様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ９ 日現在

機関番号：１４６０３
研究種目：基盤研究(B)（一般）
研究期間：2012～2015
課題番号：２４３０００５７

研究課題名（和文）高次元・構造化データに適したリンク解析的類似度尺度の研究

研究課題名（英文）Studies on link analytic similarity measures for high-dimensional/structured data

研究代表者
新保　仁（Shimbo, Masashi）

奈良先端科学技術大学院大学・情報科学研究科・准教授

研究者番号：９０３１１５８９

交付決定額（研究期間全体）：（直接経費）　　13,600,000円

研究成果の概要（和文）：近年リンク解析・機械学習分野で得られた知見を基に，自然言語をはじめとする高次元・構造化データを対象とした，検索・事例の細分類などに適した類似度尺度の特徴付けと開発を行った．特にハブ事例（数多くの他の事例の近傍に出現する事例；高次元データで出現しやすいことが知られている）が類似度尺度に与える影響に着目して調査・分析するとともに，その悪影響を軽減する方策を提案した．

研究成果の概要（英文）：Building on the latest findings in link analysis and machine learning, this research project developed and characterized various similarity measures for high-dimensional or structured data. In particular, our main focus was to investigate the influence of hubs, which can be observed both in vector space and on graph data. We proposed several techniques to reduce the emergence of hubs. The effectiveness of these techniques was evaluated on natural language processing tasks, in which the data is known to be extremely high-dimensional.

研究分野：知能情報学

キーワード：リンク解析　高次元データ　近傍検索　ハブ

1　研究開始当初の背景

最近の機械学習分野における発見に,「高次元データセットには, 他の多数の事例の近傍に位置するハブ (hub) と呼ばれる事例が出現しやすい」というものがある [Radovanović et al. 2010]. データセット中のハブは, 近傍検索・k 近傍分類・リンク解析・グラフを用いた半教師あり分類, およびその前処理である疎グラフ構築といった様々なタスクに悪影響を及ぼす可能性がある.

一方, リンク解析においては, グラフ上の全ての部分構造 (経路など) をもとに節点間類似度を計測する手法が提案されているが大規模なグラフにおいては, 単純な次数のみに依存する尺度 (一種の重要度尺度) になってしまう欠点が知られている.

2　研究の目的

リンク解析・機械学習分野で得られた知見を基に, 自然言語をはじめとする高次元・構造化データを対象とし, 検索・事例の細分類などに適した類似度尺度の特徴付けと開発を目的とする. 特にハブ事例が各種類似度尺度・タスクに及ぼす影響について調査・分析し, その影響を軽減する方策を提案する.

3　研究の方法

典型的な高次元データである自然言語データを用い, 各種のアプリケーションタスクの実験を通じて取り組む課題を発見する. ただし, 一般性のある知見を得ることが目的であるため, 個々のアプリケーションタスクに特有の問題ではなく, 多くの検索・分類問題に共通な問題に限定して分析し・改善方法を考案する.

4　研究成果

(1)　グラフ上の部分構造の総和に基づくリンク解析尺度

グラフ上の経路, あるいは森 (forest) が個々のグラフ節点・辺をどのような頻度で含むか, に基いて節点や辺の重要度・関連度・周辺グラフ密度を測定する尺度を提案した. これらの手法では, 部分構造の大きさ (経路長, 森の重み, など) をどの程度考慮するか, をパラメターを通じて制御可能であり, この特徴により, スペクトラムの異なる多様な尺度が構成可能となる. また, パラメターを適切に調整することによって,「研究の背景」で述べた大規模グラフにおける既存の類似度尺度の欠点を (この尺度の利点をある程度維持しつつ) 回避することが可能である. 得られた成果は学術誌 IEEE TPAMI, Physica A で採択され, より詳細な説明を, 本研究の一環として執筆したリンク解析技術の解説書に収めた. なお, 本研究はベルギー・ルーバンカトリック大 (Université catholique de Louvain) との共同研究として行った.

(2)　中心化によるハブの抑制

ベクトル事例間の類似度尺度として内積を用いる場合に, 中心化を行うことでハブが削減されることを理論的・実験的に示した. この成果については, 自然言語処理に関する国際会議である EMNLP にて発表した. 自然言語処理においては, 事例間類似度としてしばしばコサイン (内積の一種とみなせる) が用いられるが, 中心化は一般的には用いられることが皆無であった.

(3)　ハブの抑制の対訳抽出問題への応用

対訳抽出タスクは, 異なる言語の単語の対応を取る問題である. 最も単純な手法は, 単語を, 対訳であることが既知の複数の単語 (ベクトル要素) との共起頻度を (単元語コーパス上で) 計算して作った「共起ベクトル」の類似度 (コサイン

など) を用いて対訳となっている単語対を見つける. 我々は, この共起ベクトルに基づく手法に対してハブ抑制を行うことで (上述の中心化, および, 相互近接性 (mutual proximity) [Schnitzer et al. 2012] を用いた), より高性能であるとされてきたラベル伝搬法に基づく手法を凌駕する対訳抽出精度が得られた.

(4) 言語の構成性を捉えるための単語表現・類似度の学習

近年, 深層学習が注目される中, ベクトル空間上での単語表現を学習する手法に関する研究も盛んに行われている. なかでも, 特に複数の単語がまとまってフレーズを構成する際に, 意味が変化する現象をベクトル演算で捉える「構成性 (compositionality) のモデル化」は重要な課題である. 我々は, フレーズを構成する単語が相互に影響して意味が精製される, あるいは変化する過程をベクトル空間上の射影として実装するとともに, 単語表現の学習法と類似度を測るためのカーネルのパラメタ自動調節を組み合わせた手法を提案した. 成果は EMNLP, AAAI という自然言語処理, 人工知能に関する国際学会に採択された.

引用文献

[Radovanović 2010] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, Vol. 11, pp. 2487–2531, 2010.

[Schnitzer 2012] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research*, Vol. 13, pp. 2871–2902, 2012.

5 主な発表論文等

[雑誌論文] (計 6 件)

(1) 重藤優太郎, 鈴木 郁美, 原 一夫, 新保 仁, 松本裕治.
ハブの抑制によるコンパラブルコーパスからの対訳抽出精度の改善.
人工知能学会論文誌, Vol.31, No.2, p.E-F43_1-12, 2016. 査読あり. http://doi.org/10.1527/tjsai.E-F43 .

(2) Mathieu Senelle, Silvia Garcia-Diez, Amin Mantrach, Masashi Shimbo, Marco Saerens, and François Fouss.
The sum-over-forests density index: Identifying dense regions in a graph.
IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, No. 6, pp.1268–1274. 2014. 査読あり. http://dx.doi.org/10.1109/TPAMI.2013.227 .

(3) Ilkka Kivimäki, Masashi Shimbo, and Marco Saerens.
Developments in the theory of randomized shortest paths with a comparison of graph node distances.
Physica A: Statistical Mechanics and Its Applications, Vol. 393, pp. 600–616, 2014. 査読あり. http://dx.doi.org/10.1016/j.physa.2013.09.016 .

(4) 小嵜 耕平, 新保 仁, 小町 守, 松本 裕治.
相互 k-近傍グラフを用いた半教師あり分類.
人工知能学会論文誌 Vol. 28, No. 4, pp. 400-408, 2013. 査読あり. http://doi.org/10.1527/tjsai.28.400 .

(5) 原 一夫, 鈴木 郁美, 新保 仁, 松本 裕治.
文法的・意味的共起を利用した単語類似度の計算.
人工知能学会論文誌 Vol. 28, No. 4, pp. 379-390, 2013. 査読あり. http://doi.org/10.1527/tjsai.28.379 .

(6) 鈴木 郁美, 原 一夫, 新保 仁, 松本 裕治.
ラプラシアンカーネルによるハブの解消.
人工知能学会論文誌 Vol. 28, No. 3, pp. 297–310, 2013. 査読あり. http://doi.org/10.1527/tjsai.28.297 .

[学会発表] (計 6 件)

(1) Masashi Tsubaki, Kevin Duh, Masashi Shimbo, and Yuji Matsumoto.
Non-linear similarity learning for compositionality.
In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI '16)*, pp. 2828–2834, Phoenix, Arizona, USA. 2016. 査読あり.

(2) Kazuo Hara, Ikumi Suzuki, Masashi Shimbo, Kei Kobayashi, Kenji Fukumizu, and Miloš Radovanović.
Localized centering: Reducing hubness in large-sample data.
In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2645–2651. Austin, Texas, USA. 2015. 査読あり.

(3) Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu.
Centering similarity measures to reduce hubs.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, pp. 613–623. Seattle, Washington, USA. 2013. 査読あり.

(4) Masashi Tsubaki, Kevin Duh, Masashi Shimbo, and Yuji Matsumoto.
Modeling and learning semantic co-compositionality through prototype projections and neural networks.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, pp. 130–140. Seattle, Washington, USA. 2013. 査読あり.

(5) Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Yuji Matsumoto, and Marco Saerens.
Investigating the effectiveness of Laplacian-based kernels in hub reduction.
In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI '12)*, pp. 1112–1118, Toronto, Ontario, Canada. 2012. 査読あり.

(6) Kazuo Hara, Ikumi Suzuki, Masashi Shimbo, and Yuji Matsumoto.
Walk-based computation of contextual word similarity.
In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*, pp. 1081–1096, Mumbai, India. 2012. 査読あり.

[図書] (計 1 件)

(1) François Fouss, Marco Searens, and Masashi Shimbo.
Algorithms and Models for Network Data and Link Analysis.
Cambridge University Press. 2016 (In Press). 520 pp. ISBN 978-1-107-12577-3.

6 研究組織

(1) 研究代表者
新保 仁 (SHIMBO, Masashi)
奈良先端科学技術大学院大学・情報科学研究科・准教授
研究者番号 90311589

(2) 研究分担者
原 一夫 (HARA, Kazuo)
国立遺伝学研究所・生命情報研究センター・研究員
研究者番号 30467691

鈴木 郁美 (SUZUKI, Ikumi)
山形大学・理工学研究科・助教

研究者番号 20637730
(平成 27 年度より研究分担者)