IDMによるクチコミマイニング

松村真宏[i]
大阪大学大学院経済学研究科

1. はじめに

クチコミは何らかのキッカケがあって話題に火が付き、それが人から人へ話題が広まることによって発生する。本稿では、影響伝播モデルIDMを用いてクチコミのメカニズムをモデル化することにより、クチコミが発生してからブレイクするまでの一連の連鎖を抽出する手法について述べる。

2. IDMによるクチコミの定量化

IDMは構造化されたメッセージ（スレッド）を対象として、語、メッセージ、投稿者の影響力を算出するアルゴリズムである[松村 02]。IDMでは、メッセージM_1の影響力Inf_Mを次式で定義する。

$$Inf_{M_1} = |w_1 \cap w_2| + |w_1 \cap w_2 \cap w_3| + \cdots + |w_1 \cap w_2 \cap \cdots \cap w_m| \quad (1)$$

ここで$M_1, M_2, M_3, \ldots, M_m$は$M_1$を起点とするスレッドに含まれるメッセージ（時間順にソートしたもの）、w_iはMiに含まれる語の集合（主に主義語）を表す。

このとき、メッセージM_iに含まれる語tの影響力$Inf_{M,t}$を次式で定義する。

$$Inf_{M,t} = \frac{Inf_M}{|w_i|} \quad (2)$$

また、投稿者Pの影響力Inf_Pは次式で定義する。

$$Inf_P = \sum_{m \in \eta_P} Inf_m \quad (3)$$

ここでη_Pは投稿者Pの投稿したメッセージ集合とする。

式(1)(2)(3)で求めている影響力はいずれも語の伝播量に基づいたものであるため、クチコミ力に置き換えて解釈することも可能である。

3. クチコミの連鎖

クチコミでは、同じ話題がずっと語り継がれて広まるのではなく、それぞれの話題が次の話題のトリガーになるので、話題はどんどん遷り変わっていく。これをIDMのフレームワークから見ると、メッセージに伝播してくる語と、伝播していく語の関係に他ならない。

そこで、この関係、つまり、メッセージごとに伝播してくる語と伝播していく語のペアを求め、その関係を可視化すると図1のような有向グラフが得られる[1]。赤ノードは影響力の高い語、青ノードはそれ以外の語、リンクの向きは影響（クチコミ）の連鎖関係（矢印元の語が矢印先の語のトリガーとなる）と表している。図より「たこ焼き」「寝屋川」「割高」「高い」を中心として、「評価」「高い」「名前」「ソース」「ミナミ」にクチコミが波及していることが分かる。また、もっと細かく見ていけば、「寝屋川」の「たこ焼き」屋の「名前」を「教えて」もらったら値段が「高い」かったので「評価」が「悪い」という連鎖を把握することができる。

---

[1] 2 ちゃんねるの「大阪で最強のたこ焼き屋　その4」スレッドを解析した。

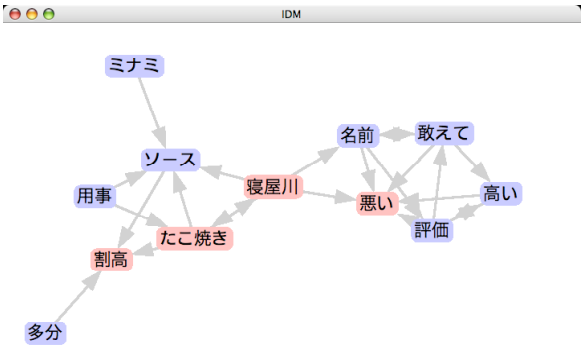

図 1. クチコミによる話題連鎖

4. その他の機能

同じ話題についても男性と女性では見ているポイントが異なっていることはよくある。例えば、男性は携帯電話の機能やカメラの解像度にこだわるかもしれないが、女性はデザインの可愛らしさや重さに注目しているかもしれない。したがって、クチコミを媒介する発言者の性別は、クチコミを理解する上で非常に重要である。IDM では機械学習を用いたメッセージごとに書き手の性別判定モジュール[Kobayashi 07]を内蔵している。

また、肯定的・否定的・中立的なメッセージがあり、それぞれの観点から分析することは対象の良い点、悪い点を把握する上で非常に重要である。IDM ではメッセージの極性判定モジュールも内蔵している。したがって、IDM では状況に応じて性別判定・極性判定を組み合わせてクチコミマイニングをすることができる。

5. 関連研究との比較

特徴語を求める有名な方法に TFIDF 法がある。TFIDF 法では語の頻度に比例する尺度なので、例えばある語の頻度が 10 倍になれば、その出現箇所がランダムでも TFIDF 値は 10 倍になる。一方、IDM では連鎖する語だけを用いて影響力を算出するので、ランダムに語を挿入しても影響力は大局的にはそれほど変化しない。また、短い間隔で頻出する語はバーストとして定式化できる[Kleinberg 02]が、IDM ではメッセージの構造を考慮している点が大きな違いである。

また、語の共起関係を取り出すアプローチはいくつも存在するが、クチコミの連鎖関係を有向グラフとして取り出せることも IDM ならではの特徴である。

また、本稿ではスペースの関係上紹介できなかったが、投稿者間の関係を表すクチコミネットワークも同様に抽出できるので、影響力のみならず構造的な特徴量を用いてインフルエンサーやインフルエンシーを見つけることも可能である。

また、同じ語でも投稿者ごとにその影響力は異なることも IDM の特徴であり、投稿者のプロファイルなどに利用できる。

6. まとめ

本稿では IDM を利用したクチコミマイニングについて概観した。今後も引き続き IDM の機能拡張に取り組みつつ、実事例にも応用していきたいと考えている。

参考文献

[松村 02] 松村真宏、大澤幸生、石塚満：テキストによるコミュニケーションにおける影響の普及モデル、人工知能学会論文誌 Vo. 17, No. 3, pp. 259—267 (2002)

[Kleinberg 02] Kleinberg, J.: Bursty and hierarchical structure in streams, Proc. 8th ACM SIGKDD (2002)

[Kobayashi 07] Daisuke Kobayashi, Naohiro Matsumura: Automatic Gender Estimation of Bloggers' Gender, Proc. 1st ICWSM, pp. 279—280 (2007)

[i] 連絡先: matumura@econ.osaka-u.ac.jp