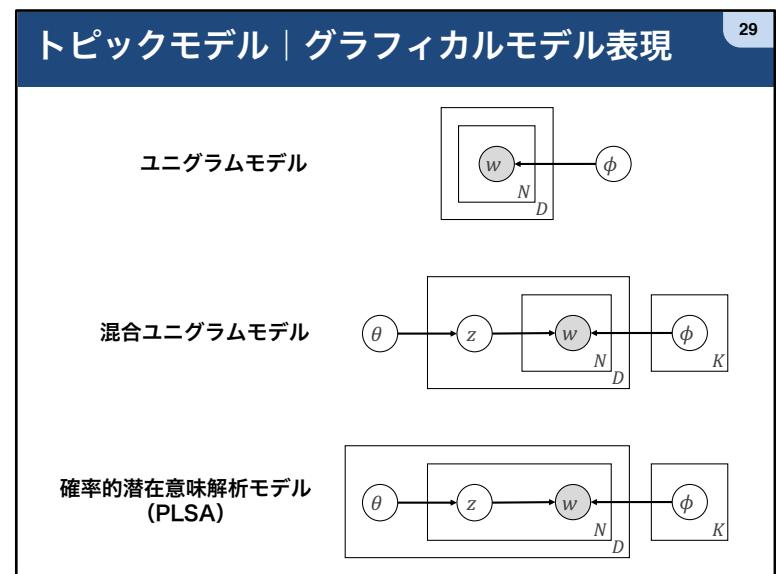
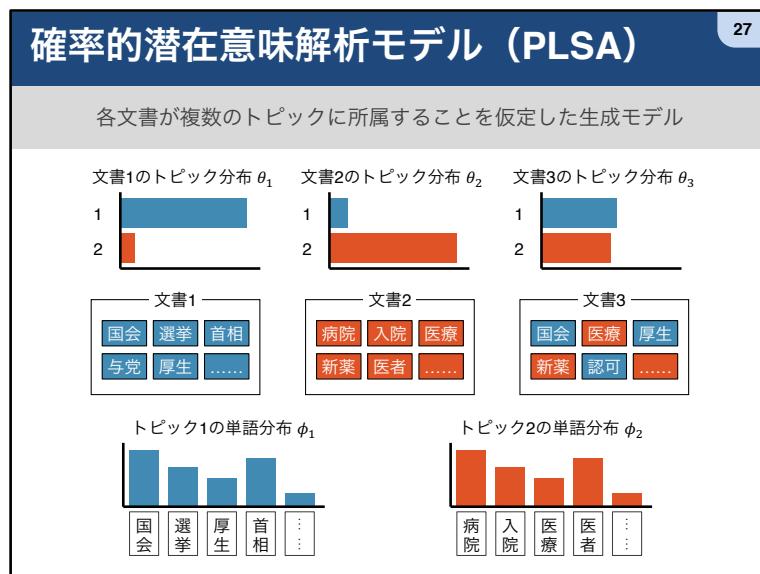
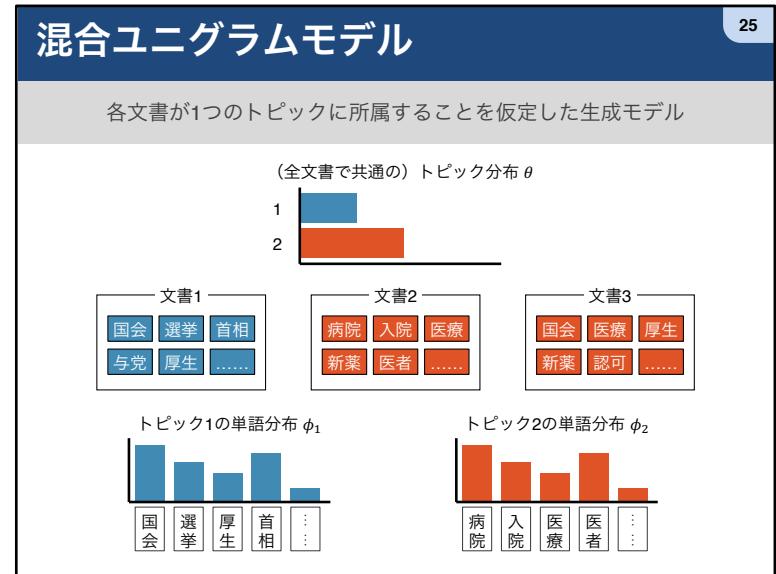
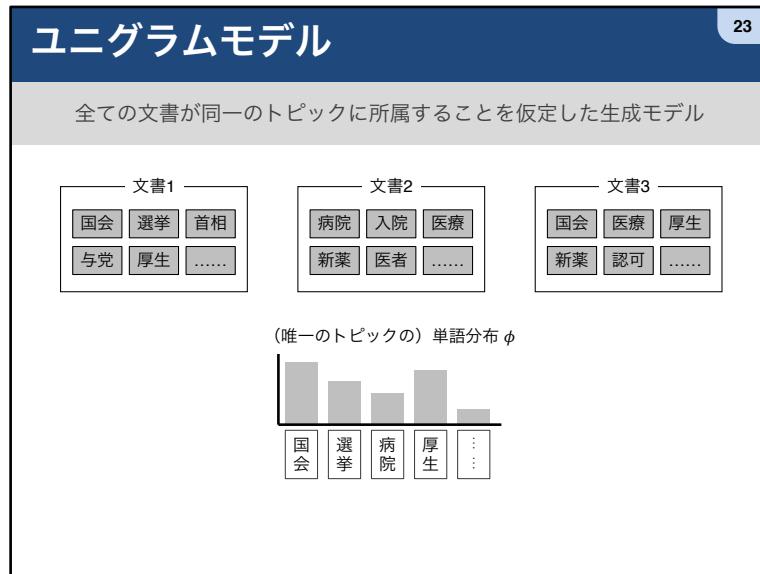
2. 学習

潜在変数と確率分布パラメータを観測データから教師なしで推定

3. 応用

クラスタリング、検索、推薦、... (未知データにも対応可能なモデルも)

前回 | 例 | トピックモデル (+確率的投薬モデル+クリックモデル)



Q&A

適切なトピック数をどう決める？

参考: 第3回で紹介した次元削減手法の場合は…

主成分分析 | 計算方法 | 寄与率

19

- 主成分の第 d 番目の主成分の寄与率
(元データの特徴をどれだけ捉えられているか)
$$C_d = \frac{\lambda_d}{\sum_{d'=1}^D \lambda_d} = \frac{\lambda_d}{\text{tr}(\mathbf{V})}$$

- 寄与率の小さな主成分の追加
› 些末な軸にまで無意味な解釈を与えてしまう
- 寄与率の大きな主成分の除去
› 重要な軸による変動を見逃してしまう

発展 | t-SNE | 利用例 [Borisov+, 2016]

23

ニューラルネットワークの中間層（数百次元）の可視化によく利用
図は元論文より引用

Figure 1: Modeling user browsing behavior on a SERP in the neural click model framework.

Figure 6: Two-dimensional t-SNE projections of the vector state s_t for different distances d to the previous click. Colors correspond to distances: black 0; blue 1; blue-green 2; green 3; green-yellow 4; red 5; grey 6. (Best viewed in color.)

次元数の決定時に
寄与率を参考にできる

可視化目的のために
2次元とすることが多い（印象）

トピックモデルの場合は...?

大きく分けて3つのアプローチ

特化

学習モデルを利用するタスクでの性能を評価

例: 文書のカテゴリ分類のためにLDAの文書-トピック分布を特徴量に使う

→ 評価対象の文書集合に対する分類精度が最も高くなるトピック数を選択

汎用

学習後にモデルの良さを評価

評価指標: Perplexity、Coherence、...

高度

学習時に最適な数を自動で決定

キーワード: ノンパラメトリックベイズモデル、中華料理店過程

Perplexity

モデル \mathcal{M} の評価用データ W^{eval} に対する perplexity の値は…

$$\text{Perplexity}(W^{\text{eval}} | \mathcal{M}) = \left(\prod_{d=1}^{D^{\text{eval}}} \prod_{n=1}^{N_d^{\text{eval}}} \frac{1}{p(w_{dn}^{\text{eval}} | \mathcal{M})} \right)^{\frac{1}{\sum_{d=1}^{D^{\text{eval}}} N_d^{\text{test}}}}$$

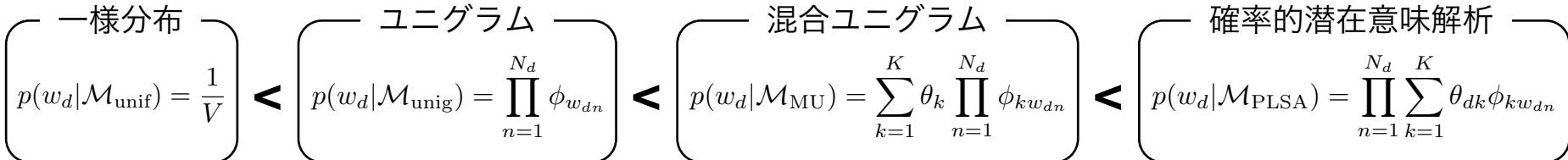
= 確率 $p(w_{dn}^{\text{eval}} | \mathcal{M})$ の逆数の幾何平均



単語の生起確率の逆数 = 予測候補として絞り込める候補の数なので

**Perplexity: 評価用データにおける予測対象の平均候補数
(低いほうが良いモデル)**

一般的な傾向 (perplexity はこの逆順)



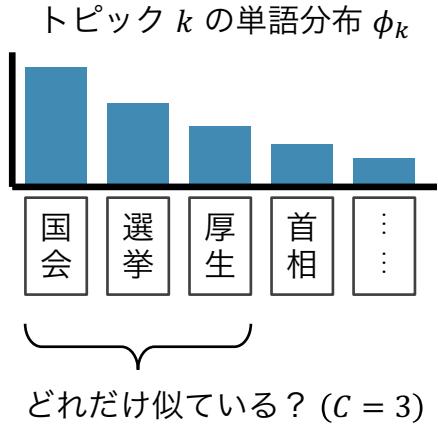
Perplexityを用いたトピック数の決定

1. データセットを訓練用・評価用に分割
2. 各トピック数 K について
 - › 訓練データを用いてモデル \mathcal{M}_K を学習
3. 評価データに対する perplexity が最小となるモデルを選択
 - › $K' = \arg \min_K \text{Perplexity}(W^{\text{eval}} | \mathcal{M}_K)$

Coherence [Newman et al., 2010]

各トピック k の一貫性を C 個の頻出単語集合 $\mathcal{W}_k = \{w_{k1}, \dots, w_{kC}\}$ の平均類似度で評価

$$\text{Coherence}_{\text{sim}}(\mathcal{W}_k) = \frac{2}{C(C-1)} \sum_{1 \leq i < j \leq C} \text{sim}(w_{ki}, w_{kj})$$



↓

全トピックの平均 $\frac{1}{K} \sum_{k=1}^K \text{Coherence}_{\text{sim}}(\mathcal{W}_k)$ が
高いならば良いモデル

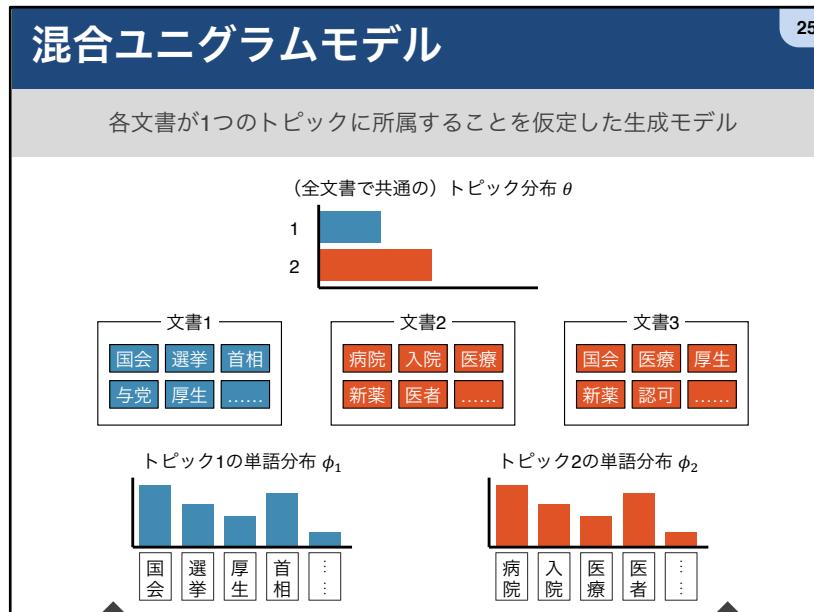
Q. 類似度 $\text{sim}(w_i, w_j)$ はどうやって計算する?

A. Wikipedia中の出現頻度 $n(\cdot)$ から計算した自己相互情報量 $\text{PMI}(w_i, w_j)$ を利用すると良い

$$\text{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad \text{where} \quad p(w) = \frac{n(w)}{\sum_v n(v)} \quad \text{and} \quad p(w_i, w_j) = \frac{n(w_i, w_j)}{\sum_{v_i, v_j} n(v_i, v_j)}$$

ノンパラメトリックベイズ

(大雑把に言えば) 無限個のパラメータを持つことのできる統計モデル



有限次元の場合の各文書の生成過程

1. 文書 d のトピック z_d を K 個の中から選択
2. 各位置 $n = 1, \dots, N_d$ について
 - a. 単語 w_{dn} を z_d の単語分布から選択

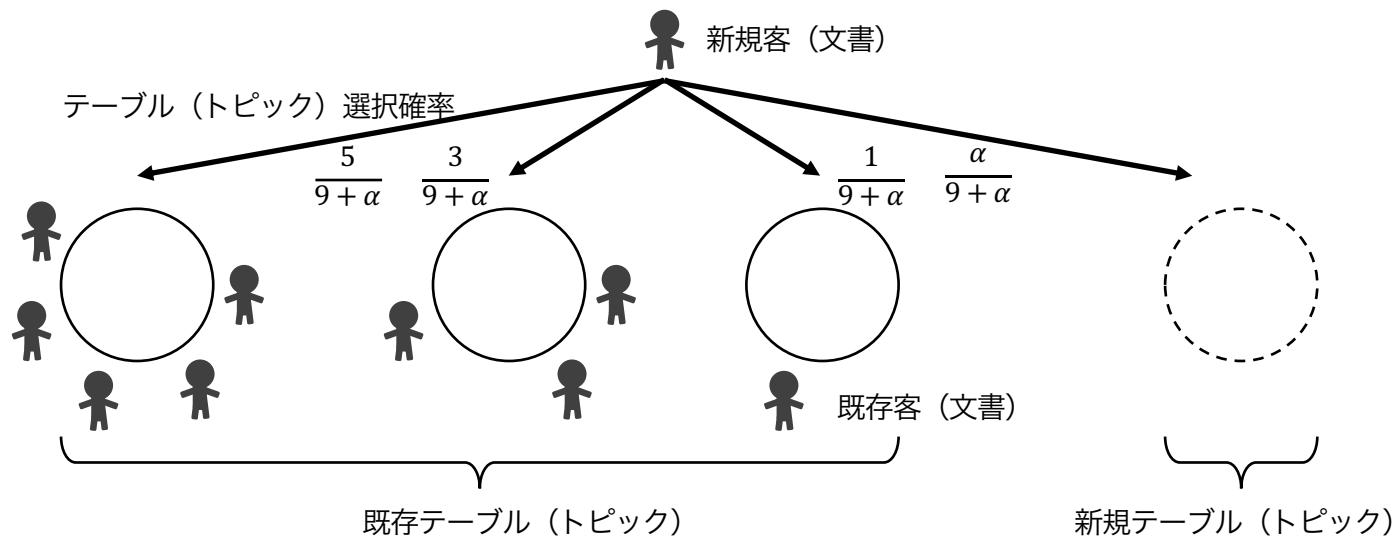
無限次元の場合の各文書の生成過程

1. 文書 d のトピック z_d を K 個の中から選択
or 新たに $K + 1$ 個目のトピックを生成
2. 各位置 $n = 1, \dots, N_d$ について
 - a. 単語 w_{dn} を z_d の単語分布から選択

中華料理店過程

集中パラメータ α の下で、 d 番目の文書のトピック z_d が k である確率は

$$p(z_d = k | z_1, \dots, z_{d-1}, \alpha) = \begin{cases} \frac{\sum_{d'=1}^{d-1} \mathbb{1}[z_d=k]}{d-1+\alpha} & (\text{既存トピック}) \\ \frac{\alpha}{d-1+\alpha} & (\text{新規トピック}) \end{cases}$$



- さらに文書中の出現単語も考慮するので $\left\{ \begin{array}{l} \text{割当文書数の多い} \\ \text{単語分布の似た} \end{array} \right\}$ トピックがサンプリングされやすい
- 既存のどのトピックとも単語分布が似ていない場合、新規のトピックが生成されやすい

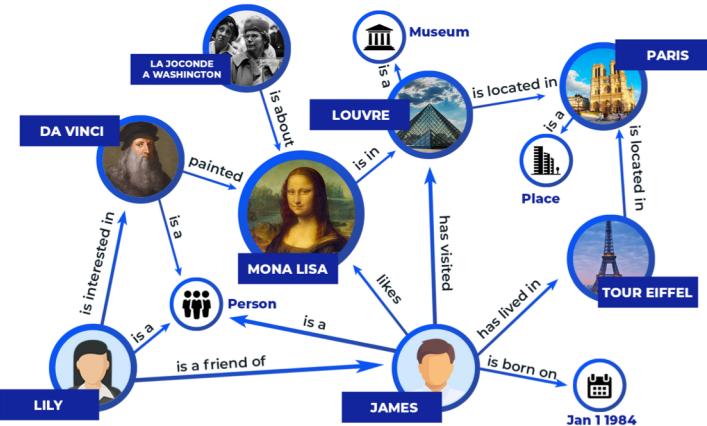
今日の話題

グラフデータ

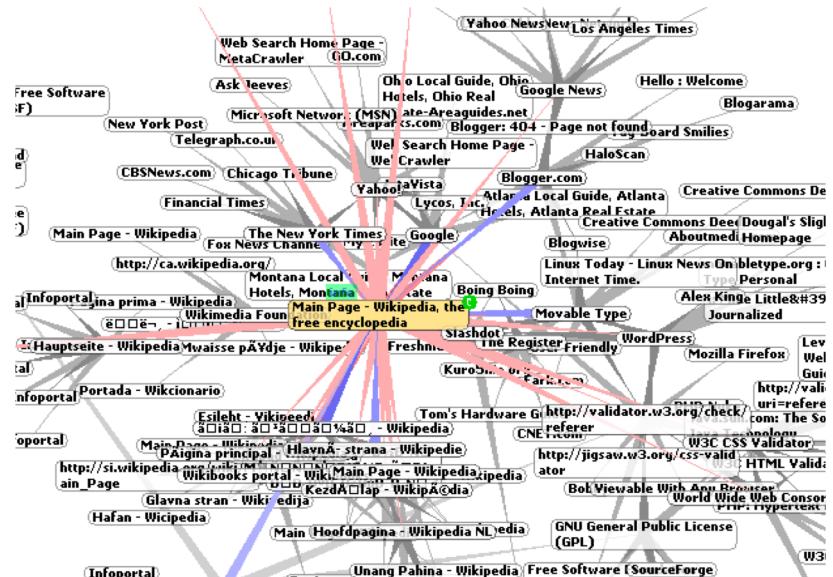
道路ネットワーク



知識ベース



Web (ハイパーリンク)



SNS



Web

検索モデルの大雑把な歴史

ブーリアン検索

- クエリと一致するものを発見
- 検索結果に順序はない

リンク解析

- アンカーテキスト
- PageRank [Page et al., 1999]

ベクトル空間モデル

- tf-idf

確率的検索モデル

- BM25

ランキング学習

- Ranking SVM
- Lambda MART

意味的マッチング

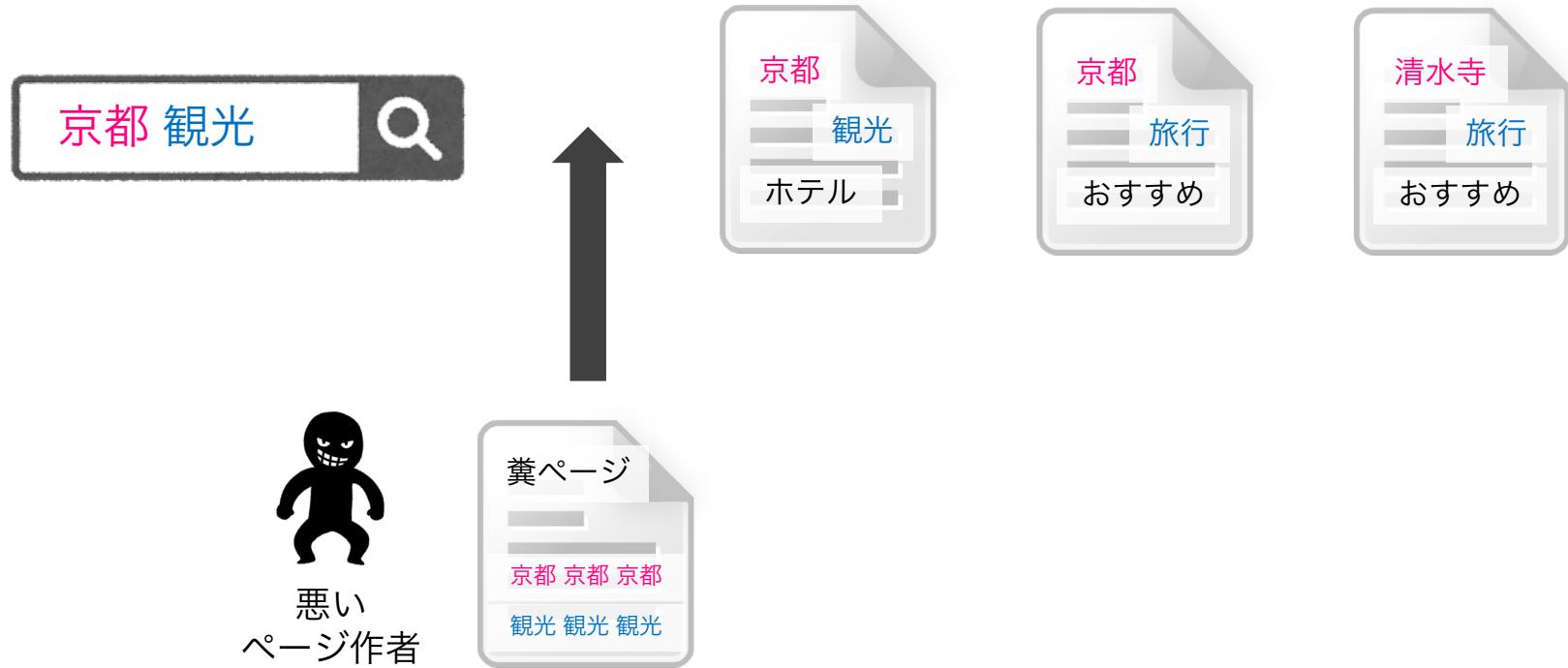
- LSI

ニューラルモデル

- ...

PageRank以前の検索モデル

検索クエリと完全 or 意味的にマッチする単語を含む文書を上位にランキング



問題点

関係のない単語を大量に埋め込むことで文書のトピックと無関係の検索でもマッチさせられる

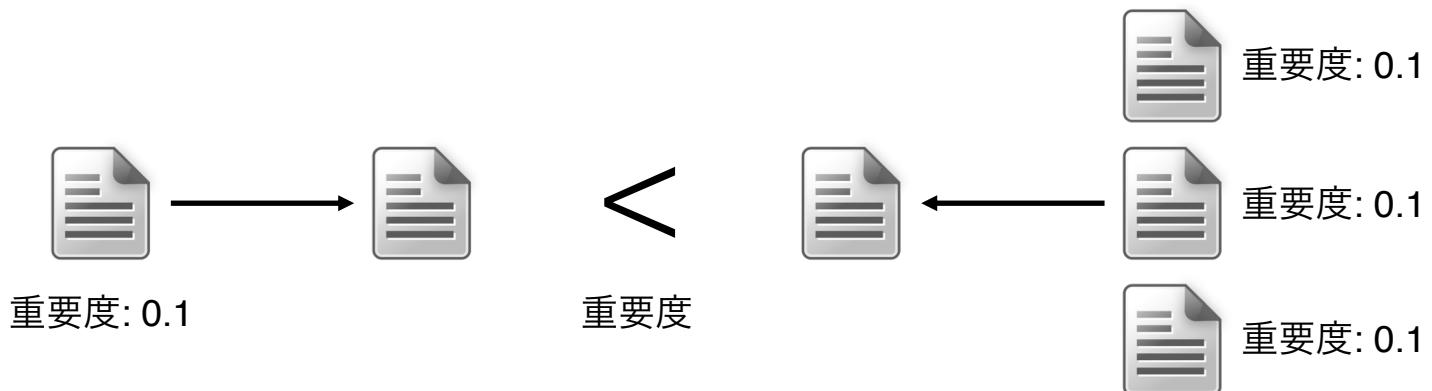
PageRankのアイデア

- 文書内の情報は作者によって自由に操作されてしまう
- 作者が直接的に操作できない文書の被リンクをランキングに利用

仮定1 重要な文書からリンクされている文書は重要



仮定2 多くの文書からリンクされている文書は重要



定義

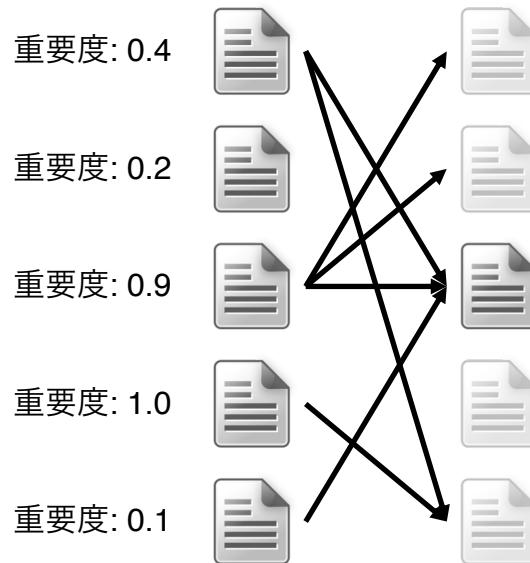
(単純な) PageRankの再帰的定義

$$\text{PR}(d) = \sum_{d' \in \text{pa}(d)} \frac{\text{PR}(d')}{|\text{ch}(d')|}$$

文書 d のPageRank値

文書 d の
リンク元文書集合

文書 d' の
リンク先文書集合



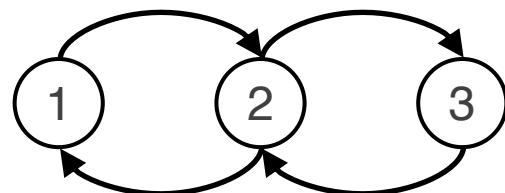
Q. この文書のPageRank値は？

計算

準備

$v = (\text{PR}(d_1), \dots, \text{PR}(d_D))^T$ とおくと前頁の定義は $v = Mv$

$$\left[\text{ただし } M \text{ の } (i,j) \text{ 成分 } m_{ij} \text{ は } m_{ij} = \Pr(j \rightarrow i) = \frac{\mathbf{1}[j \in \text{pa}(i)]}{|\text{ch}(j)|} \right]$$



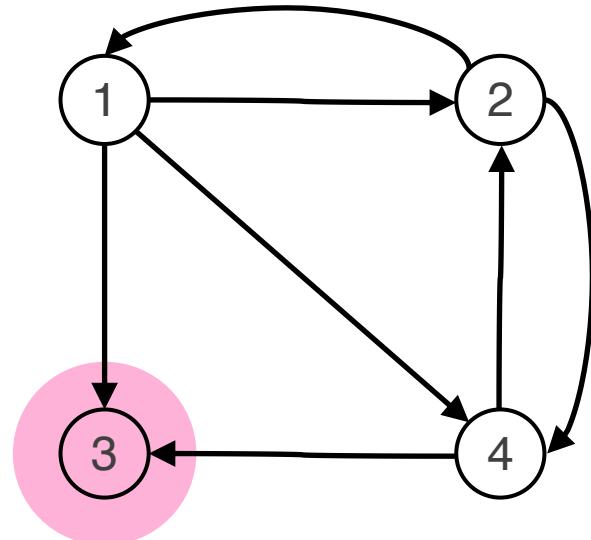
$$M = \begin{pmatrix} 0 & 0.5 & 0 \\ 1 & 0 & 1 \\ 0 & 0.5 & 0 \end{pmatrix}$$

反復

PageRankベクトルの初期値を $v^{(0)} = (\frac{1}{D}, \dots, \frac{1}{D})^T$ として

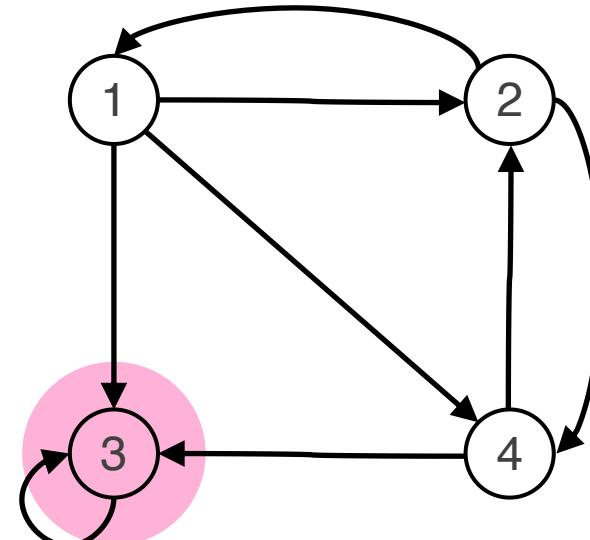
$v^{(n+1)} = Mv^{(n)}$ を収束するまで繰り返す

問題点: 実際のWebは強連結ではない



dead end

Dead endノードがあると
全ノードのPageRank値が0



spider trap

Spider trapノードがあると
当該ノード以外のPageRank値が0

対処: 無理やり強連結に

(問題対処後の) PageRankの再帰的定義

$$\text{PR}(d) = \alpha \sum_{d' \in \text{pa}(d)} \frac{\text{PR}(d')}{|\text{ch}(d')|} + (1 - \alpha) \frac{1}{D}$$

減衰率
(通常0.85)

ランダムサーファーモデルによる解釈

ランダムウォーク

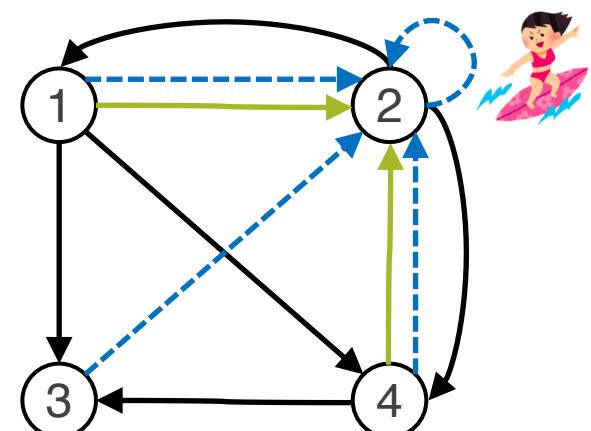
確率 α で無作為にリンクを遷移

ランダムジャンプ

確率 $1 - \alpha$ で無作為に任意の文書に瞬間移動

文書 d の $\text{PR}(d)$

長時間経過後にサーファーが文書 d に滞在している確率

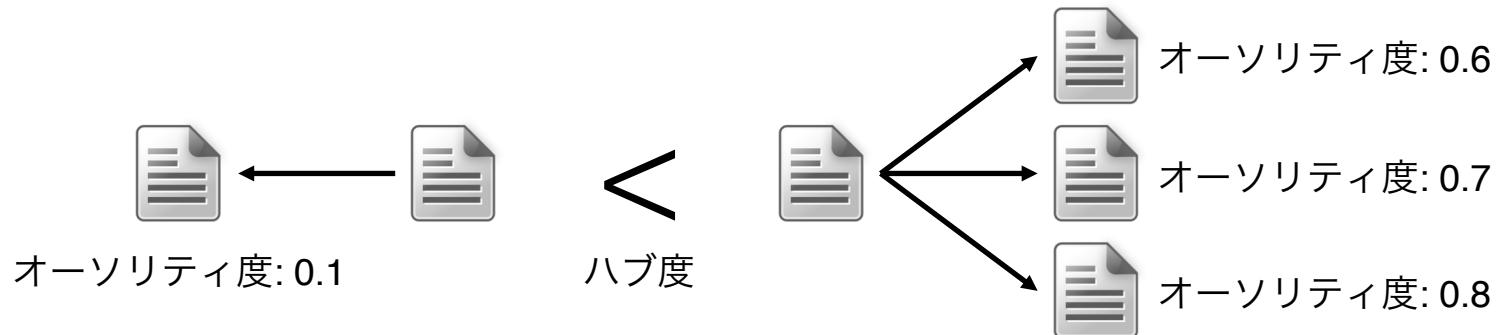


HITS (Hyperlink-Induced Topic Search) [Kleinberg, 1999]²⁴

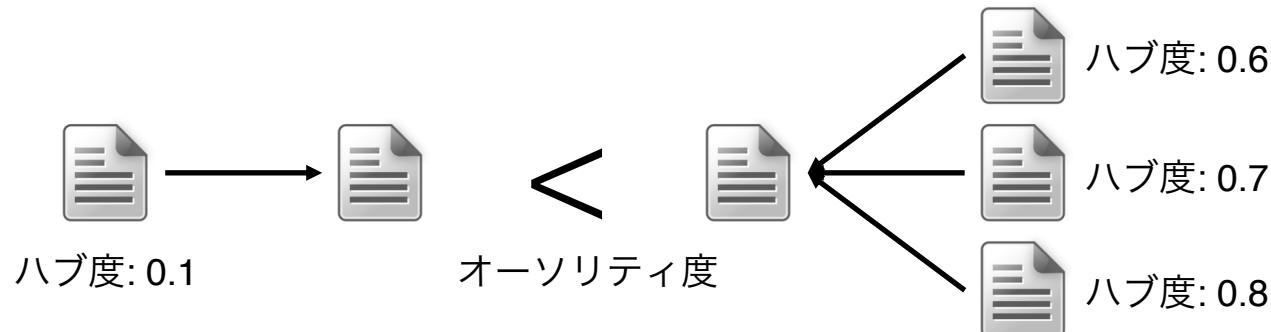
文書の良さとして2種類の性質を考える:

ハブ（良いリンクを持つ文書）とオーソリティ（良い情報を持つ文書）

仮定1 多くの良いオーソリティをリンクしている文書は良いハブ

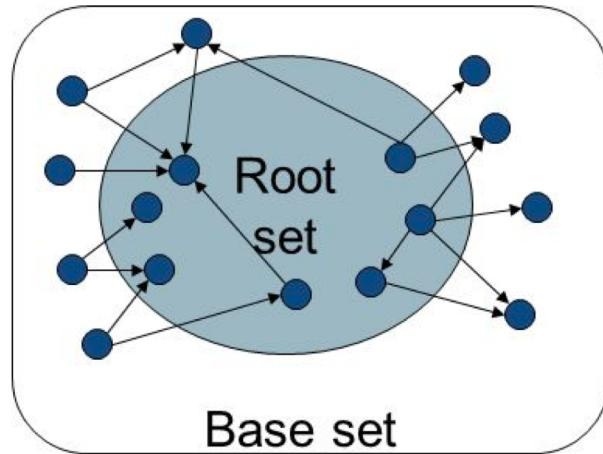


仮定2 多くの良いハブからリンクされている文書は良いオーソリティ



計算

準備



Root set

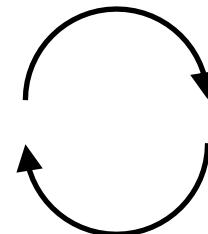
検索クエリにマッチする文書集合を取得

Base set (HITS計算対象)

Root set内の各文書のリンク元／先文書を追加

反復

$$\text{hub}(d) = \sum_{d' \in \text{ch}(d)} \text{auth}(d')$$



$$\text{auth}(d) = \sum_{d' \in \text{pa}(d)} \text{hub}(d')$$

反復の度に正規化

TrustRank [Gyöngyi et al., 2004]

- PageRankによって文書内での操作の影響は軽減できた
 - 仲間内間でリンクし合うことでPageRank値を不当に上げようという動き
- ➡ 文書のスパム判定に必要な労力をリンク解析によって減らせないか？

仮定 良い文書は悪い文書に滅多にリンクしない

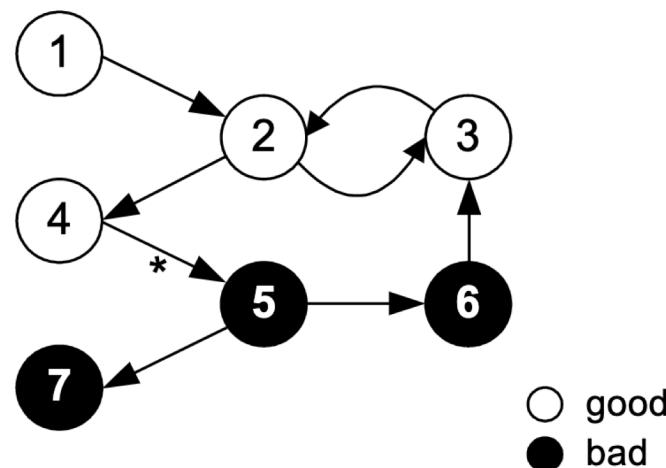
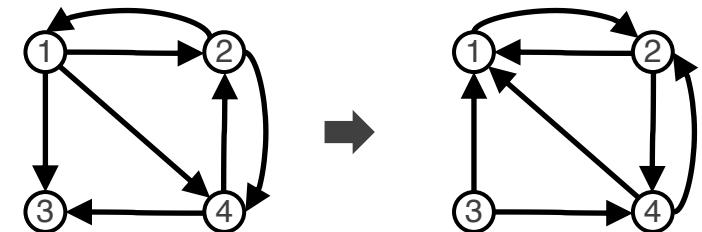


Figure 2: A web of good (white) and bad (black) nodes.

計算

シード

- **Inversed PageRank** (リンクを反転させたグラフに対するPageRank) を計算
- 以降のステップで広範囲にスコアを伝搬可能な文書を発見

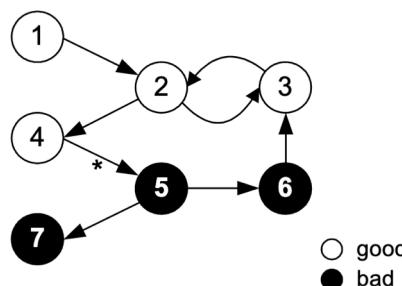


判定

IPR値が上位の文書集合に対して人手でスパム判定

反復

元のグラフ上で**Biased PageRank**を計算することでトラストスコアを伝播



ランダムウォーク

PageRankと同じ（確率 α で無作為にリンクを遷移）

ランダムジャンプ

確率 $1 - \alpha$ で非スパムのシード文書に瞬間移動

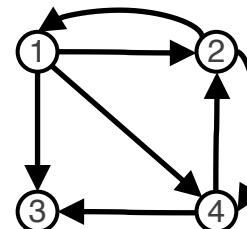
SNS

TwitterRank [Weng et al., 2010]

影響力のあるTwitterユーザをtopic-sensitive PageRankの改良により発見

グラフ

- ノード: Twitterユーザ
- エッジ: follower → followee



ユーザ2はユーザ4をフォロー

トピック

各ユーザのツイート集合を1つの文書にまとめてLDAを学習

計算

トピック t ごとにPageRankの亜種を計算

ランダムウォーク

$$P_t(i, j) = \frac{|\mathcal{T}_j|}{\sum_{a: s_i \text{ follows } s_a} |\mathcal{T}_a|} * \text{sim}_t(i, j)$$

ユーザ s_a の
投稿ツイート数

ユーザ s_i と s_j の
ツイートトピックの類似度

ランダムジャンプ

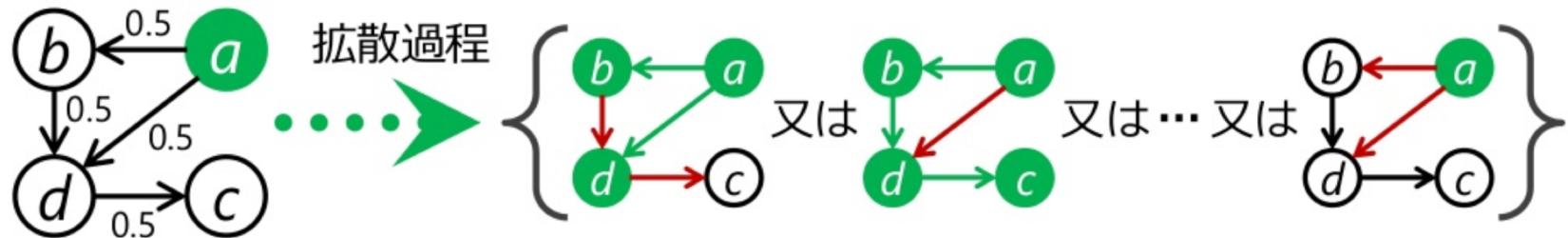
トピック t に関するツイート数に比例する割合で各ユーザに瞬間移動

Influence Maximization [Kempe et al., 2003]

SNS上での拡散行動のモデル化・予測・制御 → マーケティング

$$\operatorname{argmax}_{A:|A|=k} \mathbf{E}[A \text{ の引き起こす拡散のサイズ}]$$

(確率的モデルに従う)



劣モジュラ関数の枠組みで定式化

→ 貪欲算法が $(1 - e^{-1})$ 近似

[Nemhauser-Wolsey-Fisher. *Math. Program.*'78]

- ▶ 効率的計算の追求
- ▶ 異なるモデル
- ▶ 新たな問題設定

Random Walk Controversy [Garimella et al., 2016]

SNSにおける論争度をグラフ内のランダムウォークにより定量化

グラフ

- 所与のトピック（ハッシュタグ）に関する投稿を収集
- フォローやRT関係に基づき投稿者間にエッジを作成

分割

グラフ分割アルゴリズムMETISによりノード集合を2分割

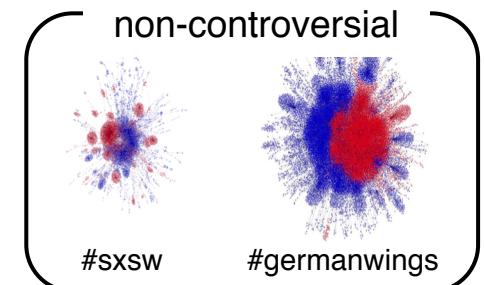
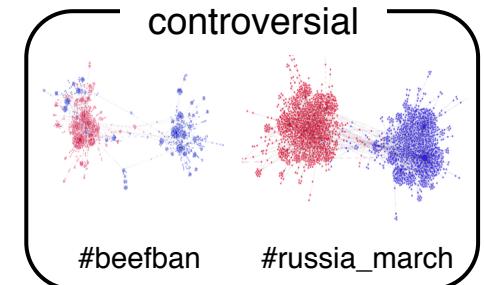
計算

仮定

論争化が進むと **グループ間交流** が **グループ内交流** に比べて少なくなる

$$\text{RWC} = P_{XX}P_{YY} - P_{YX}P_{XY}$$

$$P_{AB} = P(\text{分割 A のノードを始点} \mid \text{分割 B のノードを終点})$$



まとめ

WebやSNSのグラフデータ分析に利用される手法の紹介

- **PageRank**

- › Googleの創始者によって開発された
- › ランダムウォークとランダムジャンプから構成

- **～Rank**

- › ランダムウォークかランダムジャンプの一方 or 両方を修正していることが多い
- › グラフデータに対して再帰的な仮定がある時に使いやすい

- **HITS**

- › Web文書だけでなく2部グラフデータとの相性も良い
 - Yahoo！知恵袋の質問者と回答者
 - Flickrのユーザとタグ

参考文献

- David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 100-108.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
- Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 5 (September 1999), 604-632. DOI: <https://doi.org/10.1145/324133.324140>
- R. Lempel and S. Moran. 2001. SALSA: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.* 19, 2 (April 2001), 131-160. DOI: <https://doi.org/10.1145/382979.383041>
- Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In Proceedings of the Thirtieth international conference on Very large data bases - Volume 30 (VLDB '04), Mario A. Nascimento, M. Tamer Ozsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer (Eds.), Vol. 30. VLDB Endowment 576-587.
- Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining (WSDM '10). ACM, New York, NY, USA, 261-270.
DOI=<http://dx.doi.org/10.1145/1718487.1718520>
- David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03). ACM, New York, NY, USA, 137-146.
DOI=<http://dx.doi.org/10.1145/956750.956769>
- Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16). ACM, New York, NY, USA, 33-42. DOI: <https://doi.org/10.1145/2835776.2835792>