分布仮説に基づく語の意味変化分析の試み
～311 前後のブログを例として～

石　川　雅　弘

概要

ウェブ上には，大量のテキストデータが蓄積されている。特に一般個人が自由に情報発信できるブログやソーシャルメディアにおいては，かつてならば音声発話として消えていったような極めて属人性の高い言語使用もテキストとして蓄積されており，単純な文字列検索だけではなく，流行分析や評判分析など様々にその活用が試みられている。

高度なテキスト利用において重要な課題の一つが意味的処理である。高度なテキスト利用においては，単純な文字列上の一致・不一致だけではなくその表わす意味を適切に扱う必要があり，テキストの構成要素である語の意味の扱いも重要な課題となっている。しかし，語の意味は時間と共に変化するものである。変化の速い現代においては語の意味変化や新たな語義の獲得も速いと考えられ，例えばトレンド分析などにおいては一層その考慮が必要となろう。ブログやソーシャルメディアでは，テキストにタイムスタンプが付随しているのが一般的であり，時間を追ってそのような分析をする土台が整いつつあると言える。

我々は，ブログ記事テキストを対象として，語の意味変化をどのように捉えることができるのか，語の意味の分布仮説に基づいた語の意味表現手法である Random Indexing を用いた試行を行なった。本稿ではその手法と結果について報告する。

キーワード：テキストマイニング，ブログ，分布仮説，Random Indexing

１．はじめに

ウェブ上には大量のテキストデータが蓄積されている。新聞などの伝統的マスメディアでは情報発信が職業的発信者に独占されていたが，ブログをはじめとするソーシャルメディアでは一般個人の誰でもが自由に情報発信することができ，その結果情報爆発と形容されるほどの情報量の増加を見ている[(3)]。ソーシャルメディアによる情報発信は，新聞雑誌などとは異なり，必ずしも整理され確認された内容を含むわけではなく，かつてならば独り言や周囲の人物との音声による雑談として消えていったはずのものも多い。そのような個人的な言語使用が，タイムスタンプ付きの文字テキストとして，しかも容易に収集・利用可能な電子データとして大量に蓄積されるようになったのは

歴史上初めてのことである。整理されていないとはいえ，あるいは，だからこそ，ウェブ上のテキストデータにはその時々の一般個人の関心や意見，あるいは社会情勢が反映されていると考えられ，それらを分析することで伝統的・形式的なアンケートでは得られなかった情報を獲得できる可能性がある。そのため，例えばブログ記事の市場分析への利用などが試みられている[1][6]。

そのような分析を行なう際には，テキストの文字列としての表面的な一致性だけではなく，その意味や含意を扱う事が必要となる。テキストはその要素単位である語から成り立っており，テキストの表わす意味を扱おうとする時，語の意味をどのように扱うかを考える必要があり，さまざまな手法が提案されている。しかし，言葉の意味は時と共に変化する。特に職業記者のように統制されていない個人の言語使用にあっては，その傾向は一層強いと考えられ，ウェブ上のテキストを分析する際の重要な問題である。

また，語の意味変化を分析しようとしても，従来ならば過去の紙の文献を分析するなど，非常に長いスパンでかつ乏しい資料を元に行なうしかなかった。しかしソーシャルメディア上にタイムスタンプ付きのテキストが大量に蓄積されるようになり，より大規模かつ（従来と比較して）短いスパンでの変化を分析をする土台が整い始めたと言える。

そこで，意味の分布仮説に基づく語の意味表現手法である Random Indexing を用いてブログ記事テキストを分析することで，語の意味変化を捉える試みを行なった。本稿ではその手法と結果，および今後の課題について報告する。

2．文書と語のベクトル表現

(1) ベクトル空間モデルに基づく語の意味表現手法

自然言語処理における文法構造の解析技術は未だその精度が十分とは言えず，ソーシャルメディアのような表記のゆれの大きな文書に適用するのはさらに難しい。そのため，ブログを含むテキストデータの分析では，文書（新聞記事やブログ記事などまとまりのある独立したテキスト）を bag-of-words と呼ばれるキーワードの集合として表現する事が多い。これは情報検索や文書クラスタリングなどでも一般的な手法で，文書をそれが含むキーワードの集合（要素の重複出現を許すため，正確にはマルチ集合，すなわち bag）で表すものである。文書を表わすキーワード集合は，効率的に扱うためにベクトルとして表現され，文書集合は行列として扱われるため，ベクトル空間モデルと呼ばれる[10]。

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。

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

C の要素 c_{ij} は，単語 （キーワード）i の文書 j における出現数や重みであり，列ベクトルが文書を，行ベクトルが単語を表す。二つの単語に対応した二つの行ベクトルについて考えると，より多くの文書で共起する単語同士ほどより類似したベクトルとなることが期待できる。また，同じ単語を多く含む文書ほど，対応する文書ベクトルが類似することが期待できる。ベクトル空間モデルでは，単語や文書の類似度はユークリッド距離や内積，コサイン尺度など，ベクトル間の演算で定義される。ここで，行列 **C** の行数 m は文書集合全体での単語の異なり総数であり，一般的に非常に大きな数値となる。それに対して一つの文書に含まれる単語数は小さいため，行列 **C** は巨大な疎行列となり，効率的な扱いが困難になることが多い。

そのため，意味的処理に適したより効率的な単語（文書）ベクトルの生成手法が提案されており，代表的なものに LSI（Latent Semantic Indexing），word2vec, Random Indexing などがある。LSI は構築した単語－文書行列 **C** を特異値分解することで，**C** より低次元で密な，かつより意味的関係を反映した行列を作成する[10]。しかし LSI で必要とされる特異値分解は，その計算量から扱える行列 （すなわち文書集合）のサイズに限界がある。word2vec [8]は，ニューラルネットへの入力および教師出力として文書集合を与え，それを学習させることで単語のベクトル表現を獲得する機械学習に基づく手法である。数百次元と効率的なベクトル表現を生成できる上，ベクトル間の加減算が意味的な演算になっているとして注目されているが，単語の意味変化を追跡するような漸増的生成には適さない。

本稿で示す手法は，単語ベクトルの漸増的生成に適した Random Indexing に基づいている。

(2) Random Indexing

Random Indexing[5]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。意味の分布仮説では，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）。

私　は　旅行　に　行き　たい

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表現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「旅行」の前後 2 語を文脈とすると，「私」,「は」,「に」,「行き」の索引ベクトルの合計が「旅行」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。結果として，類似した文脈に出現する語は類似したベクトル表現を得る。

各語に索引ベクトルを割り当てる時点では，各語間にどのような意味的関係があるかは不明であるため，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すると，単語の異なり総数に等しい m 次元が必要となり，超高次元の疎な単語ベクトルとなってしまい，効率的な処理ができなくなる。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，各語に m' 次元の擬直交ベクトルを割り当てる事で m' 次元の効率的な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ことができる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選択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[2]，本稿では論文[7]で提案された手法を用いている。

3．提案手法

(1) Random Indexing に基づく単語の日別ベクトルと累積日別ベクトルの生成

ブログやソーシャルメディアで発信された情報には，発信の日付や時刻であるタイムスタンプが付随するのが一般的である。そのため，テキストデータ群を時系列上に整列するのはたやすい。そこで，ブログ記事集合 D から日別のテキスト集合を作成し，D_1，D_2, ……，D_T とする。D_t は第 t 日目のタイムスタンプを持つテキストの集合である。

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，ここでは簡単のためある文書に含まれる単語の文脈はその文書全体であるとする。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，これを単語の日別ベクトル（daily vector）と呼ぶ。また，第 t 日目までの全てのテキストデータから得られる w_i の単語ベクトルは

$$V_i^{(t)} = \sum_{j=1}^{t} v_i^{(j)} = \sum_{j=1}^{t}\sum_{d\in D_j}\sum_{w\in d} w \text{ の索引ベクトル}$$

であり，これを単語の累積日別ベクトル（cumulative daily vector）と呼ぶ。このままでは出現数の多い単語ほど大きなベクトルを持つ事になるため，長さ 1 に正規化したものをそれぞれ

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

と置き，単位日別ベクトル（unit daily vector），単位累積日別ベクトル（unit cumulative daily vector）と呼ぶ。単位累積日別ベクトルはその時点までの全文書から生成した単語ベクトルであるので，これを単に（単語 w_i の日付 t における）単語ベクトルと呼ぶこととする。

単語ベクトル $\hat{V}_i^{(t)}$ の変化を追跡することで，その単語の意味の変化を検出することを試みるのが本稿の目的である。

4．実験

(1) 追跡対象とするキーワード

一般に大規模な文書集合に現われる単語の異なり総数は膨大である。その全てを追跡してもそこから意味変化のあった単語を発見するのは容易ではない。今回の実験は，そのような網羅的な実験に向けての予備実験として対象とするキーワードを一つに絞って行なう。また，提案手法の有効性を確認するためには，追跡期間内に観測可能な程度の意味的な変化があったと考えられるキーワードを選定する必要がある。分布仮説によれば単語の意味の変化はそれが用いられる文脈の変化として表われる。そこで追跡期間中にその単語の出現する文脈が（傾向として）大きく変化したと考えられる単語を選択することとした。追跡対象としたのは「福島」である※1。

(2) 実験データと前処理

実験は goo ブログ[4]から収集した34756ブロガーの中から，2011年３月11日より前に「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5921ブロガーの全ブログ記事を対象に行なった。追跡期間は2004年４月１日から2012年３月31日である※2。

文書（記事）を bag-of-words として表現するためには，記事テキストから単語を切り出す必要があるが，ここでは形態素解析ソフト MeCab (v0. 97)[9]により名詞と判定された形態素を単語と見做すこととする。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。なお，MeCab 用辞書としては IPA 辞書を用いた。

対象とした5921ブロガーによる追跡期間内のブログ記事総数は8,495,464，出現する単語の異なり総数は65,362,090である。基本的なベクトル空間モデルで表現する場合，約555テラ要素という巨大な疎行列になる。しかし今回追跡するキーワードは「福島」一語であるため，「福島」を含む記事だけに絞り込んでも結果は変わらない。絞り込んだ結果75,768記事，8,511,075単語となった。基本的なベクトル空間モデルの場合，約644ギガ要素という依然として効率的処理は難しい巨大な疎行列となる。

図１は，横軸に2004年４月１日を起点とした日数，縦軸に各日のタイムスタンプを持つ記事数を示したものである。図２には記事の累積数を示した。グラフから，一見して2011年３月11日を境に記事数が大きく増加していることが見てとれるが，これは東日本大震災および福島第一原発事故によるものであろう。

また，それ以前で最大の記事数を記録した2010年５月29日とその前後は，５月28日に沖縄県普天間基地移設問題を巡り当時の福島瑞穂大臣が更迭された件に関するものであった。Random Indexing は，このケースのように文字列上は一致してもその意味は異なる多義語などの扱いに問題があるとされるが，本稿ではこの点については扱わない。

※1　不幸な事故により，この語はそれ以前と以後で出現する文脈に大きな変化があったことは明らかである。風評と共に語られる事も多く，実験の対象とするのは適切ではないという指摘もあり得る。しかし，その語の意味，すなわち出現文脈の変化を観察することは，いわゆる風評の発生や収束を知る手がかりになるものであり，またその対策を考えるための手掛りを得ることにもなると考えられ，有意義なことだと考える。

※2　goo ブログは 2004年３月に開設された。

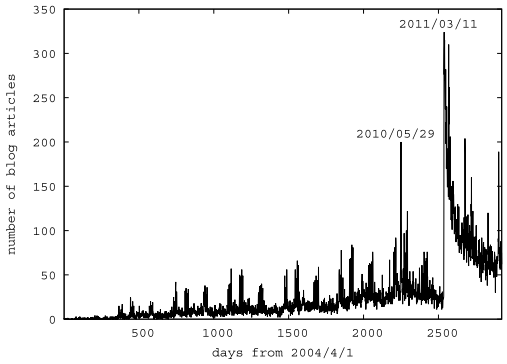

図1「福島」を含んだ日別ブログ記事数
（2004年 4 月 1 日起点）

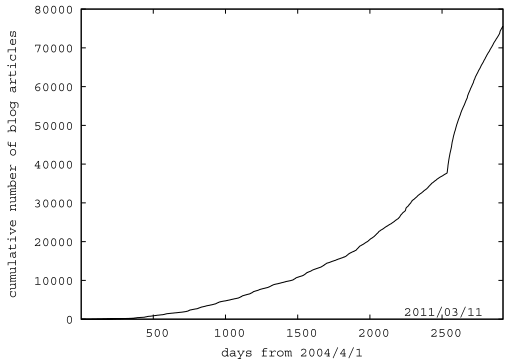

図2「福島」を含んだブログ記事累積数
（2004年 4 月 1 日起点）

(3) 結果と分析

提案手法を適用して，各日の単位日別ベクトルと単位累積日別ベクトルを生成した。得られたベクトルを用いて，その内容に変化が現われた点を検出できないか試みた。

一般的にベクトル空間モデルでは二つのベクトル $\vec{u}, \vec{v}$ の類似度をコサイン尺度で表わす。

$$cosine(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|}$$

コサイン尺度は二つのベクトルが為す角が小さいほど，すなわち類似したベクトル同士ほど 1 に近づく。図 3 に示すのは，当日とその前日の単位日別単語ベクトル間のコサイン尺度である。ブログ開設間も無い間は記事数も少なく安定しないため，2006年 2 月26日（この日以降は対象記事数が0の日は無い）を起点とした。図 4 は前日までの単位累積日別ベクトルと当日の単位単語ベクトルの

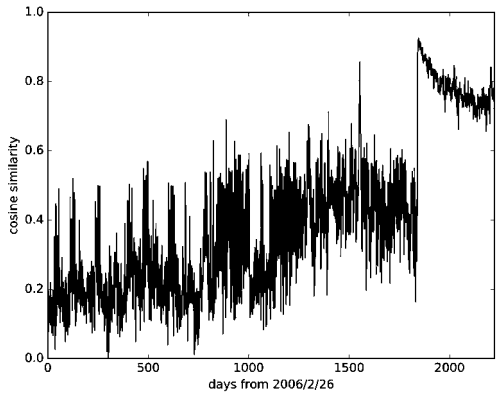

図3　第 t-1 日と第 t 日の日別ベクトル間コサイン類似度（2006年 2 月26日起点）

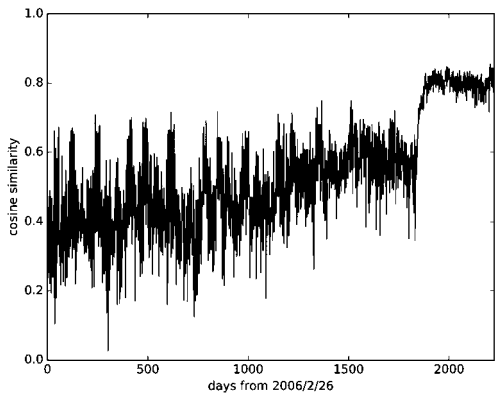

図4　第 t-1 日の累積日別ベクトルと第 t 日の日別ベクトルの間のコサイン類似度

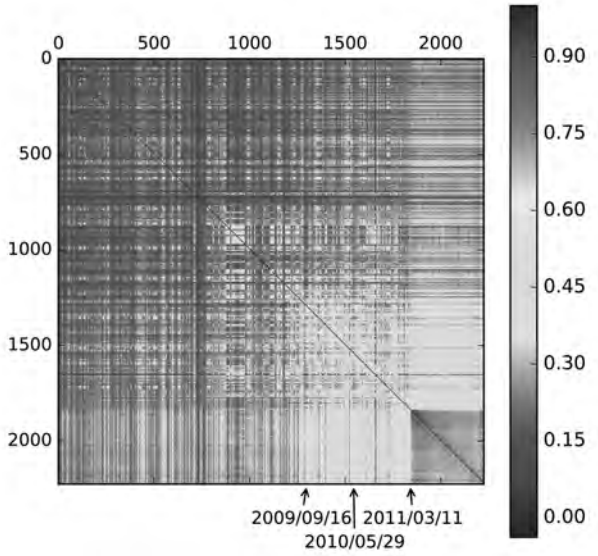

図５　日別ベクトル間のコサイン尺度のヒートマップ

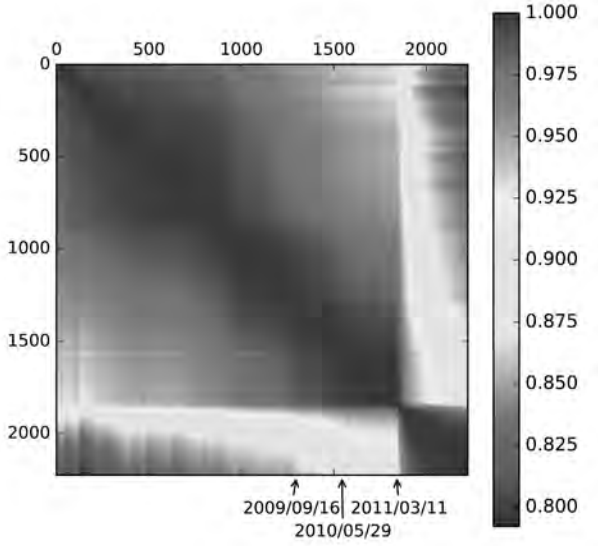

図６　累積日別ベクトル間のコサイン尺度のヒートマップ

コサイン尺度である。どちらのグラフからも，2011年３月11日に大きな変化があり，追跡期間中その傾向が継続している事が分かる。

隣接した二つ同士だけではなく，全体的な関係を見るため，2006年２月26日以降の全てのベクトル間のコサイン尺度をヒートマップとして可視化したのが図５，６である。これら二つのヒートマップは，言わば対称な類似度行列を可視化したものである。そのため，対角線上に浮かび上がるコサ

イン尺度の高いブロックは，類似したベクトルが並んだ範囲を表すと考えられる。

両グラフから，2011年３月11日以降は明らかに一つの類似ブロックを形成していることが見て取れる。すなわち，それ以前と以後では対称キーワードの出現文脈が大きく異なり，またそれ以後はほぼ類似した文脈に出現していることが分かる。

図 1で示した記事数が特に増加した 2010年５月29日を含む 2009年９月16日から2011年３月11日の期間にも若干ぼやけてはいるがブロックが確認できる。すでに指摘した通り2010年５月29日は当時の大臣である福島瑞穂氏が更迭された日であるが，2009年９月16日は鳩山由紀夫内閣が発足し，福島氏が入閣した日である。この事からこのブロックは福島瑞穂氏に関するものだと推測できる。元々意図していた「同じ」キーワードの意味変化ではなく，同じ文字列で別々のものを指す二つのキーワードが混合したケースではあるが，その文字列が出現する文脈の変化を検出することには成功していると言える。

５．まとめと今後の課題

本稿では，分布仮説に基づく単語ベクトル生成手法である Random Indexing を利用した単語の意味変化の追跡・検出手法を提案した。対象としたデータや意味変化を追跡したキーワードは限定的であるが，実験によりその有効性を示した。

単語をベクトルとして表現するモデルにおいては，意味変化の検出はベクトルを要素とした時系列データの変化を検出することを意味する。スカラー値を要素とした時系列データの変化の検出は様々な手法が考案されているが，ベクトル要素の場合に適用できるものは少ない。今回は隣接した二つのベクトルのコサイン尺度のグラフと全ベクトル間のコサイン尺度のヒートマップにより変化を可視化し，人の目により確認したが，今後網羅的かつ自動的な検出実験を行なうためには，人の目によらない数値的な検出手法が必要である。

また，今回対象としたブログ記事に現れる単語の異なり総数は8,511,075であるが，対象キーワードを含まない記事も入れると65,362,090と膨大なものとなる。網羅的な実験を行なうためにはこのスケールに対応できる効率的なものでなければならない。単語の意味変化の継続的で網羅的かつ自動的な追跡と検出を行なうために，手法の空間・時間計算量の解析も進めたい。

（いしかわ・まさひろ　メディア社会学科）

参考文献

（１） 加藤淳一　石川雅弘，大量ブログ記事をデータとした市場セグメンテーションの半自動的分析手順，日本オペレーションズリサーチ学会　2011年春季研究発表会，2011年３月.

（２） 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.

（３） 喜連川優，情報爆発のこれまでとこれから，電子情報通信学会誌 Vol. 94, No. 8, pp 662–666, 2011.

（４） goo ブログ，http://blog.goo.ne.jp/.

（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） 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.

（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） 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, pp 3111–3119, Curran Associates, Inc. 2013.

（９） MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.

（10） Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval,* Addison Wesley, 1999.

An experiment to analyze word sense changes based on the distributional hypothesis, using blog texts generated before and after the 3.11 accident.

Masahiro Ishikawa

We live in the era of information explosion. There is a massive amount of text on the Web. Especially in social media, huge amount of text is generated by users day by day. There are many studies trying to use them for sentiment analysis, market analysis, and so on. In such analysis, treatment of text meaning is essential. Text is composed of words, thus word sense treatment is essential. However, word meanings undergo changes. A word can acquire new word senses, or become obsolete. Thus, we should consider word sense changes over time. In this paper, we propose a method to detect word sense changes over time. The proposed method uses Random Indexing technique, which is based on the distributional hypothesis of word meanings. The result of the first experiment on timestamped blog texts is also presented.

Key words: Text Mining, Blog, Distributional Hypothesis, Random Indexing