ISS-A-053

単語ベクトルを用いた企業間類似度の時間的遷移の検出

荒谷 優也† Gohourou Didier† 桑原 和宏†† 黄 宏軒††
†立命館大学大学院情報理工学研究科 ††立命館大学情報理工学部

1. はじめに

近年，活動拠点を複数の国で持つような多国籍企業が多く存在している．そのような企業では，新たな場所での事業展開の際に，関係する企業の調査が必要である．特に，時間が経つにつれて，企業間の関係がどのように変化したかを知ることは，それぞれの企業の背景を理解する上で重要である．本研究では，時間的遷移に伴って，企業間の類似度がどのように変化するのかを検出することを目的とする．

2. 企業間のコサイン類似度の時間的遷移

word2vec[1]を使うことで，テキストコーパス内の単語のベクトル表現を求めることができる．本研究では，英語版Wikipediaのスナップショットをもとに，企業間のコサイン類似度を求める．

予備実験として，FacebookとTwitter，AmazonとGoogle，AppleとGitHubの，3つの企業間のコサイン類似度の時間的遷移を求めた．テキストコーパスとして使用したのは，2002年1月1日， 2003年8月24日，2005年4月15日，2006年12月6日，2008年7月28日，2010年3月20日，2011年11月10日，2013年7月2日，2015年2月22日，2016年10月14日の英語版Wikipediaのダンプデータで，wikitm[1]を用いて作成したものである．図1に，企業間のコサイン類似度の遷移を示す．

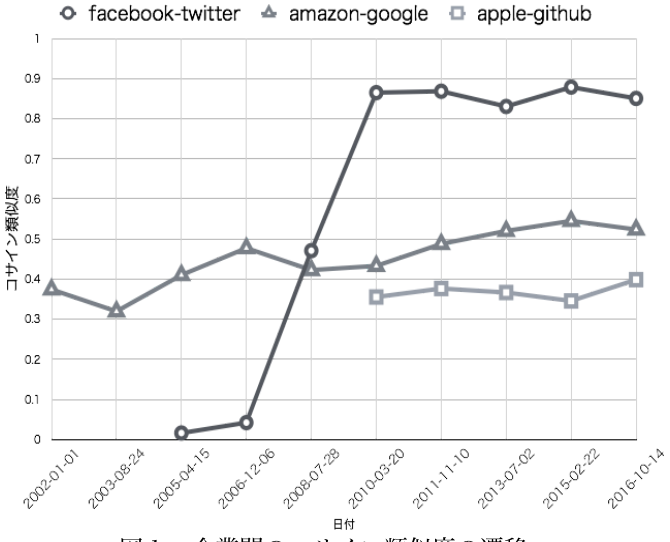

図１．企業間のコサイン類似度の遷移

3. コサイン類似度の時間的遷移

3.1 企業間の変化の検出例

2003年8月24日から2006年12月6日において，AmazonとGoogle間の類似度が約0.158上昇している．これは，2004年11月にAmazonによりAmazon Simple Queue Serviceを開始したことと，2006年3月にAmazon Web Services の提供を初めたことが原因であると考えられる．すなわち，Amazonがクラウドコンピューティング事業に進出することで，企業間の類似度が上昇したと解釈できる．このように企業間の類似度が時間的に変化することから，企業に関するイベントが発生したことが推測できる．

3.2 ページ内容の変化

AppleとGitHub間の類似度は，2010年3月20日から2015年2月22日にかけて， 最大0.021の変化がみられるが，これは日に日に変化していくテキストコーパスの影響によるノイズだと思われる．そのため，企業間の類似度の変化として，意味があるもののみを取り出すためには，テキストコーパスの細かい変化によるノイズの除去を行うことが必要である．

3.3 ページ作成などによる影響

2006年12月6日から2010年3月20日にかけて，FacebookとTwitter間の類似度は， 約0.823増加している．これは，何かしらの出来事が起こったためではなく， Wikipedia内に存在していなかった企業のページが作成された，もしくはページが大幅に記述された事によるものであると考えられる．確かに類似度が大きく変化しているが，これは企業間で何かしらの動きがあったのではないため，このような変化は，フィルタリングする必要がある．

4. おわりに

本稿では，単語ベクトルを利用して，企業間の関係の時間的遷移を検出する手法を提案した．予備実験として， Wikipediaをテキストコーパスとして，いくつかの企業間の類似度の時間的遷移を求めた．その結果，類似度の変化と企業間の出来事の対応がとれる可能性があることがわかった．

今後の課題として，fastText[2]など別の手法での類似度算出，類似度変化の要因の検出，ページ作成の影響やノイズの除去を試みる予定である．

参考文献

[1] Tomas Mikolov, et al., Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems 26, pp. 3111-3119, 2013.
[2] Piotr Bojanowski, et al., Enriching word vectors with subword information, CoRR, Vol. abs/1607.04606, 2016.

[1]https://github.com/semlab/wikitm