人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-05

ブログテキストの分析に基づく語の意味の経時変化可視化の試み

An Experimental Result on Visualization of Word Sense Changes by Blog Text Analysis

石川 雅弘 [1*]
Masahiro ISHIKAWA[1]

[1] 高崎健康福祉大学
[1] Takasaki University of Health and Welfare

Abstract: More than a decade have passed since blog or SNS became common. Massive amount of user-generated text has already been accumulated on the web. Many researchers are trying to analyze accumulated texts trying to exploit them in many fields. In such analysis, treatment of text meaning is important. Text is composed of words, thus word sense treatment is essential. However, word meanings undergo changes, thus we should consider word sense changes over time. In this paper, we present a method to detect and visualize word sense changes. The proposed method uses Random Indexing technique, which is based on the distributional hypothesis of word meanings. The result of an experiment on blog texts is also presented.

1 はじめに

Web や SNS，レビューサイトなどの普及により，記者や作家などの職業的文章生産者だけではなく，一般個人により生産された大量のテキストの蓄積が進んだ．そこにはかつてならば音声発話として消えていったような個人的な意見や感情の表明も含まれており，流行分析や評判分析，マーケット分析など様々な活用が試みられている [1]．今後もテキストデータの蓄積は継続的に拡大していくと考えられ，その有効活用を考える必要がある．

自然言語で記述されたテキストの活用において重要な課題の一つが意味処理である．そこでは，テキストの文字列としての一致・不一致だけではなく，その表す意味を適切に扱う必要があるが，その基礎として，単語の意味の扱いが重要である．

しかし，単語の意味は時間とともに変化する可能性がある．変化の速い現代においては単語の意味変化や新たな語義の獲得も速いと考えられるが，職業記者のように統制されていない一般個人の言語使用においては，その傾向は一層強いであろう．また，例えばトレンド分析など経時変化に対する感度が重要な分析においては，その考慮が一層重要である．企業や商品に対するイメージや人気の変化も，企業名や商品名の利用文脈の変化としても表れると考えられ，単語の意味・利用文脈の変化を分析することはマーケティングなどにおいても有用性があると考えられる．

このような観点から，ブログデータを対象として，単語の意味変化の検出とその可視化を試みた．本稿ではその手法と結果を報告し，今後の課題を検討する．

*連絡先：高崎健康福祉大学 健康福祉学部 医療情報学科
〒 370-0033 群馬県高崎市中大類町 37-1
E-mail: ishikawa@takasaki-u.ac.jp

2 単語のベクトル表現

自然言語処理おいては，単語や文書を数値ベクトルとして表現するベクトル空間モデルが一般的である [2]．基本的なベクトル空間モデルでは，n 個の文書の集合を $m \times n$ 単語–文書行列 $\mathbf{C}$ で表現する．

$$\mathbf{C} = \begin{pmatrix} c_{11} & \cdots & c_{1j} & \cdots & c_{1n} \\ \vdots & \ddots & & & \vdots \\ c_{i1} & & c_{ij} & & c_{in} \\ \vdots & & & \ddots & \vdots \\ c_{m1} & \cdots & c_{mi} & \cdots & c_{mn} \end{pmatrix}$$

$\mathbf{C}$ の要素 c_{ij} は，単語 i の文書 j における出現頻度や TF-IDF 値などの重みであり，列ベクトルが文書を，行ベクトルが単語を表す．また，単語や文書間の（非）類似度は，ベクトル間のユークリッド距離やコサイン尺度などで表わされる．このように一つの単語や一つの文書が一つの次元を構成する 1-of-k スタイルの表現では，行列 $\mathbf{C}$ は巨大な疎行列となり効率的な処理が難しい．また意味的処理にも適さないことから，LSI(Latent

Semantic Indexing)[3] などの手法により，より低次元で密な行列に変換されることが多い．しかし，全文書を入手してから一括して処理をする必要があるなど，時間を追って単語の意味変化を分析するためのベクトル表現方法としては適さない．

word2vec[4] や Random Indexing[5] では，巨大な疎行列を介さずに最初から密でより低次元な単語の分散表現を生成できる．word2vec は生成した単語ベクトル間に加法構成性があると見られるという点でも注目されているが，Random Indexing はより単純な計算で単語ベクトルを生成できる上，文書集合が増加した時の漸増的計算も容易であるという利点があり，時間とともに変化する単語ベクトルの生成手法として適している．また，できるだけ大規模で網羅的な処理を行う場合でも，処理の分散と結果の集約が容易という利点がある．そのため，本研究では Random Indexing を用いて単語ベクトルを生成し，それを分析することで意味の変化の検出と可視化を試みる．

2.1 Random Indexing

Random Indexing は，単語の意味の分布仮説に基づく単語ベクトル生成手法である [5]．分布仮説では，ある単語の意味はその出現文脈に現れる他の単語群により決定されるとされる．例として次のような文章を考える（ここでは分かち書きした各部を単語とする）．

春 は 桜 が 咲き ます

ここで各語には「索引ベクトル」と呼ばれる固有のベクトルが割り当てられているとする．この時，各語の前後 k 語の範囲をその語の文脈とし，文脈中の各語の索引ベクトルを合計することでその単語のこの出現における「文脈ベクトル」を得る．文脈ベクトルは単語のその文脈による意味付けである．例えば単語の前後 2 語までを文脈とすると,「春」,「は」,「桜」,「が」,「咲き」の各索引ベクトルの合計が「桜」のこの出現における文脈ベクトルである．ある単語のベクトル表現は，その単語の全文書における全出現の文脈ベクトルを合計することで得られる．

単語に索引ベクトルを割り当てる時点では単語の意味も類似性も不明のため，索引ベクトルは互いに直交であることが望ましい．しかし，1-of-k スタイルで各語に互いに直交なベクトルを割り当てると，単語の異なり総数に等しい m 次元が必要となり，疎な高次元ベクトルとなってしまう．

Random Indexing では，索引ベクトルの次元を単語の異なり総数 m より小さな値 m'（$\ll m$）とし，各語の索引ベクトルとして m' 次元の擬直交ベクトルを割り当てる事で m' 次元の単語ベクトルを生成する．これは高次元ベクトル空間では次元数より遥かに多い擬直交ベクトルが存在するという性質を利用している．ここで $\vec{u},\vec{v}$ が擬直交ベクトルであるとは，$\vec{u}\cdot\vec{v}\approx 0$ を意味する．

なお，一定の条件を満たしたランダムなベクトルを生成することで擬直交ベクトル群を得られることが示されており [6]，本研究では論文 [7] で提案された手法を用いる．

Random Indexing では，単語ベクトルは文書集合全体におけるその単語の全出現の文脈ベクトルを単純に合計することで生成できる．そのため，時間とともに文書集合が増加する場合でも，逐次的に新たな文書における文脈ベクトルを求め，それを過去の文書集合から計算された単語ベクトルと合計することで最新の単語ベクトルを求められる．したがって，新たな文書における文脈ベクトルとそれを合計した最新の単語ベクトルが時間とともにどのように変化するかを分析することで，単語の意味変化を捉え得ると考えられる．

3 提案手法

3.1 単語ベクトルの生成法

本稿で分析対象とするのは，Web から収集したブログ記事テキストである．ブログ記事には作成日時が付されているため，記事集合を一定期間ごとに分割し時間順に整列することができる．

対象とするブログ記事の集合を D とし，それらを月別に分割し $D_0, D_1, \cdots, D_{T-1}$ とする．D_t は第 t 月目に生産されたブログ記事集合である．

文書集合から Random Indexing に基づいて単語ベクトルを生成するには，単語の文脈を前後何単語の範囲とするかを決める必要があるが，今回は簡単のため単語が含まれるブログ記事テキスト全体を文脈とする．すなわち単語 w_i の D_t から求めた単語ベクトルは

$$v_i^{(t)} = \sum_{d\in D_t}\sum_{w\in d} w \text{ の索引ベクトル}$$

となり，これを単語 w_i の第 t 月における月別ベクトルと呼ぶ．また，第 t 月目までの全てのブログ記事から得られる w_i の単語ベクトルは

$$V_i^{(t)} = \sum_{j=0}^{t} v_i^{(j)} = \sum_{j=0}^{t}\sum_{d\in D_j}\sum_{w\in d} w \text{ の索引ベクトル}$$

であり，これを単語の第 t 月における累積月別ベクトルと呼ぶこととする．第 t 月における w_i の単語ベクトルは，累積月別ベクトルを長さ 1 に正規化したものであり，下式で与えられる．

$$\hat{V}_i^{(t)} = \frac{V_i^{(t)}}{\|V_i^{(t)}\|}$$

本稿の目的は，単語ベクトルの変化を追跡することで，分布仮説に基づく単語の意味，すなわち利用文脈の変化を分析し，その可視化を試みることである．

3.2 変化の追跡と検出方法

ベクトル表現された単語間の類似度としては，一般的に用いられているコサイン尺度を採用する．長さが 1 に正規化された二つの単語ベクトル $\vec{u}, \vec{v}$ のコサイン尺度は下式で与えられる．

$$cosine(\vec{u}, \vec{v}) = \sum_i u_i v_i$$

単語の意味が恒常的であれば，二つの期間 s, t における単語ベクトル $\hat{V}_i^{(s)}, \hat{V}_i^{(t)}$ はほぼ等しいことが期待でき，類似度は

$$cosine(\hat{V}_i^{(s)}, \hat{V}_i^{(t)}) \approx 1.0$$

となる．逆に意味に変化があれば類似度は低下する．したがって，各時点の単語ベクトルについて，同じ単語の過去の時点でのベクトルとの類似度を追跡することで意味変化を検出できる可能性がある．

3.3 変化内容の分析方法

単語ベクトルは，文脈中に共起した単語集合の索引ベクトルを合算したものであり，単語ベクトルの変化は共起する単語集合が変化したことを意味する．したがって，共起する単語集合のクラスター構造の変化を分析することで，意味変化の内容を知ることができると考えられる．

単語ベクトルのクラスタリングには，SOM (Self-Organizing Maps) [8] の一種である Batch Map を用いる．SOM はデータを二次元空間や一次元空間上に整列したセルに写像し，データ空間上のクラスター構造を低次元空間上に「再現」する．そのため高次元空間中のクラスター構造の可視化に利用される．本研究では，SOM によるクラスタリング結果の可視化手法としては [10] で提案した極座標ヒストグラムと面グラフを修正した手法を用いる．これらについては 4 で述べる．

本研究で用いる Batch Map はバッチ学習型の SOM であり，一般的な逐次学習型の SOM と比べて計算効率が良い．また，後述する近傍半径が 0 の場合には k-means クラスタリングと一致する．また，一般的な SOM や k-means クラスタリングではデータ間の非類似度としてユークリッド距離を用いるが，ここでは類似度としてコサイン尺度を用いる．従って，本研究で用いる SOM は Dot Product Batch Map[8] であり，近傍半径が 0 の場合 spherical k-means クラスタリングに一致する．データの写像先のセルは k-means クラスタリングのクラスターに対応するが，SOM ではそれらの間に二次元または一次元上の隣接関係が与えられ，隣接したセルには類似したクラスターが配置されるようにクラスタリングが行われる．そのため，得られたクラスター間の類似の度合いを測ることができる．本研究では、環状に配置されたセルを用いる．

3.3.1 Dot Product Batch Map

Dot Product Batch Map によるクラスタリング手順を示す．ここで，セルの数は k-means クラスタリングにおける k であり，求めるクラスター数に対応する．また，セルは環状につながっているものとする．

1. 各セルの中心ベクトルを初期化する．
2. 各データをコサイン尺度が最も小さいセルに割り当てる．
3. 各セルの中心ベクトルを，近傍半径 R 内のセルに割り当てられた全てのデータの平均ベクトルで更新する．
4. 収束するまで (2), (3) を繰り返す．ただし手順 (3) の半径 R は大きな値から徐々に減少させる．

4 実験

本節では，実際のブログデータを対象とした単語ベクトルの変化と変化内容の可視化例を示す．ただし，追跡期間において明らかに使用文脈に変化が生じたと考えられる単語である「福島」のみを対象とした．この単語の使用文脈の分析は，社会が受けたインパクトの大きさや風評の広がりと収束の様子を知るためにも有意義だと考える．

4.1 データセット

2011 年から 2012 年にかけて goo ブログ [9] の新着記事 RSS で捕捉した 34756 ブロガーのうち，2011 年より前に「福島」を含む記事を投稿しており，かつ 2012 年 3 月 11 日以降も記事を投稿している 5714 ブロガーが 2010 年 1 月 1 日から 2012 年 3 月 31 日までに投稿した記事を対象とした．

記事テキストは MeCab(v0.97)[11] により形態素解析を行ない，名詞と判定された語のみを抽出し分析に用いた．ただし代名詞，非自立語，接頭辞，接尾辞，数詞，サ変接続詞は除いた．また，出現数が 10 未満のものと 1 文字のみのものも除いた．なお，MeCab 用辞書としては IPA 辞書を用いた．

最終的に対象となったブログ記事数は 3,690,657，単語の異なり総数は 507,532 である．

また，月ごとの変化を追跡するために，2010 年 1 月から 2012 年 3 月までのひと月ごとにデータセットを分割した．

4.2 単語ベクトルの生成

3.1 で示した手順に従い，全ての単語の全期間の月別単語ベクトルと累積単語ベクトルを作成した．索引ベクトルは 200 次元であり，したがって単語ベクトルも 200 次元である．

4.3 自己類似度の変化の可視化

まず,「福島」ベクトルの第 $t-1$ 月と第 t 月の月別ベクトル間類似度，第 0 月と第 t 月の累積月別ベクトル間類似度を求める．結果を図 1 に示す．横軸は 2010 年 1 月を第 0 月とした月数である．一見して第 14 月，すなわち 2011 年 3 月に，それ以前とも以後とも大きく異なる月別ベクトルが生成されていることが分かる．また，それに伴い累積ベクトルにも急激な変化が生じている．

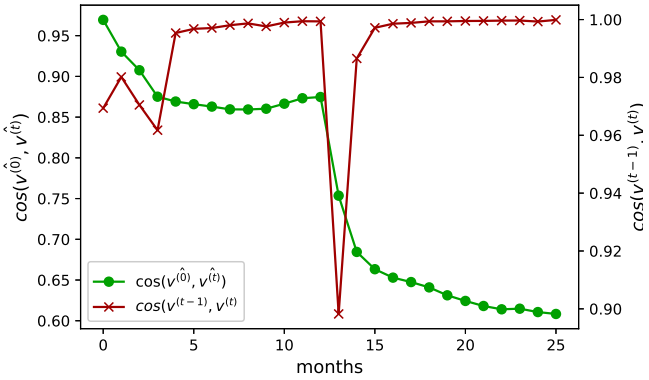

図 1: 「福島」ベクトルの自己類似度の変化

比較のため，同時期に「福島」ほどの変化は生じなかったと考えられる「沖縄」の自己類似度の様子を図 2 に示す．第 14 月に「福島」のような変化は見られないことが分かる．

4.4 極座標ヒストグラムによるクラスター構造の可視化

図 1 から，第 14 月に「福島」の出現文脈が大きく変化したことは分かった．ここからはその変化の内容を分析する．

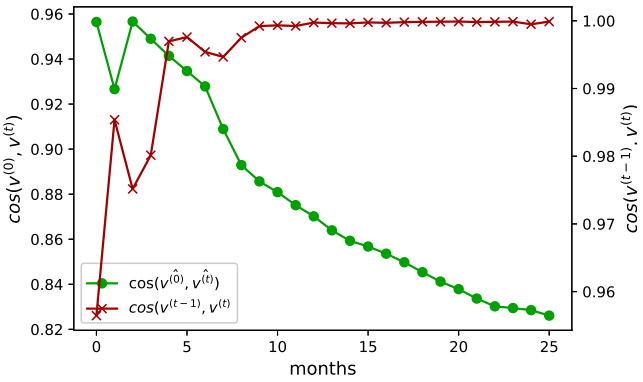

図 2: 「沖縄」ベクトルの自己類似度の変化

文脈の変化とはすなわち文脈内で共起する単語集合の変化である．そこでまず，最終月における全ての単語ベクトルをクラスタリングした結果を図 3 に示す．SOM は，k-means クラスタリング同様クラスター数 (k) を指定する必要があるが，本実験では $k = 128$ とした．図の各バーが一つのクラスターを表し，バーの長さはそのクラスターを構成する単語の数を表す．隣接した二つのバーの間の角度は，二つのクラスターの中心の分離度を示す．すなわち，角度が小さいほど二つのクラスターの中心同士の類似度が大きく，狭い角度に多くのクラスターが密集している範囲は，実際には大きな一つのクラスターが，大きする k を与えられたため過剰に分割された可能性がある．また，隣接した二つのバーの角度が小さいほど類似した色が配されている．

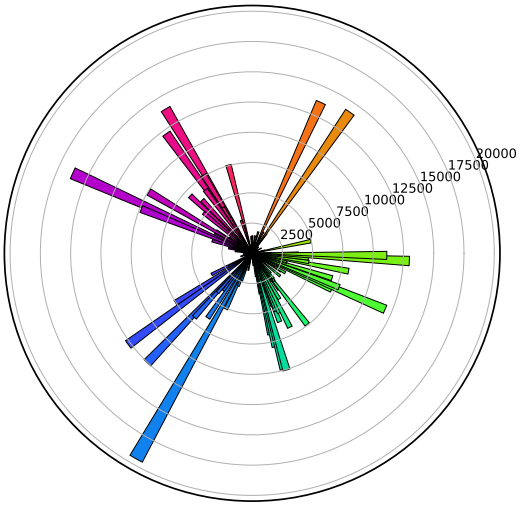

図 3: 最終月における全単語ベクトルのクラスター構造（バーの長さはクラスターに属する単語の数）

この図からは，全単語群が全体的にいくつかの大きなクラスターに別れ，さらに大クラスターが細分され

ているらしい様子が分かる．

4.5　文脈とその変化の可視化

図 3 のクラスター構造をベースに,「福島」の文脈の変化の可視化を試みる．

図 4 に示すのは，2011 年 2 月と 3 月それぞれについて,「福島」と共起した単語のみに絞って描いた極座標ヒストグラムである．ただしバーの長さは，共起した単語のうちそのクラスターに属する単語がその月に出現した頻度である．彩色については図 3 とは異なる方針を取り，隣り合ったクラスターがなるべく異なる色を持つようにした．これは，密集した領域でクラスターサイズの変化があった場合，類似色では判別しづらいからである．

このヒストグラムは，いわば「福島」のその月の文脈を可視化したものであり，異なる期間のヒストグラムを比較することで，その変化を見ることができる．図中矢印で示したように，2 月には大変小さかった二つのクラスターが 3 月には爆発的に増大していることがわかる．図 1 の自己類似度チャートの第 14 月における大きな変化はこのクラスターが原因だと考えられる．

4.6　面ヒートマップによる文脈変化の可視化

図 5 に示すのは，図 4 と同じデータの全期間分を一枚の面グラフとして表示したものである．横軸が月数，縦軸は 128 個のクラスターの相対サイズを積み上げたものである．また，各期間の各クラスターの相対サイズを表す領域は相対サイズの値でヒートマップ風に彩色した．青がサイズが小さいことを，赤が大きいことを表している．

面グラフは，全期間に渡る全クラスターの変化の様子を一枚で可視化できる点が優れているが，面積だけで表しているとクラスター数が多い場合には判別し難いという問題があったが，ヒートマップ風彩色により変化を把握しやすくなっている．

図中楕円で示したように，第 14 週において二つのクラスターが増大していることがわかる．これは図 4 で増大していた二つのクラスターと同じものである．

これらのクラスターを精査することで，実際にどのように単語の使用文脈が変化したのかを分析できる．

5　まとめと今後の課題

ブログや SNS などの一般ユーザーが生産した大量のテキストデータの蓄積を背景に，テキストデータの利活用が進んでいる．しかしテキストデータの活用のためにはまずは単語の意味を適切に扱う必要がある．しかし単語の意味そのものも時間とともに変化するため，変化があった時にそれを捉えられる必要がある．

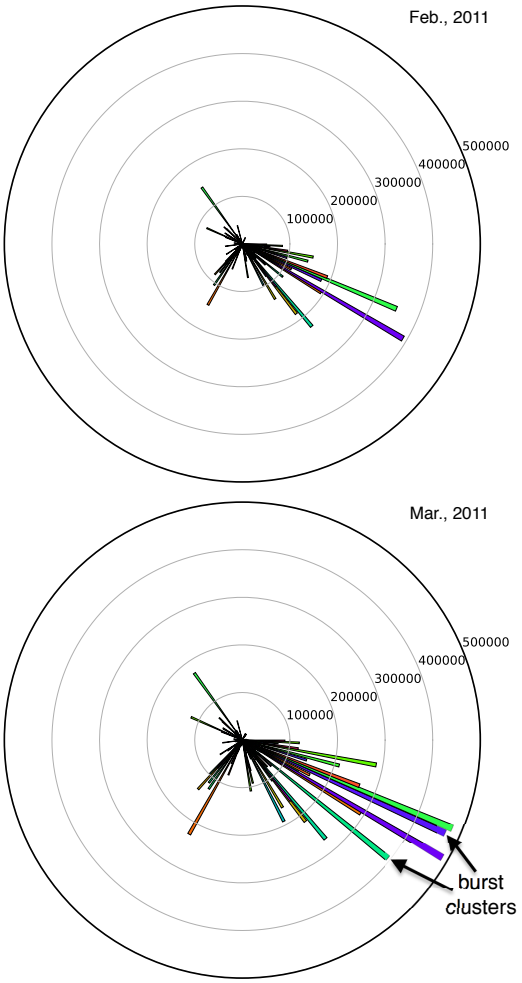

図 4: 2011 年 2 月・3 月にそれぞれ「福島」と共起した単語群のクラスター構造（バーの高さはクラスターに属する単語のその期間における出現頻度）

そこで分布仮説に基づいた単語ベクトル生成手法である Random Indexing を利用して，単語の意味の経時変化の可視化を試みた．対象としたデータセットの期間内において大きな変化があったと考えらえた「福島」を例として，自己類似度チャート，Batch Map による単語クラスタリング，極座標ヒストグラムと面ヒートマップによる可視化例を示した．

例えば商品名をターゲットにこのような分析を行えば，商品の人気や評判がどのタイミングでどのように変化したのかを知るのにも役立つと考えられる．

しかし今回分析の対象としたのは，極めて例外的に短時間で大きな文脈変化があったと考えられた単語であり，その他の一般的な単語の意味変化を捉え可視化できるかどうかは明らかではない．その確認のために

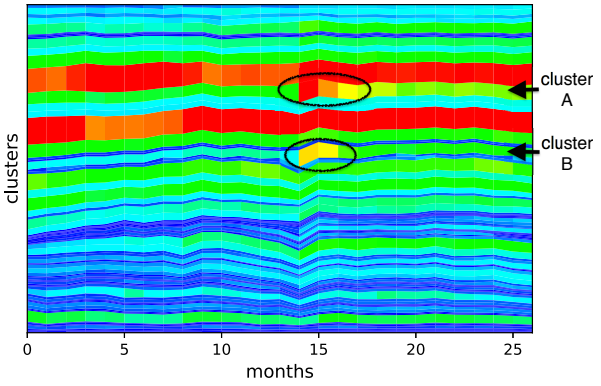

図 5: 「福島」と共起した単語群のクラスター構造の面グラフ・ヒートマップ表現（幅と色は相対クラスターサイズを表す）

は，データセットの量，特に期間の延長が必要であり，より長期にわたってより穏やかに変化する単語の分析を行う必要がある．今回使用したブログデータをベースに，今後それ以降の記事の収集を進めたい．

また，今回のように特定の単語をターゲットに据えて分析するのではなく，変化のあった可能性のある単語そのものを発見するには，ヒストグラムやヒートマップはあまり役立たないであろう．単語の数は膨大であり，その全てのチャートを作成して目視することはできないからである．変化のあった可能性のある単語を自動的に検出し，それらについてのみチャートを作成し目視するのが現実的である．

したがって，変化を自動検知する手法を考える必要があるが，これは各単語の自己類似度を追跡することで可能であると考えており，今後取り組みたい．

極座標ヒストグラムや面ヒートマップの基礎となるクラスタリングにも問題が残る．

今回の実験では最終期間における累積ベクトルを用いてクラスタリングを行い，過去の各期間のヒストグラムやヒートマップはそのクラスターに各期間の単語を割り当てて作成した．しかし，単語ベクトルは変化しているというのが本研究の出発点である．したがって，最終期間の累積ベクトルと，過去の時点での「同じ」単語の累積ベクトルは異なっている可能性がある．表層の文字列が同じだからと言う理由で異なる期間の単語を「同じ」ものとして扱うことはできないはずである．しかしそれらを別のものとして扱うためには，表層文字列とは別の「概念」を表す記号が必要となり，煩雑となる．また，現在扱っているデータセットの期間は短いため，このような配慮が実際に必要になる単語はほとんどないと考えられ，試みとしての本実験では追求しなかった．将来的により長期に渡る意味変化の追跡を行う中ではこの点も考慮したい．

参考文献

[1] Junichi Kato, *Customers' Needs for Digital Terrestrial Television Broadcasting: An Analysis of Weblog Data*, Proceedings of The 8th International Conference on Innovation and Management, pp.1093–1096, 2011.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[3] Christos H. Papadimitriou et al., *Latent Semantic Indexing: A Probabilistic Analysis*, In Proc.of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp159–168, 1998.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, *Distributed representations of words and phrases and their compositionality*, Advances in Neural Information Processing Systems 26, pp3111-3119, Curran Associates, Inc. 2013.

[5] Sahlgren, Magnus, *An Introduction to Random Indexing*, In Proc. of the Methods and Applicatons of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, 2005.

[6] Kanerva P, Kristofersson J, Holst A, *Random indexing of text samples for latent semantic analysis*, In Proc. of the 22nd Annual Conference of the Cognitive Science Society, p.1036, 2000.

[7] Dimitris Achlioptas, *Database-friendly Random Projections*, In Proc. of the 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp274–281, 2001.

[8] Teuvo Kohonen, *Self-Organizing Maps, Third Edition*, Springer-Verlag, 2001.

[9] goo ブログ, `http://blog.goo.ne.jp/`.

[10] Masahiro Ishikawa, *Visualizing Cluster Structures and Their Changes over Time by Two-Step Application of Self-Organizing Maps*, Proceedings of the 2011 International Workshop on Behavior Informatics at the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011), pp.160–171, Shenzhen, China, May 2011.

[11] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.