時系列テキストストリームからの単語共起を使った新情報検出法

New Event Detection with Time Subtraction Co-occurrence Words

竹田隆治 Takaharu Takeda　高須淳宏 Atsuhiro Takasu

何がわかる？

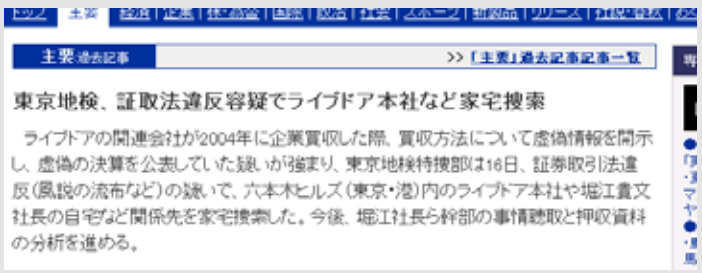
日々配信されるオンラインニュース　その記事を見て一連のニュース中の話題の流れの中で特に重要な展開などがああたところ検出する。　天災と復興、事件と犯人逮捕、などといった新情報の即時的検出を目標とする

どんな研究？

既知の話題に対しての関連性と、それらに対しての乖離度、新規性という、旧来は相容れない概念として捉えられてきたが、本研究では新規性と関連性という二つの尺度を独立した値として両立させることが可能であると実証する

状況設定

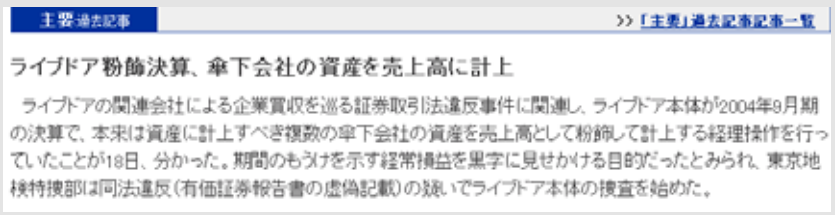
通常のテキストマイニングではある話題について言及した記事を集めることにみの集中し、話題に関する細かい変遷についてまでは深く感知しない


主要過去記事　>>「主要」過去記事記事一覧

東京地検、証取法違反容疑でライブドア本社など家宅捜索

ライブドアの関連会社が2004年に企業買収した際、買収方法について虚偽情報を開示し、虚偽の決算を公表していた疑いが強まり、東京地検特捜部は16日、証券取引法違反(風説の流布など)の疑いで、六本木ヒルズ(東京・港)内のライブドア本社や堀江貴文社長の自宅など関係先を家宅捜索した。今後、堀江社長ら幹部の事情聴取と押収資料の分析を進める。

2006/1/16　NIKKEI NET


主要過去記事　>>「主要」過去記事記事一覧

ライブドア粉飾決算、傘下会社の資産を売上高に計上

ライブドアの関連会社による企業買収を巡る証券取引法違反事件に関連し、ライブドア本体が2004年9月期の決算で、本来は資産に計上すべき複数の傘下会社の資産を売上高として粉飾して計上する経理操作を行っていたことが18日、分かった。期間のもうけを示す経常損益を黒字に見せかける目的だったとみられ、東京地検特捜部は同法違反(有価証券報告書の虚偽記載)の疑いでライブドア本体の捜査を始めた。

2006/1/18　NIKKEI NET

マンション２０軒とホテル１軒の耐震性に問題が…	→	ホテル1軒の営業休止…
カトリーナで被害を受けた石油関連施設…		ＩＥＡの加盟各国が石油の戦略備蓄を緊急放出…
郵政民営化法案の再提出に向け、….		11日午後の衆院本会議で可決、参院に送付される…

上記のような例は既存研究ではうまく検出できなかったり、または時間的に十分後になってからでなければ検出できなかったりしたこのような重要な変化を捉える

研究状況

手法の有効性

- データ量に対して線形時間で処理ができる
- 話題を形成する新しい語がなくても、新情報だと判定できる
- 単語共起の差分として現れる単語対の中には新情報を示す説明的な単語対が現れる

一連のニュース中で共通して現れる語	差分として現れる新出単語対
耐震 強度 偽装 マンション	一連 刑事 広域 態勢 確立 立件 詐欺
弁護士 府警 西村 比例 近畿	臨時 離党 受理 除籍 辞職
税制 減税	伊吹 文明 懇談 酒税 ビール 節税 原料 種類 簡素 普通
大相撲 横綱 青龍 制覇	偉業 成し遂げ　輪島 抜い 歴代

提案手法では文書レコードを単語共起の集合として表現する。

文書レコード中に出現する単語を並べたリスト Di

$$D_i = \{W_1, \ldots W_n\}$$

この文書 Di 中で前後 n 語以内に出現する全ての２単語のペア

$$CW_i = \{(W_i, W_j) \mid W_i, W_j \in D_i, i \pm n = J, |n| < N\}$$

提案手法では、新文書と(累積)既知単語対集合との比較で新情報を検出する。

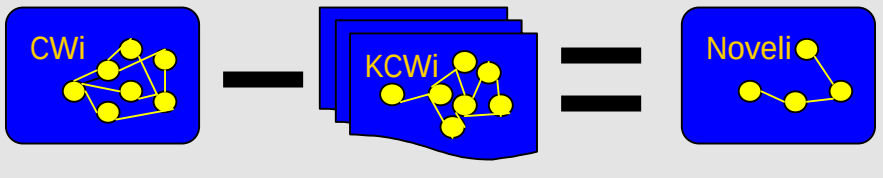
S_d=累積既知文書の幅(document-window)

既知単語対 KCW_i

新出単語対 $Novel_i$

$$KCW_i = \bigcup_{j=i-S_d}^{i-1} CW_j$$

$$Novel_i = CW_i - KCW_i$$



NII
連絡先：高須淳宏 国立情報学研究所　コンテンツ科学研究系　教授
TEL：03-4212-2519　FAX：03-3556-1916　Email：takeda_takaharu@grad.nii.ac.jp , takasu@nii.ac.jp