DEIM Forum 2019 P1-125

ジオタグ付きツイートを用いた方言分布の可視化

小笠原奈々† 王 元元† 河合由起子††,†††

† 山口大学工学部知能情報工学科 〒 755–8611 山口県宇部市常盤台 2-16-1
†† 京都産業大学情報理工学部 〒 603–8555 京都府京都市北区上賀茂本山
††† 大阪大学サイバーメディアセンター 〒 567–0047 大阪府茨木市美穂ヶ丘 5 番 1 号
E-mail: †{v017ff, y.wang}@yamaguchi-u.ac.jp ††kawai@cc.kyoto-su.ac.jp

あらまし 近年，SNS の普及によりテキストによるコミュニケーションが促進されてきた．日本では特に Twitter はくだけた言葉でコミュニケーションをとることができる．日本各地に住む多くの人が Twitter を利用しているにも関わらず，行ったことのない地域に住む人たちの人柄の傾向を把握することは困難である．そこで本研究では，引越し先や旅行先を選ぶ指標にするために，各地域のツイートに含まれている方言に着目し，方言と位置情報に基づく異なる地域同士の関連性発見および可視化を目指す．本論文では，日本国内のジオタグ付きツイートを対象とし，まずは地域ごとのツイートからユーザの出身が顕著に現れる方言を抽出する． 次に，自分の慣れ親しんだ都道府県などに対し，その都道府県の方言の発言場所を観測しマッピングすることで，方言が使われている地域を発見できるシステムを提案する．最後に，ツイートやタグクラウドを用いた可視化システムを構築し，被験者アンケートによる可視化システムの有用性を図る評価実験を行った．

キーワード Twitter，方言，ジオタグ付きツイート

1. はじめに

近年，SNS の普及によりテキストによるコミュニケーションが促進されてきた．日本では多くの SNS を用途によって使い分ける傾向がある．特に Twitter(注1)はくだけた言葉を使やすい SNS の 1 つとして多くの日本人が利用している．Twitter の投稿には喋り言葉をそのままテキストにしたような投稿が多いため，方言を含む喋り言葉が多く存在する．

しかし，SNS が普及するまで方言についての研究は現地の人々を取材することでしか行うことができず，方言の話されている地域が日本でどういった分布になっているのかを知ることは難しかった．また，口語は時の流れにより変化していくものであるのに加え，周りの環境によって各人の方言も変わっていくので，現在の方言の分布がどうなっているかをリアルタイムで知ることは難しい．本研究では，Twitter に投稿されたツイートの方言分布をジオタグを用いて可視化し明確化することで，自分の慣れ親しんだ都道府県などの方言が現在使われている地域を発見できるシステムを提案する．

本論文の構成は以下のとおりである．次章では方言やツイート分析・可視化に関する関連研究を紹介し，3 章では，ツイートの方言分析の方法について述べる．4 章では，本研究での方言分布の可視化方法について述べる．5 章では，提案した方言分布の可視化システムの実行例と評価実験の結果を示す．最後に，6 章でまとめと今後の課題について述べる．

2. 関連研究

地域特性についての関連研究として，坂本ら [1] は地域ごとの特性を算出し地域ごとに情報を推薦する研究を行った．Kamimura ら [2] は，Twitter に投稿されるジオタグ付き投稿に含まれる単語の空間的局所性をリアルタイムに解析することにより，最新の地域情報を表す単語を抽出した．Lee ら [3] は，Twitter へのジオタグ付き投稿を用い，特定の空間領域における投稿数やユーザ数の急激な変化に基づき，夏祭りや花火大会など局所的に人が集中するようなイベントの検出を行った．

SNS のテキストデータを用いて方言について研究した例は稀である．深谷ら [4] は方言を含む投稿から地域性に基づいた話題発見を目指した．彼らは文末の形態素について方言判定を行ったが，本研究では語尾ではなく，4 文字以上の特徴的な単語に着目している点が異なる．また，文末が多少変化しても意味を汲み取ることはできるが，全く知らない単語が出てきたとき意味を推測するのが難しい場合があるため，地域独自の単語に着目した．

近年，ソーシャルメディアサイトの発展とスマートフォンの普及によりジオタグ付きデータが急速に増加している．それに伴い，ジオタグを用いたマッピングについての研究が国内外で広く取り組まれている．志土地ら [5] はジオタグを用いて時空間をクラスタリングした．平久江ら [6] はジオタグを用いてマイナー観光地を推薦した．王ら [7] はジオタグの情報に加え，投稿位置の高さをツイート内容から算出した．酒井ら [8] は Twitter のジオタグを用いて注目を集めているツイートの時空間的な変遷を分析した．本研究では一定期間の投稿を対象にしており，投稿時間の変遷を考慮した方言抽出が必要である．Cheng ら [9] の研究では，ジオタグ付きツイートをもとに位置情報が付与されていないツイートの発信場所を確率的に推定するための手法などが提案されている．Watanabe ら [10] は，位置情報が付与されていないツイートの発信位置を推定し，ローカルイベント

(注1)：https://twitter.com/

表 1　大阪府の 4 文字以上の方言

都道府県	方言
大阪	あかんたれ, あほくさい, あらしょ, あんじょー, いかーしめへん, いちびる, いてこます, いんじゃんほい, うっとこ, えげつない, おえはん, おはよーおかえり, おもしゃい, おもろい, おもんない, おーきに, かめへん, かんてき, かんにん, がめつい, ぎょーさん, ぐっすり, けったいな, けなりー, こそばす, ごっかぶり, ごわへん, ごんたくれ, しょーもない, じゅんさいな, せわしない, せーだい, だんない, ちゃーる, でぼちん, どんならん, なんぎや, なんしか, なんでれでー, にくそい, にっちょ, のんもる, へてから, ぼちぼち, まいっぺん, めっちゃ, めばちこ, もむない, よってに, わらかす

をリアルタイムに発見する手法を提案した．ジオタグの付いたツイートは全体の 10 パーセントほどであるが，これらの研究を元により今後より多くのツイートを対象にすることを検討している．

大量のツイート情報を分析する研究が活発に行われており，リアルタイムに投稿される大量の Twitter データを閲覧者が容易に理解できるような可視化手法やインタフェースが必要となる．ジオタグ付きツイートの可視化についての関連研究も多く取り組まれている．Ghanem ら [11] と Magdy ら [12] の研究では，ジオタグ付きツイートを集約した情報を時刻ごとに地図上で可視化するインタラクティブなシステムが提案されている．Antoine ら [13] と Jatowt ら [14] の研究では，過去と未来のツイートの時空的関連性を分析・可視化し，過去・現在の的確な情報把握が可能なインタフェースが提案されている．

文字の可視化についての関連研究も多くなされている．松浦ら [15] はタグクラウドを用いて SNS に自らが投稿した古い投稿を想起するシステムを提案した．本研究ではタグクラウドを用いて表示する内容は方言辞書の内容の全てであり，アルゴリズムによって表示するものを決定していない．

3. ツイート分析に基づく方言抽出

3.1 方言抽出

多数の方言が収録されている goo 方言辞書 (注2) を用いて方言が話されているとされる都道府県名と方言を抽出し，方言辞書を作成した．方言辞書の一部である，大阪府の方言の例を表 1 に示す．

本研究では 3 文字以内の方言は別の意味を表す文字列と一致することが多いため除外した．たとえば,「仲間に入れる」という意味を持つ北海道の方言「かてる」は一般的に使われる「勝利することができる」という意味の「勝てる」のひらがな表記までも抽出してしまい適合率が落ちる．表 1 のように 4 文字以上の方言になると意味が唯一である単語が大多数であり，抽出された投稿内容では方言として使用されている場合がほとんどであった．4 文字以上の方言は約 8000 ツイートが該当した．

(注2)：https://dictionary.goo.ne.jp/dialect/

3.2 ツイート分析

本研究では，2016 年 10 月 1 日から 2016 年 10 月 31 日の 1ヶ月分の日本全国のジオタグ付きツイートを対象にしている．なお，ひらがなを 1 文字以上含むツイートで，bot のような大量の同じ内容のツイートを除いた，約 43000 件のツイートを対象にした．

ツイートには，データ格納時に自動で振り分けるユニーク ID，ユーザに対して自動で割り当てられるユーザ ID，ユーザのスクリーン名，@ をつけてメンションを送ることのできるユーザ名，ツイートの内容，発信した場所の緯度と経度，ユーザのプロフィール画像の URL，ツイートの発信時刻の情報が付随している．まず，MeCab を用いてツイートの内容を形態素解析した．ベースとなる MeCab 辞書には mecab-ipadic-NEologd (注3) を用いたが，作成した方言辞書の単語をユーザ辞書として加えて登録し，優先順位を上げることで方言が間違って形態素解析されないようにした．

4. 方言分布の可視化

本研究では，ツイートの方言情報を提示する一手法として，ツイートから抽出された方言を単純にランキング順に提示するなども考えられるが，複数の語を一覧的に表示する手法として代表的な手法であるタグクラウド [16] は，重要度によって大きさを変えた複数の語をアルファベット順に配列し効率的に表示する手法であり，方言を含むツイートを瞬時に把握することができるため，提案システムに採用した．

4.1 方言を用いたタグクラウド生成

本研究では，タグクラウドを生成するために，前章で作成した方言辞書を用いてツイートに含まれる頻出方言を抽出する．具体的には，*TF* 手法に基づきツイートを下記の式より，方言の出現頻度を算出し重みを付与する．

$$TF = \frac{\text{方言 } i \text{ の出現回数}}{\text{すべての方言の出現回数}}$$

これにより，発信されたツイートに含まれる方言を抽出し，重みが付与された方言を用いて各場所のツイートの方言を集約したタグクラウドを生成する．タグクラウドで方言の重みに応じてフォントサイズを設定し，表示する．

4.2 タグクラウドを用いた可視化

出現頻度の高い方言をタグクラウド (注4) として表示し，その場所に言及した都道府県別の方言を提示する．図 1 に可視化インタフェースを示す．図 1 はユーザが大阪の地図を閲覧している場合であり，タグクラウドの欄には大阪府の方言のタグクラウドが表示され，頻出した方言ほど大きく表示されている．重みに応じて色分けし，地図のピンと対応させる．色やタグの大きさは重みによって指定したが，タグの配置はランダムに決められている．

(注3)：https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md

(注4)：http://geoapi.heartrails.com/api.html

図 1　可視化システムのインタフェース

5. 評　　価

5.1　ツイート方言抽出の検証

ツイート方言抽出では，形態素解析したツイートのうち，3.1 節で作成した方言辞書の単語を含むもの 7645 件を抽出した．なお，店名に方言が含まれる場合を除いた．各単語に対する正解データの判定は，20 代の大学生 5 人が goo 方言辞書を参照しながら，主観的評価に基づきそれらの単語が方言の持つ意味合いとして使われているかどうかで評価を行う．5 人中 3 人以上が適しているとした単語が正解データとした．これにより判別したツイート全てにおける方言の抽出精度を適合率として以下の式によって算出した．

$$\text{適合率} = \frac{\text{手動で検出した方言ツイートの数}}{\text{システムによって検出された方言ツイートの総数}}$$

3.1 節で作成した方言辞書の単語を含むツイートを抽出した結果，適合率は 51.2% であった．抽出した方言を含むツイートの例を表 2 に示す．5 文字以上の方言ツイート 241 件のみを抽出した結果，適合率は 75.5% であった．このことから，少ない文字の方言を含むほど形態素解析の精度が落ちてしまうことがわかった．誤って判別したものの例の 1 つに,「だんだん」という方言がある.「だんだん」は広島県の方言で「ありがとう」，佐賀県では「いつも」という意味をもつ方言であるのに対し，ツイート中から検出されたものは「徐々に」の意味合いで使われていたため,「だんだん」を含むツイートを間違って抽出した．

件数が多いがほとんどが適合していないものに「あります」もあった.「あります」は山口県の方言で「〜ございます」という意味だが,「〜存在します」という意味で使われているものがほとんどで，7645 件中 1539 件の間違った検出をしたため，大幅に適合率を落とす原因の 1 つとなった．特に多く使われていた方言は大阪府で「とても」を意味する「めっちゃ」であった．最も多くの種類の方言を抽出することができた都道府県は大阪府であった．また，最北端，最南端である北海道，沖縄も多数の方言を抽出することができた．さらに,「冷たい」を意味する「しゃっこい」は北海道と青森県の,「山盛り」を意味する「てんこもり」は北海道と滋賀の方言であることから，都道府県境を超えて使われている方言や，離れた都道府県で共通の方言があることがわかった．同音異義語も数多く検出された．大阪府の方言が最も多く 2751 件で，内訳は表 3 のようになった..

表 2　方言を含むツイートの例

方言を含んだツイート	方言	都道府県名
八ヶ岳中央農業実践大学校の直売所に来ました。紅葉旅行なんだけど、天気が悪いので、のっけから予定が違う……^^; @八ヶ岳中央農業実践大学校 https://t.co/WXvouohcbU	のっけから	岐阜県
えげつないアメリカンムキムキと太鼓の達人@セガ横浜中華街 https://t.co/Ydb7Vs3hAp	えげつない	大阪府
焼肉定食くださいな〜(@旬の肴安) https://t.co/rpF6cAdOz3 https://t.co/1dgVNencot	くださいな	東京都
昨日のタスクバー楽しみました！ タスクで DJ する時は、うまく 言えないけど、「楽しい」と 「気持ちいい」が混ざってて、 いつもほっこりする #タスクバー #dj #djlife… https://t.co/wTZ2JOUprF	ほっこりする	京都府
宿ボロいけど値段安いからしゃーない (@水戸第一ホテル本館 in 水戸市茨城県) https://t.co/fTtcbOKCBx	しゃーない	大阪府

表 3　大阪府の 4 文字以上の方言

方言	件数
めっちゃ	2562 件
おもろい	100 件
おおきに	30 件
えげつない	14 件
ぼちぼち	13 件
しょーもない	11 件
おもんない	7 件
なんしか	4 件
まいどー	3 件
がめつい	3 件
あんじょー	1 件
めばちこ	1 件
へてから	1 件
他	0 件

5.2　システムの実行例

可視化システムのインタフェースは図 1 に示した通りである．OpenStreetMap(注5)はフリーの地理情報データを作成することを目的としたプロジェクトである．本研究では，OpenStreetMap に Leaflet を用いて，方言を含むツイートが投稿された位置を地図上で表示した．4.2 節で作成したタグクラウドを地図の右に表示した．下には方言に一致するツイートをユーザ名と投稿時間と一緒に表示した．以上の機能により，ユーザはタグクラウドを見ることで，その地域によく使われる方言を確認でき，関連のあるツイートを見ることで，その地域に関する方言ツイートを発見でき，効率的かつ効果的な情報提供につ

(注5)：https://openstreetmap.jp/

(a)

(b)

図 2　システムの実行例

ながる．

システムを実行する一連の流れは以下のとおりである．

（1）　ユーザが都道府県名を選択する．

（2）　その都道府県の方言一覧，投稿された場所のピン，投稿内容と投稿日時が表示される．

（3）　タグクラウドから方言をユーザがクリックする．

（4）　その方言に該当するピンとツイートが表示される．

図 2 に実行した際のシステムの様子を示した．左の画面は，左上のプルダウンで大阪府を選択し，大阪府の全ての方言ツイートのピンとツイート一覧を表示している時の様子である．この画面のタグクラウドのうち,「えげつない」をクリックした時の様子が右の画面である．この時，タグクラウドとピンが黄緑色である「えげつない」のみのツイートとピンが表示されるようになる．青い丸で囲まれていることから分かるように，ツイート一覧は「えげつない」を含むツイートのみが表示されていることがわかる．ピンは黄緑色のもののみになっており，これは全て「えげつない」を含んだツイートのジオタグから取得した投稿位置を表している．大阪府の方言であるが大阪府からの投稿はなく，関東圏でより多く投稿されていることがわかる．

5.3　可視化システム有用性の評価

本節では，4 章で実装した可視化システムを 20 代男性 4 名と 30 代女性 1 名の被験者に見てもらい，下記の設問項目についてアンケートを実施した．主に研究目的である新たな方言分布の発見可能性について質問した．アンケートの内容は以下の通りである．

$Q1$　興味のある都道府県が発見できた

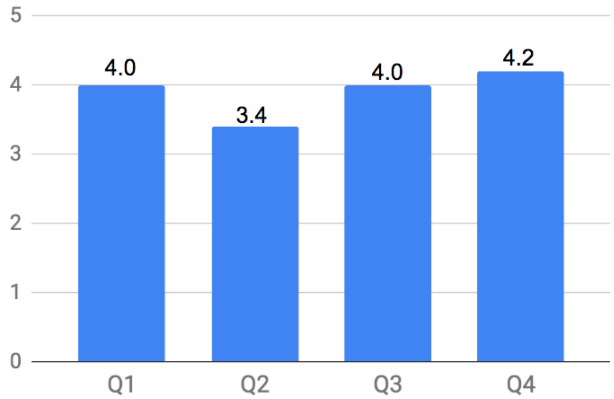

図 3　Q1 から Q4 の 5 段階評価の評価結値

$Q2$　その都道府県に行ってみたいと思った

$Q3$　現地の人と話してみたいと思った

$Q4$　1 つの都道府県に対し，よく使われている方言がどんなものかわかった

$Q5$　UI について使いにくいと感じた箇所があれば書いてください

$Q6$　ほしい機能があれば書いてください

これらアンケートは，$Q1$ から $Q4$ は 5 段階評価（1. 全くそう思わない，2. そう思わない，3. どちらともいえない，4. そう思う，5. とてもそう思う）で評価してもらい，その他の質問 $Q5$ と $Q6$ は詳しくわかるように自由に記述できるコメント欄を設けるようにし，それらアンケート結果を用いて，提案した可視化システムの有用性を評価する．5 段階評価による $Q1$ から $Q4$ の評価の平均値を図 3 に示す．全体的に良い評価であったが $Q2$ の「行ってみたいと思った」のみ点数が低かった．この原因としては，地図を見ただけでは周辺情報がわかりにくく，行く目的が明確に定まらないためだと考えられる．本研究の目的である $Q1$ の「興味のある都道府県が発見できたか」について平均値 4 の評価を得られたため，研究目的を概ね達成したといえる．$Q4$ については最も高い平均値 4.2 の評価を得られた．よく使われている方言についてはタグクラウドによって文字の大きさで表現されている他，地図上のピンの数を見ることで理解できるため，高い評価を得られた．

$Q5$ の記述評価について，以下のような意見を得られた．

- 送信ボタンが小さい
- 地図の表示部分がもう少し大きいと良い
- タグクラウドの配置が出現順だと良い
- 出現頻度が曖昧
- 方言の意味がわからない

送信ボタンや地図のサイズについての指摘があったため，UI について再検討する必要がある．また，本システムでは方言の意味を表示する機能がないため，ユーザは使いにくいと感じたことがわかった．タグクラウドによって表示した単語は，ランダムに配置されるようになっているが，ソートして見やすくしたいという意見があったため，今後，タグクラウドでの単語配置の決定方法も検討する必要がある．

$Q6$ の記述評価について，以下のような意見を得られた．

- 方言を使用する者同士の会話を見られる機能
- 一般的な地図サービスにある地名検索機能
- 方言の意味説明機能
- 五十音順，自分の現在地から近い順などの並べ替え機能

$Q5$ でも方言の並び替えについての指摘と方言の意味がわからないとの指摘があったため，優先的に追加すべき機能として，方言の並べ替えができる機能，方言の意味を説明する機能が必要である．

6. おわりに

本研究では，ジオタグ付きツイートから方言を抽出し，ジオタグと結びつけた方言ツイートの投稿場所を可視化することによって，自分の慣れ親しんだ都道府県や興味のある都道府県に対し，その場所の方言が使われている地域を発見できる可視化システムを提案した．また，評価実験ではシステムによって現在の方言の利用状況を理解し，各都道府県に興味を持ってもらえるかを検証した．その結果，ツイートから方言を抽出できることを確認できた．しかし，4 文字以上の方言が適切に抽出されている割合は 51.2% にとどまった．可視化システム有用性の評価では概ね高い評価を得られた．提案した可視化システムが各都道府県に興味を持ってもらうことに貢献できることを確認した．

今後の課題として，ツイートから適切に方言を抽出する方法を考える必要がある．たとえば，MeCab で使用する方言に，品詞などを登録し，適切なアルゴリズムによって重み付けをすることにより，精度が上がると考えられる．また，本研究では 4 文字以上の方言を対象にしたが，4 文字以内で表現される方言も方言と正しく見抜くアルゴリズムについっての検討が必要である．さらに，日々言葉は変化しており，死語になるものや新たに誕生する方言もでてくると考えられる．ジオタグ付きツイートから新たな方言を発見し，動的に方言辞書を作成することを検討する必要がある．

謝　辞

本研究の一部は，JSPS 科研費 15K00162，17K12686 の助成を受けたものである．ここに記して謝意を表す．

文　献

[1] 坂本 宏祐, Lim Jeongwoo, 新田 直子, 中村 和晃, 馬場口 登: マイクロブログを用いたリアルタイム地域情報の推薦, DEIM Forum 2018, D1-2, 2018.

[2] T. Kamimura, N. Nitta, K. Nakmura, and N. Babaguchi, “On-line Geospatial Term Extraction from Streaming Geotagged Tweets,” Proc. International Conference on Multimedia Big Data, pp.322329, 2017.

[3] R. Lee, S. Wakamiya, and K. Sumiya, “Discovery of Unusual Regional Social Activities Using Geo-tagged Microblogs,” World Wide Web, Vol.14, No.4, pp.321349, 2011.

[4] 深谷 大樹, 有馬 直也, 河合由起子, 湯本 高行: ジオタグツイートの文体分析に基づく話題抽出と可視化, DEIM Forum 2017, P2-1, 2017.

[5] 志土地由香, 井手一郎, 高橋友和, 村瀬洋: 時空間的な投稿数を考慮した密度に基づく適応的な時空間クラスタリング手法, DEIM Forum 2015, A2-4,2016.

[6] 平久江知樹，早川智一，疋田輝雄: マイクロブログにおけるジオタグのクラスタリングを用いたマイナー観光地抽出手法の改良, DEIM Forum 2018, H1-5,2018.

[7] 王元元，安井豪基，丸山直樹，河合由紀子，秋山豊和，角谷和俊: 複合施設におけるツイートの時空間分析に基づくタグクラウドを用いた可視化システム, 2018.

[8] 酒井 達弘, 田村慶一, 北上 始: 時空間的な投稿数を考慮した密度に基づく適応的な時空間クラスタリング手法, DEIM Forum 2016, A2-4, 2016.

[9] Cheng, Z., Caverlee, J., and Lee, K.: You Are Where YouTweet: A Content-based Approach to Geo-locating Twitter Users,inProc. of 19th ACM International Conference on Information andKnowledge Management (CIKM 2010), pp. 759768 (2010)

[10] Watanabe, K., Ochi, M., Okabe, M., and Onai, R.: Jasmine: A Real-time Local-event Detection System Based on Geolo-cation Information Propagated to Microblogs, inProc. of 20th ACMInternational Conference on Information and Knowledge Manage-ment (CIKM 2011), pp. 25412544 (2011)

[11] Ghanem, T. M., Magdy, A., Musleh, M., Ghani, S., andMokbel, M. F.: VisCAT: Spatio-temporal Visualization and Aggrega-tion of Categorical Attributes in Twitter Data, inProc. of 22nd ACMSIGSPATIAL International Conference on Advances in GeographicInformation Systems (SIGSPATIAL 2014), pp. 537540 (2014)

[12] Magdy, A., Alarabi, L., Al-Harthi, S., Musleh, M.,Ghanem, T. M., Ghani, S., and Mokbel, M. F.: Taghreed: A Systemfor Querying, Analyzing, and Visualizing Geotagged Microblogs, inProc. of 22nd ACM SIGSPATIAL International Conference on Ad-vances in Geographic Information Systems (SIGSPATIAL 2014), pp.163172 (2014)

[13] Antoine, E., Jatowt, A., Wakamiya, S., Kawai, Y., andAkiyama, T.: Portraying Collective Spatial Attention in Twitter, inProc. of 21st ACM SIGKDD Conference on Knowledge Discoveryand Data Mining (KDD 2015), pp. 3948 (2015)

[14] Jatowt, A., Antoine, E., Kawai, Y., and Akiyama, T.:Mapping Temporal Horizons, Analysis of Collective Future and Pastrelated Attention in Microblogging, inProc. of 24th InternationalConference on World Wide Web (WWW 2015), pp. 484494 (2015)

[15] 松浦翔, 松本若樹, 村上晴美: タグクラウドを用いた記憶の想起支援, 情報処理学会第 77 回全国大会, 3N-08,2018.

[16] Martin J. Halvey, Mark T. Keane: An Assessment of Tag Presentation Techniques, Proc. of the 16th international conference on World Wide Web (WWW 2007), pp. 1313-1314,