

LOD × 既存予測アルゴリズム

越川兼地^{1*} 1

¹ 電気通信大学 大学院情報システム学研究科

¹ University of Electro-Communications Graduate School of Information Systems

Abstract: Linked Open Data Challenge Japan 2012 のアイデア部門のエントリー作品について説明する。本アイデアは、既存の予測アルゴリズムと LOD とのマッシュアップ案を幾つか提示し、マッシュアップにより生まれるであろう新たな価値と必要な LOD について考察するものである。

1 本稿について

本稿は Linked Open Data Challenge Japan 2012¹ のアイデア部門へのエントリー作品の説明文書である。本アイデアはイベント:International Asian LOD Challenge Day²でのグループからの投稿であり、グループの構成員は越川 兼地³、森田 武史⁴、西村 悟史⁵、長野 伸一⁶から構成されている。(敬称略)

参考までにイベント当日に本アイデアのイメージをまとめたものを図 1 に示す。

2 本アイデアについて

LOD の特徴である分野を超えたデータ統合が容易であるという点に着目し、既存の予測アルゴリズムの精度に寄与するであろう情報を LOD から補う形でのマッシュアップ案を複数提示する。具体的には、既存アルゴリズムとマッシュアップすべき LOD 情報の組み合わせを提示し、新たに生まれるであろう価値及び必要な LOD について考察する。

これらのマッシュアップ案が実現することができれば、日々の行動の意思決定時の材料が増え、確かな行動選択につながり、我々の日々の暮らしがより豊かになると信じている。次の 3 節にいくつかの既存予測アルゴリズムと LOD のマッシュアップ案について述べる。

3 既存予測アルゴリズムと LOD のマッシュアップ

3.1 過去の事象間の因果関係から起こりうる事象の予測

Radinsky ら [1, 2] は過去のニュース記事からを機械学習・データマイニング・オントロジーの技術を用いて、自然言語から事象間の因果関係に着目したネットワークを構築し、150 年間の過去のニュース記事を対象に学習及び評価を行ない、人間よりも高い精度でニュースを予測できたことを示した研究である。この手法を実用化するためには、随時更新されているソーシャルメディア・マスメディアの情報を入力として実際に事が起こる前に予測システムが対象者に予測した結果を伝えなくてはならない。そうでないと、たとえ予測結果の精度が正しかろうと、起こりうる事象に備えることが出来ずに予測の意味がなくなってしまうので。この手法の実用化のために何よりも最新のニュース記事を確保する必要がある。またニュース記事だけでなく、ソーシャルメディアから得られる情報についても同様な処理を行うことで、ニュース記事がカバーしていない領域に対しての事象予測が大いに期待できる。

3.2 病気の流行予測

Aramaki らはインフルエンザの流行度を Twitter⁷ の情報を元に把握するモデルを提案している [3]。この研究の主なアイデアはインフルエンザに関連する語 ("flu" など) が含まれるツイートに対し、患者が発信したものかどうかを SVM で分類する手法 [4] を使っていることである。この手法に対し、患者の所在地、症状が現れた時間などの病気に関する詳しい情報を得ることができれば、地域的な感染経路の把握および時系列での深い分析が期待でき、効果的な感染の予防につなげることになる。

*連絡先：電気通信大学 大学院情報システム学研究科 社会知能情報学専攻 大須賀・田原研究室

E-mail: k-koshikawa@ohsuga.is.uec.ac.jp

¹<http://lod.sfc.keio.ac.jp/challenge2012/index.html>

²<http://www.ei.sanken.osaka-u.ac.jp/jist2012/IALOD-ChallengeDay.htm>

³所属：電気通信大学 情報システム学研究科 社会知能情報学専攻 大須賀・田原研究室

⁴所属：青山学院大学 社会情報学部

⁵所属：大阪大学 産業科学研究所 旧溝口研究室

⁶所属：東芝 研究開発センター 知識メティアラボラトリー

⁷<https://twitter.com/>

表 1: 各予測アルゴリズムの改良に必要な情報 (情報源・情報鮮度・特に必要な LOD)

	因果関係	病気	交通情報	システムトレード	被害情報	需要	犯罪
主な情報源	マスメディア ソーシャルメディア	○ △	- ○	○ ○	○ ○	○ ○	○ ○
情報鮮度(3段階)	そこそこ重要	重要	とても重要	とても重要	とても重要	重要	そこそこ重要
特に必要な LOD	事象情報	病気情報 患者の状況	交通機関 ユーザの行動情報	企業/銘柄 事象情報 消費者の声	災害情報 地理情報 被災者の状況	商品情報 消費者の声	犯罪情報 地理情報 犯罪者目撃情報

とができると考えている。3.1 節で紹介した予測アルゴリズムの場合と同じく、情報の鮮度が良いほど早い段階で予防に対処できるため、鮮度の良い解析に適した情報資源が必要となっている。

3.3 特定場所の交通情報を把握した上での最適な経路推薦

Nguyen ら [5] は東日本大震災時における首都圏の交通網が完全に停止したパニック時を想定し、ソーシャルメディアから特定の場所の状況を把握することで適切な行動（経路）推薦を行えることを示した。この手法は震災に限らずに日常に頻繁に起こりうる。電車遅延における最適な代替経路の探索などにも適応できると考えおり、その実現のためには、各種交通機関（鉄道、バス、道路状況など）の最新ステータスの情報とソーシャルメディアなどのボトムアップ型の情報を組み合わせることで実現に近づけると考えている。

3.4 システムトレード分野の予測アルゴリズム

システムトレードの分野においても効率的な資産運用を目的とした様々な予測アルゴリズムが提案されている [6]。その中でも Bollen ら [7] はソーシャルメディア（Twitter）上から読み取れるユーザの感情情報を着目し、ある一日の大多数のユーザの感情傾向が“平穏”であると判断された場合、3日後の株価（ダウ平均株価）と正の相関があることを発見し、その特徴を利用した特徴モデルを提案している。ニュース記事を利用した手法 [8, 9]、インターネット株式掲示板 [10] などの情報を手がかりにした世の中の事象を表現しているテキスト情報から予測モデルに落としこむ手法が近年のトレンドとなっている [11]。

前節で挙げたアルゴリズムと同様に、マスメディアとソーシャルメディア双方の実世界に関する事象情報が株価予測に利用されていたが、精度向上のためには情報の鮮度とともに、株価予測の分野ならではの整備された情報源を LOD 化して、既存 LOD との組み合わせた情報を用いることでさらなる精度向上が期待できると考えている。例えば、企業情報 LOD からある企

業の各営業所（営業所存在した場合）に関する情報と地理情報 LOD を紐付け、地理情報 LOD から営業所の緯度経度が取得できる。地域別のインフルエンザなどの疫病の流行度を表す LOD（存在する場合）を用いており、営業所の付近が疫病の該当地域かどうかが判断でき、疫病が流行っている地区では営業がまともにできないことが想像でき、株価に響く要因であると考えることができる。このように情報を組み合わせて、より着いた先の情報が株価と相関があるものかどうかを調べ、重要な因子を算出することで株価予測アルゴリズムに寄与できると考えている。

3.5 その他分野の予測手法

その他分野として、“災害の被害情報予測”，“需要予測”，“犯罪予測”などの予測手法に関しても LOD 情報が寄与できると考えている。災害の被害情報予測については、過去の災害規模と被害情報を LOD として整備することで、より繊細な被害情報予測が可能になる。また早めに被害想定地域の人々に伝えることで、2次災害の被害を抑えることができると考えている。

需要予測においては、株価予測アルゴリズムと同様に商品の需要を決定する要素は多岐にわたるため、関連する情報を LOD として整備することで、重要な要素を解析する工程が容易になり小売業の店舗における最適な商品発注量の算出精度向上への寄与、商品開発においての的確なマーケティング戦略案を練ることにつながると考えている。

最後に犯罪予測に関して、アメリカにおいては、過去の犯罪データに基いて次の犯罪が起きる可能性の高い地域と時間帯を計算機で予測するシステムを運用している。[12] 今後日本においても徐々にアメリカのような犯罪予測に関する仕組みが整備され、導入されていくと予想される。この犯罪を予測する精度においても他の予測アルゴリズムと同様に、他分野の LOD を組み合わせることで精度向上が期待でき、犯人逮捕率向上とともに事前に想定地域に対して警告を出すことで、被害者を減らすことになるとを考えている。

4 おわりに

本稿では、既存の予測アルゴリズムと LOD とのマッシュアップ案をいくつか提示し、マッシュアップにより生まれるであろう新たな価値と必要な LOD について考察した。表 1 に前節で紹介したアルゴリズムの精度向上に必要な情報源・情報鮮度・特に必要な LOD についてまとめたものを示す。

この表及び、前節の考察から情報鮮度の良い情報が未来予測にはかなり重要な情報であることがわかる。また“事象情報”、“患者の状況”、“ユーザの行動情報”、“消費者の声”、“被災者の声”、“犯罪者目撃情報”といったような実世界の出来事を表した情報が未来予測もしくは精度向上には必要不可欠であることがわかる。しかし、このような世の中の事象情報を構造化した状態で表したデータは世に公開されていないのが現状であり、このようなデータを世に整備するのが早急の課題であることが本考察から導けた。

また 3 節では、各予測アルゴリズムに対し有用だと思われる LOD を挙げたが、具体的にどのようにアルゴリズムに用いるのかについては、考慮しきれていない。その具体的な方法論を考え実際にアルゴリズムをサービスとして実現することが求められる。

謝辞

本稿を書くにあたり、International LOD Challenge Day でのグループメンバーとの刺激的な議論及び有用なコメントを頂きました森田 武史先生、西村 悟史様、長野 伸一様に心より感謝いたします。またこの素晴らしいイベントの運営をされている LOD Challenge Community の皆様、イベントで関わったすべての参加者の皆様に感謝いたします。

参考文献

- [1] Kira Radinsky, Sagine Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *WWW*, pp. 909–918, 2012.
- [2] Kira Radinsky and Eric Horvitz. Mining the web to predict future events. In *WSDM*, pp. xxx–xxx, Feb 2013. to appear.
- [3] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Influenza patients are invisible in the web: Traditional model still improves the state of the art web based influenza surveillance. 2012.
- [4] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *EMNLP*, pp. 1568–1576, 2011.
- [5] The-Minh Nguyen, Takahiro Kawamura, Yasuyuki Tahara, and Akihiko Ohsuga. Building a time series action network for earthquake disaster. In *ICAART (1)'12*, pp. 100–108, 2012.
- [6] ファイナンスにおける人工知能応用研究会 (SIG-FIN) Wiki 関連論文サーバイ資産運用計画. <http://goo.gl/owY6q>.
- [7] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *J. Comput. Science*, Vol. 2, No. 1, pp. 1–8, 2011.
- [8] M.-A. Mittermayer and G. F. Knolmayer. Text mining systems for market response to news. *A Survey, Technical report, University of Bern*, 2006.
- [9] R. Schumaker and H. Chen. A discrete stock prediction engine based on financial news. *IEEE Computer*, Vol. 43, No. 1, pp. 51–56, 2010.
- [10] 丸山健, 梅原英一, 謝訪博彦, 太田敏澄. インターネット株式掲示板の投稿内容と株式市場の関係. *証券アナリストジャーナル*, Vol. 46, 11・12, pp. 110–127, 2008.
- [11] 和泉潔, 後藤卓, 松井藤五郎. テキスト分析による金融取引の実評価. *人工知能学会論文誌*, Vol. 26, No. 2, pp. 313–317, 2011.
- [12] FAST: Future Attribute Screening Technology. http://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_st_fast.pdf.

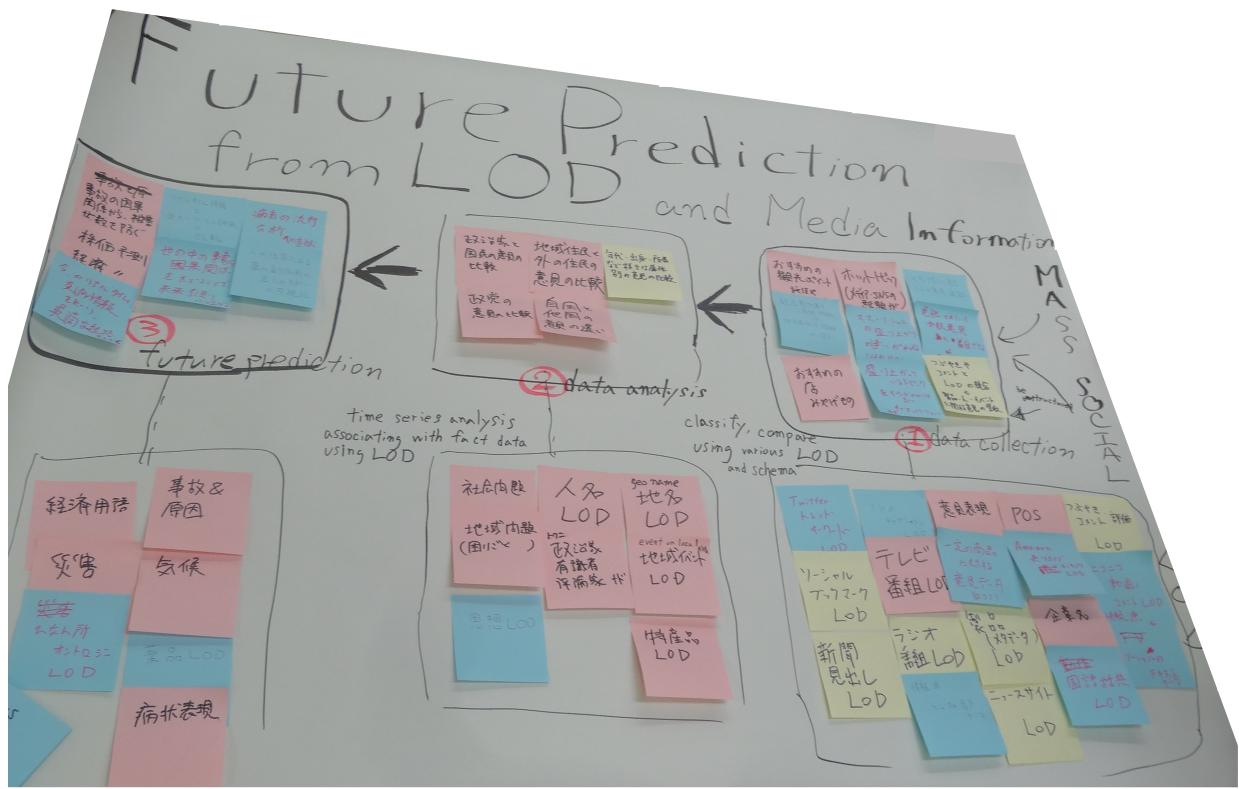


図 1: イベント:「International LOD Challenge Day」グループメンバーで作成した「予測アルゴリズム × LOD のマッシュアップ案」イメージ