社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

マイクロブログにおける対話ネットワークと投稿内容を併用したユーザ推薦に関する一考察

岡本　大輝†　　豊田　正史†　　喜連川　優†,††

† 東京大学大学院情報理工学系研究科
† 東京大学大学院情報理工学系研究科
† 東京大学大学院情報理工学系研究科，国立情報学研究所
E-mail: †hokamoto@tkl.iis.u-tokyo.ac.jp

あらまし　ソーシャルネットワーク上でユーザに他のユーザを推薦するユーザ推薦は，多くのソーシャルメディアにおいて提供されている．ユーザ推薦は，ユーザ間のコネクション形成を促進するうえで重要であり，ユーザは多くのユーザとコミュニケーションを取り，効率的な情報収集を行うことが可能になる．ユーザ推薦はグラフにおけるリンク予測問題の一種であり，ソーシャルネットワークのグラフ構造に着目した手法，発言内容に基づく手法等が提案されているが，これらを併用することでさらなる精度向上が期待される．本論文では，マイクロブログサービスの１つである Twitter 上において，ユーザ間の対話関係を表すグラフ構造を用いた Random Walk 手法と，発言内容の類似度を併用した手法を 2 種類提案しその精度を比較した．
キーワード　ユーザ推薦，ランダムウォーク，マイクロブログ，

User Recommendation in Microblogs based on Interaction Networks and Contents

Hiroki OKAMOTO†, Masashi TOYODA†, and Masaru KITSUREGAWA†,††

† Graduate School of Information Science and Technology,The University of Tokyo
† Graduate School of Information Science and Technology,The University of Tokyo
† Graduate School of Information Science and Technology,The University of Tokyo, NII
E-mail: †hokamoto@tkl.iis.u-tokyo.ac.jp

Abstract　User recommendation on social networks is one of the most important service on social media. It generates more links between users, and users can communicate with more friends and efficiently correct information. User recommendation is one of the link prediction problems. There have been several methods based on graph structure and contents of users' posts, but these features have not been well combined in previous methods. In this paper we propose two user recommendation methods that utilizes a random walk technique combined with contents similarities. We apply these methods to a Twitter dataset and compare their accuracy.
Key words　user recommendation, random walk, micro blogs, tf-idf,

1. ま え が き

近年，社会生活における Web の重要性が増し，人々は Web 上でも社会活動や経済活動を行うようになった．そうした中で，web ユーザに対する情報推薦は，Web サービスにおいて非常に重要な技術となっており，その精度を高める研究は数多くなされている．例えばユーザに商品を推薦することで，購買行動が活発になる．また，ユーザにユーザを推薦することで，ユーザ間にコネクションが生じ，サービスの利用率を高めることが可能となる．

一方で，web 自体の肥大化と多様化に伴い，web ユーザの数が増加し，web の使い方も様々に変化・多様化している．そうした中で，推薦の手法もまた，様々なサービスに適応する必要がある．このような環境の中で，これまで盛んに研究されてきた，ソーシャルネットワークのグラフ構造のみ着目するのではなく，グラフに依拠しない新たな指標に基づく推薦手法と併用を試みる．

そこで本稿では，ソーシャルネットワークにおけるユーザ推

薦を目的とし，ソーシャルネットワークのグラフ構造と，グラフに因らない特徴としてユーザが投稿するテキストの内容の両方に着目し，それらを併用したユーザ推薦手法の実験と，その成果を示す．

以降，第 2 章では，ユーザ同士の関係から構築されるグラフに着目したユーザ推薦手法や，ユーザの特徴に基づいたユーザ推薦手法について，先行研究の紹介を行う．それらを踏まえて第 3 章では，Twitter におけるユーザ推薦を目的とし，今回実施した実験に使用したデータの詳細について述べる．第 4 章では，グラフ構造とツイート内容を併用した今回の実験の詳細と，実験結果の評価方法について述べ，第 5 章でその結果と考察について述べる．第 6 章では，それらを総括し，今後の課題について述べる．

2. 関 連 研 究

近年，Web 上においても，様々なコンテンツやユーザを，効率的に，かつ的確に推薦する手法について，多種多様な研究が行われている．特に，特定の環境下のユーザに対して，同一のサービスを利用するユーザを推薦する際には，そのネットワーク構造に着目したアプローチをされることがある．

2.1 ユーザ推薦における一般論

ある対象に対し，同一の環境下 (同じ学会に属している，同じサービスを利用している，等) の別の対象を推薦することは，対象間のリンク予測問題に帰着する．具体的には，同一の環境下にある二つの対象の間に，近い将来何らかの関係性 (リンク) が構築されると予測した場合，対象のそれぞれに相手を推薦するというものである．同一環境下でのリンク予測の活用例，ないし実験や検証に用いる環境として，SNS への利用や生体ネットワークへの適用が代表的である [5].

SNS においては，同一サービスにおいて利用者間の関係性を分析し，2 者が将来的にリンク関係 (主に友人関係) を形成するか予測し，高評価となったユーザを推薦する．生体を構成するタンパク質同士の相互作用の有無を分析する．バイオインフォマティクス分野で盛んに研究されている．

以下では，データ構造の表現に則り，推薦候補および被推薦の対象となるオブジェクトをノード，オブジェクト同士のリンク関係をエッジと表現する．

また，ノード A とエッジを形成しているノードの集合を $\Gamma_{(A)}$ として，$\Gamma_{(A)}$ の要素のノードをノード A の隣接ノードと表現する．

2.2 様々な推薦手法の紹介

ユーザ推薦においては，グラフを利用した様々な手法が考案されている．

最もシンプルな推薦手法として，2 つのノードに共通する隣接ノードの数を数えて，数が多い順に推薦するというものがある [2] [1]．また，共通する隣接ノードの数に，分母として両者の隣接ノードの集合の論理和をとった Jaccard's coefficient という指標がある [2]．推薦候補に隣接ノードが多い場合，無用にスコアが高くなることを防ぐ．

ノード A とノード B に共通する隣接ノードであるノード C の隣接ノードが少ないほど，C による A と B 近縁性が高まるとした Adamic Adar の評価指標がある [3]．Jaccard's coefficient と同じく，多数の隣接ノードを保有するノードによって，近縁性が不当に高まることを抑制することができる．

$$Score(A,B) = \sum_{\Gamma_{(A)} \cap \Gamma_{(B)}} \frac{1}{log(\Gamma_{(C)})} \tag{1}$$

推薦の対象となるノードから離れたノードも評価できる指標として，ノード A からノード B へ，距離 l で移動する経路の集合 $paths^{(l)}_{A,B}$ を用いた Katz の指標がある [11].

$$Score(A,B) = \sum_{l=1}^{\infty} \beta^l \cdot |paths^{(l)}_{A,B}| \tag{2}$$

一般的に，β は 0 以上 1 未満の値をとり，β が小さいほど，A, B 間の短距離経路が重視される．

2.3 Random Walk

推薦される対象 (target) となるノードを s として，s の隣接ノードへとランダムで移動する動点の挙動を考える．動点はエッジに従ってさらに隣接ノードへの移動を繰り返し，最終的に動点が存在する確率が高いノードを s に推薦する [4]．s と同じ距離に存在する複数の候補ノードの中でも，より s に近しい対象を推薦する手法として，よく用いられている．

2.4 Supervised Random Walk

Backstorm らは，facebook において友人関係をエッジとして，エッジの形成時間，インタラクションの回数，友人申請の方向などをエッジの特徴量として，特徴量を元に機械学習を行い，エッジにウエイトを付与してランダムウォークを行う Supervised Random Walk を提案した．[7]

エッジの特徴量は多岐にわたるが，マイクロブログにおいては主要なコンテンツとなる，ユーザの発言・投稿の内容については触れられていなかった．

2.5 Content Based Method

数は少ないものの，グラフ構造に着目せず，ユーザの特徴によってユーザ推薦を行う試みもある．Hannon らは，Twitter におけるユーザ推薦において，ユーザのツイート内容と，フォロー/フォロワーリストをユーザの特徴量と見做し，content based の手法で推薦を行った．[10]

結果的に，2-hop(隣接ユーザの隣接ユーザ) まではフォロー/フォロワーリストが有効に作用するが，より広範なグラフ構造には言及していなかった．

3. 提 案 手 法

Random Walk を用いた推薦手法は，ネットワーク構造を用いて近接性の高いユーザをランキングすることにより推薦を行うが，これまでに提案された手法では，ユーザの発言内容までは考慮されていなかった．また，発言内容を用いた推薦手法ではネットワーク構造が十分に考慮されていなかった．そこで、,

論文では，Random Walk による推薦手法を基にして，発言内容を考慮した推薦を行うよう拡張する手法を 2 種類提案し比較実験を行う．

1 つ目の手法は、Random Walk における遷移確率に，発言内容の類似度を考慮した重み付けを行うものであり，2 つ目の手法は、Random Walk によるランキングに発言内容の類似度によるランキングを混合する手法である．以下にその詳細を述べる．

3.1 発言内容の類似度を考慮した Random Walk

本手法は、Random Walk 手法における遷移確率を，ユーザ間のメンション回数や発言内容などを考慮して重み付けを行う．Random Walk 手法としては，推薦対象となるユーザ s を起点とし，友人関係のリンクをたどる Randam Walk を行い，各ステップにおいて一定の確率で s に戻る，Random Walk with Restart 手法 [4] を基にする．遷移確率行列を M とすると，定常確率のベクトル$\vec{r}$は以下で表わされる．

$$\vec{r}_{k+1} = (1-\alpha)M\vec{r}_k + \alpha\vec{r}_0 \tag{3}$$

ただし，$\vec{r}_0$ は，ユーザ s のみ 1 で残りは 0 となるようなベクトルである．α は，各ステップにおいて s に戻る確率を表し，今回の実験では Supervised Random Walk 手法 [7] において用いられていた 0.3 を値として用いた．

遷移確率には，各隣接ノードに等確率で遷移する一様な重み付けに加え，以下の 2 種類を用いる．

- メンション回数によるウエイト

ユーザ間のメンション回数は親密性を表すため，回数が多いほど高い確率で遷移するような重み付けを行う．ユーザ A，ユーザ B 間のメンション回数を $n_{(A,B)}$ として，ウエイト $w_{m(A,B)}$ を以下のように定義する．

$$w_{m(A,B)} = 1 - \frac{1}{1+n_{(A,B)}} \tag{4}$$

- ユーザ間の発言内容の類似度を用いた重み付け

各エッジにおいて両端のユーザの発言内容から名詞を抽出し，$tf \cdot idf$ で重みづけした特徴ベクトル $\vec{t}_A, \vec{t}_B$ を用いて，コサイン類似度に基づくウエイト $w_{t(A,B)}$ を以下のように定義する．

$$w_{t(A,B)} = \frac{1}{2}\left(1 + \frac{\vec{t}_A \cdot \vec{t}_B}{|\vec{t}_A||\vec{t}_B|}\right) \tag{5}$$

以上の重み $w_{m(A,B)}, w_{t(A,B)}$ を線形結合して，エッジの重みを以下のように計算する．

$$w_{(A,B)} = \begin{cases} \beta w_{m(A,B)} + (1-\beta) w_{t(A,B)} \\ 0 \quad (A, B \text{ 間にエッジが無い場合}) \end{cases} \tag{6}$$

遷移確率行列においては，あるユーザから隣接ユーザに対して，この重みに比例した確率で遷移を行うよう正規化を行う．

3.2 ランキング混合手法

前項の Random Walk において，名詞のコサイン類似度を考慮しない，すなわち $\beta = 0$ として得られた，各推薦候補ノードの定常確率を $\vec{r}$ として，ターゲットユーザ A に対する推薦候補ユーザ B のスコアを，

$$Score(A,B) = \beta \log(r_B) + (1-\beta)\log(cos(A,B)) \tag{7}$$

として，全ての推薦候補ユーザに対してスコアを付し，A に推薦する．

図 1 と図 2 に，全てのターゲットユーザ (90 ユーザ) に対する推薦候補ユーザ (約 22,000 ユーザ) の，定常確率とコサイン類似度の分布を示した．横軸を対数にすることで山型を描くため，定常確率とコサイン類似度の対数の値を使用した．

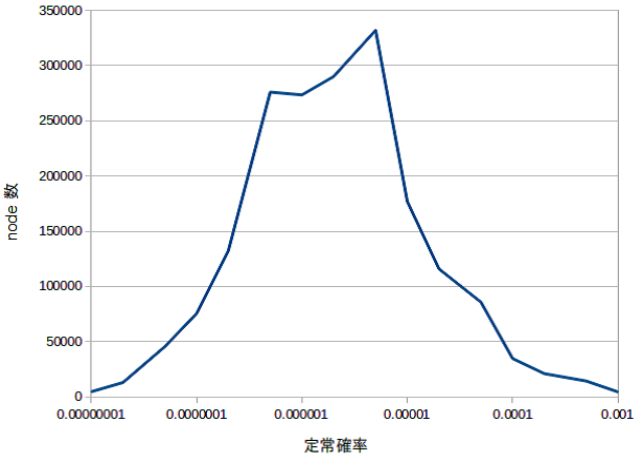

図 1　Random Walk ($\beta = 0$) における
ノードの定常確率と分布

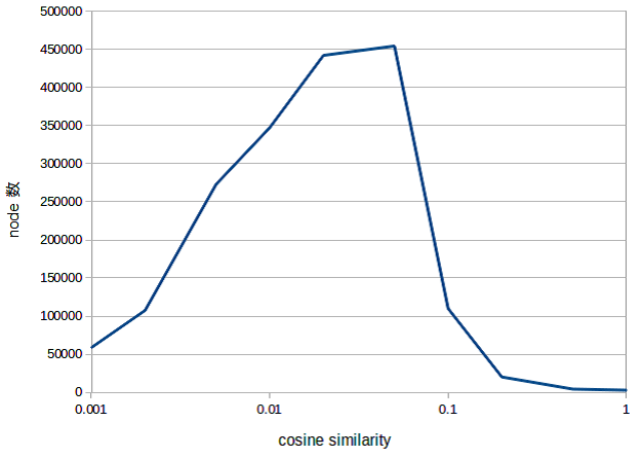

図 2　各ノードの cosine similarity と分布

3.3 評 価 方 法

Random Walk，Ranking based method のそれぞれで，prec@10，AUC の 2 つの指標を用いて精度を評価する．

prec@10

スコアが高くなった上位 10 ユーザのうち，後半期間でエッジが形成されるユーザの平均数．今回のデータでは，推薦のターゲットユーザが後半期間に新たに 10 以上のエッジを形成するため，推薦の精度の評価としてこの値を使用する．

AUC

推薦候補のスコアとランキングが与えられた際に，スコアの閾値を変えながら，閾値を超えたユーザを positive，閾値以下のユーザを negative として，表 1 における

$$\frac{TRUEpositive}{TRUEpositive + FALSEnegative} \tag{8}$$

を縦軸の値，

$$\frac{FALSEpositive}{FALSEpositive + TRUEnegative} \tag{9}$$

を横軸の値として，ROC 曲線をプロットする．

表 1　Candidate user division

	後半期間にエッジが形成される	後半期間にエッジが形成されない
スコアが閾値以上	TRUEpositive	FALSEpositive
スコアが閾値未満	FALSEnegative	TRUEnegative

ROC 曲線と横軸が為す面積が AUC であり，一般的に 0.5 以上の値となる．prec@10 がランキングの上位のみを考慮することに対し，AUC はランキングの下位の推薦候補も加味する．

4. データセット

今回の実験において，Twitter 社が提供する API を用いて，ツイートを収集した．著者の使用している Twitter アカウントを起点として，ソーシャルグラフ上で近傍に存在するユーザがフォローしているユーザを順次集めて，最終的に 60,000 ユーザの tweet データを取得した．

その中で，特に 2012 年 10 月から同年 12 月まで (以下，前半期間) のデータと，2013 年 1 月から 2013 年 3 月まで (以下，後半期間) のデータを抽出した．それぞれでデータで，ユーザをノードとして，ユーザが別のユーザに mention (@(ユーザ名) の文字列を含む tweet) をした，あるいは受け取った場合にエッジが形成されるものとした．

以上の方法でグラフを作成し，以下の条件を満たすユーザの間で，推薦を行う．

- 自動投稿アカウントではない
- 前半期間で 100 件以上の tweet を取得できた
- 前半期間でトラッキング対象ユーザの中に隣接ユーザが存在する

さらに，上記条件を満たすユーザのうち，後半期間で新たに隣接ユーザとなるユーザが 10 以上存在するユーザをアクティブなユーザとして，ランダムに 90 ユーザを抽出し，推薦の target user とした．

なお，事前分析として，後半期間に新たにエッジを形成する (後半期間で初めてメンションを行う) ユーザが，前半期間のグラフではどれほどの距離であったのか，を調査した (図 3)．

図 3 より，今回のデータで新しく形成されるエッジは，2-hop に 40%程度が含まれ，3-hop までで 80%程度が含まれる．新しく形成されるエッジが 2-hop に集中していると，common neighbors，jaccard's coefficient 等も有用であるが，今回は 3-hop 以上の探索が必要と考えられる．なお，ターゲットユーザが含まれるグラフから孤立しているユーザにエッジが形成された場合は，横軸の値を-1 としてプロットしている．

参考として，Supervised Random Walk の実験 [7] では，2-hop までに 90%のユーザが集中しており，近傍ノードの重要性に大きな違いがあることがわかる．

表 2　Dataset detail

	all user	candidate user	target user
ユーザ数	46884	22422	90
前半期間の平均隣接数	8.83	11.31	43.64
後半期間の平均隣接数	7.89	10.24	56.71
後半期間で新たに観測される隣接数	1.06	2.21	16.70

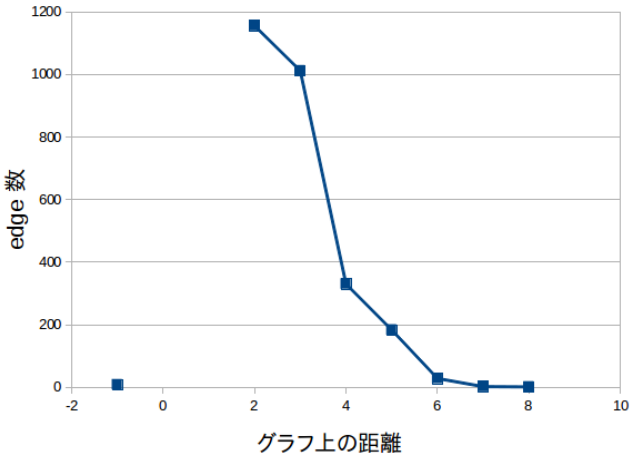

図 3　グラフ上距離別の形成エッジ数

5. 実験結果と考察

Random Walk(RW)，Rankin based method(RBM) のそれぞれで，β の値を変えながら prec@10 と AUC の値を調べた．なお，ベースラインとしてエッジにウエイトをかけない一般的な Random Walk(RW) の結果も併記する．

表 3　prec@10 and AUC

	prec@10	AUC
base line	1.395	0.9165
RW($\beta = 1.0$)	1.430	0.9163
RW($\beta = 0.56$)	1.500	0.9168
RBM($\beta = 0.86$)	1.488	0.9186

図 4 を見ると，$\beta > 0.5$ の範囲ではいずれの手法にも大きな変化は見られないが，$\beta < 0.5$ の範囲，すなわち名詞コサイン類似度による影響を強くした場合は，精度が大きく落ちていることがわかる．その場合 Ranking based では，グラフ上の距

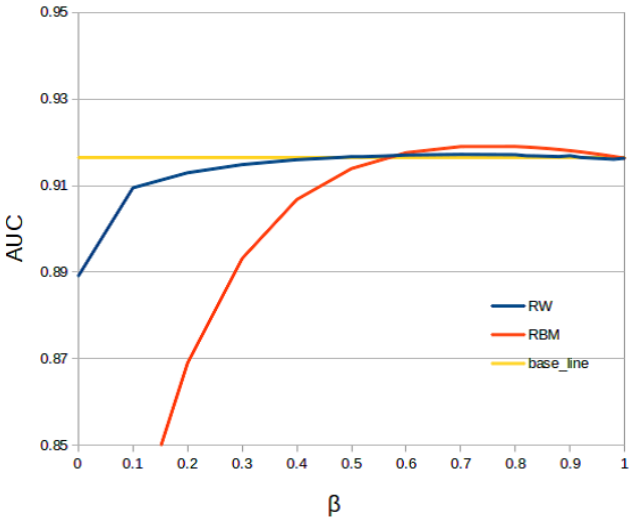

図 4　β の値による AUC の変化

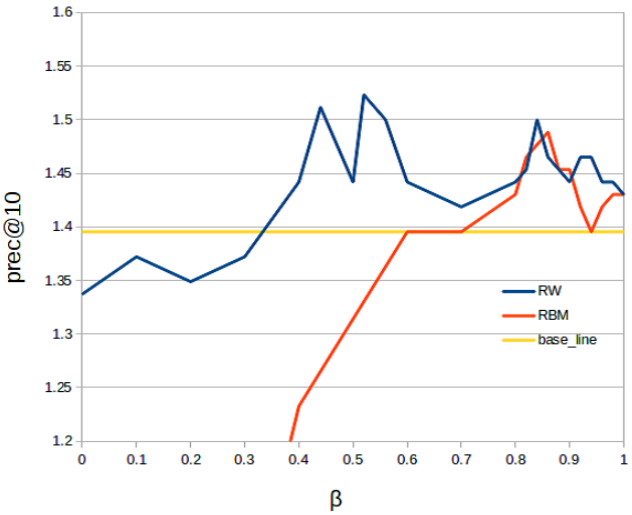

図 5　β の値による prec@10 の変化

離を問わず，ツイート内容が似ているユーザが優先的に推薦されるが，Random Walk のほうが推薦の手法としては妥当であることを示している．

図 5 を見ると，prec@10 はベースラインに対して最大 1.4%ほどの向上が見られた．RW，RBM のいずれでも $\beta = 1.0$ の場合が最大値とならないことから，Random Walk による推薦に，テキストのコサイン類似度による評価を加味することに，意義があると述べることができる．

しかしながら，precision も 15%止まりであり，依然として改善の余地が見受けられる．また，今回は β の値を細かく変化させながら精度を見積もったが，特に prec@10 は β の値が少し変化するだけで乱高下することがあり，的確な変数の設定方法が課題である．

6.　今後の課題

以上の結果を元に，今後の研究課題及び改善点を述べる．

今回は個々のエッジ形成 (ユーザが新しく別のユーザにメンションをする行為) の原因には触れず，グラフ構造や発言内容を機械的に処理して，推薦の精度の向上を試みた．今後はエッジ形成が発生する場面を抽出し，メンションを発する/受ける理由を考察し，その際の特徴を利用することで，推薦の精度の向上につなげたい．

今回は Twitter の API の仕様上，サンプリング期間のメンション回数と発言内容を特徴量として利用したが，発言の頻度，時間帯，following/followed の関係，Twitter の機能であるハッシュタグ等，様々な特徴量の活用が考えられる．また，正解たる推薦候補の特徴，高スコアとなりながら不正解となる推薦候補の特徴等，より詳しく特徴を分析し，精度向上の手掛かりを得たい．

今回の Random Walk の手法では，β の値を固定して，90 のターゲットユーザ，22,000 の推薦候補に対して，3.2GHz のシングルプロセッサで約 6 時間の計算を要した．実際には，より多くのユーザの間で推薦をすることになるため，何らかの高速化が必要である．

文　　献

[1] M.E.J.Newman, Clustering and preferential attachment in growing networks, Physical Review E - Volume 64, 2001.
[2] G. Salton, M. J. McGill, Introduction to modern information retrieval, 1986.
[3] Lada A Adamic, Eytan Adar, Friends and neighbors on the Web, Social Networks, 2003.
[4] Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, 1998.
[5] David Liben-Nowell, Jon Kleinberg, The Link-Prediction Problem for Social Networks, The American Society for Information Science and Technology, 2007.
[6] Jonathan L. Herlocker, Joseph A. Konstan, Lorgen G. Terveen, John T. Riedl, Evaluating collaborative filtering recommender systems, ACM Transaction on Information Systems, 2004,
[7] Lars Backstorm, Jure Leskovec, Supervised random walks: predicting and recommending links in social networks, WSDM, 2011,
[8] Fouss F., Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation, IEEE Transactions on Knowledge and Data Engineering, 2007.
[9] Jure Leskovec, Daniel Huttenlocher, Jon Kleinberg, Predicting Positive and Negative Links in Online Social Networks, WWW, 2010.
[10] John Hannon, Mike Bennett, Barry Smyth, Recommending twitter users to follow using content and collaborative filtering approaches, RecSys, 2010.
[11] Katz L., A new status index derived from sociometric analysis, Psychometrika 18(1), 1953.
[12] Glen Jeh, Jennifer Widom, SimRank: a measure of structual-context similarity, KDD, 2002.