2E4-NFC-01b-5

テキストから得られる複数特徴量を融合する絵本類似探索法

Picture Book Similarity Search by Fusing Multiple Linguistic Features

服部 正嗣 *1　藤田 早苗 *1　青山 一生 *1
Takashi Hattori　Sanae Fujita　Kazuo Aoyama

*1NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

Searching for books similar to a favorite book may be helpful on choosing picture books for a child. In this paper, we propose a similarity search method for picture books which finds similar books to a favorite book; we carry out searches based on multiple linguistic features, and fuse the results. Experimental results show that the picture books which appear in the fused result tend to be ranked high at each result regarding the corresponding feature.

1. はじめに

絵本の読み聞かせは幼児の言語発達を促進することが知られている [1]．より高い効果を得るためには幼児に合った絵本を選ぶことが重要と考えられる．本稿では幼児の気に入った絵本に類似した絵本は幼児に関心高く受け入られやすいという仮説の下に，就学年齢程度までの絵本を対象に，入力した絵本に類似した絵本を出力する類似絵本探索法を提案する．

絵本の類似性を定義する際，テキストのみに着目しても，語彙や一文あたりの長さなどの複数種類の特徴を抽出可能である．複数の中から単一種類の特徴のみを選ぶ場合，最もユーザの直感にあった類似の絵本を探索できる特徴の選択は自明ではない．また，多くの絵本について適切だった種類の特徴が例外的な絵本には適さない場合も考えられる．例えば，一般に絵本に登場する語彙は絵本の対象年齢が就学年齢に近づくほど難しいものになる傾向があるため，語彙に基づく類似性を定義すれば入力絵本と近い年齢層向けの絵本の探索が期待できる．しかし，赤ちゃん向けのように使用語彙の難易度が幅広く，保育者が文章を読むことを前提としている絵本の中には比較的難しい語彙が含まれる場合もあり，このような絵本と類似する絵本を探索する場合には語彙に基づく類似性は適切ではない．

一方で，複数の特徴を組み合わせて探索を行うことにより個々の特徴の得手不得手を克服するアプローチも提案されている [2, 3, 4]．絵本のように種々の観点を持つ複雑なオブジェクトは，単一種類の特徴より複数種類の特徴を用いることによって，より適切に表現されると考えられる [4]．本稿では，テキストから抽出された複数の特徴を併用する類似探索を行う方法について考察する．特徴段階で融合した後に類似探索を行う方法および個々の特徴で類似探索を行った結果を融合する方法をそれぞれ絵本に適用し，その効果について考察する．

以下，2 節において対象データと，絵本から類似探索に特徴ベクトルを抽出する方法を述べる．3 節ではグラフ索引型類似探索法について述べた後，複数の特徴を用いてグラフ索引型類似探索を行う際の特徴の融合方法について考察する．4 節において提案法について述べ，5 節で絵本データベースに提案法を適用した評価実験とその考察について述べる．最後に 6 節で結論を述べる．

連絡先: 服部 正嗣，NTT コミュニケーション科学基礎研究所，〒 619-0237，Tel:0774-93-5335，Fax:0774-93-5155，hattori.takashi@lab.ntt.co.jp

2. 対象データと特徴ベクトル作成

紀伊国屋書店グループ全体で販売された絵本のうち，2010 年度および 2011 年度に売り上げ上位であったものから 835 冊を選び，絵本データベースとする．以下，絵本データベースの各絵本から 4 種の特徴を抽出して，それぞれについての特徴ベクトルを得る方法について述べる．

2.1 テキストからの特徴抽出

絵本データベース中の各絵本から（1）形態素の原形，（2）形態素の品詞大分類，（3）原形が所属する意味クラス，および（4）書誌情報の 4 種類の特徴を抽出する．

絵本一冊ごとにテキストの形態素解析を行う．形態素解析には，京都テキスト解析ツールキット KyTea[5] を用いる．解析モデルは絵本用に構築したものを用いる [6]．ただし，品詞体系は IPA 品詞体系を用いる．次に，形態素の原形とその品詞大分類を抽出し，それぞれの出現回数をその絵本における *tf*（text frequency）値として記録する．形態素が名詞であった場合，その名詞が所属する日本語語彙大系 [7] の意味クラスを抽出する．意味クラスは 2,710 カテゴリのシソーラスで定義されている．このシソーラスの上位 4 レベルを図 1 に示す．このシソーラスは 12 レベルまでの深さを持つ非平衡型階層構造である．レベル 1 は《1:**名詞**》であり，レベル 12 は《1961:**農作業**》，《1993:**出演**》などを含む．名詞の中には複数の意味クラスに所属し文脈によって異なる役割を果たすものも存在する．この種の名詞を扱う際には文脈を考慮した曖昧性除去は行わず，ある名詞が n 種の意味クラスに所属する場合はそれぞれの意味クラスが $1/n$ 回出現したものとみなす．出現した意味クラスに上位意味クラスが存在する場合は，その上位意味クラスも同じ回数分だけ出現したものとみなす．再帰的に出現回数記録し，すべての名詞について意味クラス《1:**名詞**》に達するまで階層構造を上方に移動し，経由した意味クラスの出現回数を加算する．最終的に記録された各意味クラスの出現回数をその絵本における各意味クラスの *tf* 値と定義する．また，絵本の書誌情報として，著者，挿絵画家，訳者などの人名および出版社名をそれぞれの *tf* 値を 1 として記録する．更に，絵本データベース中に出現した形態素の原形の種類数 N_{term} を記録し，N_{term} 種類の原形それぞれについて *idf*(inverse document frequency) 値を算出し記録する．品詞大分類，書誌情報についても同様にそれぞれの種類数 N_{pos}, N_{bib} と品詞大分類，書誌情報一種ずつの *idf* 値を算出し記録す

る．意味クラスについては種類数 N_{sc} を同様に記録する．*idf* 値については，1 に満たない値であっても非零の *tf* 値を持つ意味クラスはその絵本に出現したものとみなして，その意味クラスの *idf* 値を算出する．

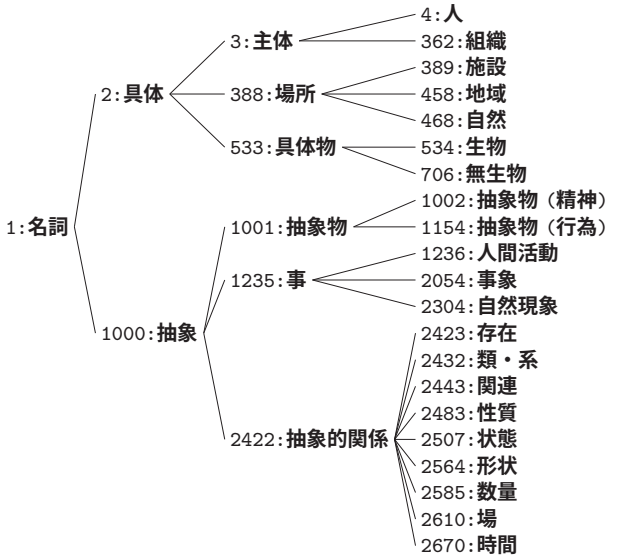

図 1: 日本語語彙大系の上位 4 階層 (一般名詞シソーラス)

2.2 特徴ベクトル

絵本データベース中の i 冊目の絵本の形態素の原形についての特徴ベクトルを，下記のように定義する．

$$\mathbf{F}_{term}(i) \equiv (term_1, \cdots, term_k, \cdots, term_{N_{term}}). \quad (1)$$

ここで，k 番目の要素 $term_k$ は，絵本データベース中 k 種類目に現れた形態素の原形についての i 番目の絵本の *tfidf* 値である．同様の方法で i 冊目の絵本の品詞大分類についての特徴ベクトル $\mathbf{F}_{pos}$，意味クラスについての特徴ベクトル $\mathbf{F}_{sc}$，書誌情報についての特徴ベクトル $\mathbf{F}_{bib}$ を定義する．

3. 類似探索

複数の特徴を併用して大規模データから類似探索を行う方法は高速性に優れていることが望ましい．この要件を満たす探索法の 1 つに，グラフを索引する類似探索法がある [8]．また，複数の特徴の各々に対する重要性を探索実行時にユーザが自由に決定できることは，ユーザの利便性向上の観点から望まれる．これら 2 要件を考慮し，ここではグラフを索引とする高速な類似探索法について述べ，その手法に適した特徴の融合方法について検討する．

3.1 グラフ索引型類似探索法

社会学やネットワーク科学の分野において研究されてきた，多くの現実のネットワークに観られる普遍的な性質の 1 つにスモールワールド性がある．スモールワールド性を有するネットワークは，任意の 2 頂点間の平均最短パス長は非常に小さい．更に，そのネットワークの中には decentralized algorithm が 2 頂点間に存在するパス長の短い経路を発見できるナビゲーション機能を有するものがあることが知られている [9, 10]．これらの性質はネットワークにおける効率的な探索を可能にする．

この探索上の特長に着目し，スモールワールド性を探索に応用する類似探索法が研究されている [8, 12]．これらの方法では，スモールワールド性を有し任意のオブジェクト間の平均最短パス長が小さいネットワーク構造を，与えられたオブジェクト集合（頂点集合）と類似度定義に基づいて人工的に構築する．構築されたネットワーク構造は類似探索時に索引として用いられるためグラフ索引と呼ばれる．本稿では，グラフ索引を用いる類似探索法を，グラフ索引型類似探索法と呼ぶ．グラフ索引型類似探索法は，オブジェクト集合と類似度定義からグラフ索引を作成する前処理段階（オフライン）と入力されたクエリに類似するオブジェクトをグラフ索引を用いて高速に探索・出力する探索段階（オンライン）の 2 段階からなる．具体的なグラフ索引としては，各オブジェクトを頂点としその最類似の k 個の頂点を連結した *k-NN*(Nearest Neighbor) グラフを用いる方法が提案されている [8, 11]．また，更なる高速化のために，*k-NN* グラフの次数を低減した degree-reduced *k-NN*（*k-DR*）グラフを用いる方法も提案されている [8, 12]．

3.2 複数特徴の同時利用

複数種類の特徴を用いた探索法としては，探索前の類似度を融合する方法や，特徴ごとに類似探索を行った後にそれぞれの結果を融合する方法などが挙げられる [2]．この 2 つのアプローチについて概説し，グラフ索引型類似探索との相性について考察する．

3.2.1 類似度段階での融合

複数種類の特徴の各々が構成する特徴空間に定義される類似度を融合することによって，オブジェクト間の総合的な類似度を定義するアプローチである．具体的な類似度の融合法としては，類似度の重み付き線形和を用いる方法 [4]，MAX，MIN などの特徴的な値を採用すると行った方法などが挙げられる．ここでは簡易ながら高い効果が期待できる重み付き線形和によって融合する場合を考える．

オブジェクト集合 X，オブジェクト $x, y \in X$ 各々を表現する n 種類の特徴ベクトルの組 $(\mathbf{F}_1^x, \mathbf{F}_2^x, \cdots, \mathbf{F}_n^x)$ と $(\mathbf{F}_1^y, \mathbf{F}_2^y, \cdots, \mathbf{F}_n^y)$ とが与えられ，同種の特徴ベクトルのコサイン類似度を融合して総合的な類似度を定義する場合を考える．特徴ベクトル $\mathbf{F}_i^x$，$\mathbf{F}_i^y$ のコサイン類似度を $S_i(x,y)$ と表すと，$S_i(x,y) = \mathbf{F}_i^x \cdot \mathbf{F}_i^y / \parallel \mathbf{F}_i^x \parallel \cdot \parallel \mathbf{F}_i^y \parallel$ である．このとき，x, y の総合的な類似度 $S(x,y)$ を式 (2) に示す各種類の類似度 $S_i(x,y)$ の重み付き線形和で与える．

$$S(x,y) = \sum_{i=1}^{n} w_i \cdot S_i(x,y) \ , \quad \sum_{i=1}^{n} w_i = 1 \ \text{ and } \ w_i \geq 0 \ . \quad (2)$$

本稿の場合，特徴ベクトルは x, y 各々について $\mathbf{F}_{term}$，$\mathbf{F}_{pos}$，$\mathbf{F}_{sc}$，$\mathbf{F}_{bib}$ の 4 種類あるので，2 つの絵本 x, y の類似度 $S(x,y)$ は，原形の類似度 $S_{term}(x,y)$，品詞大分類の類似度 $S_{pos}(x,y)$，意味クラスの類似度 $S_{sc}(x,y)$，および書誌情報の類似度 $S_{bib}(x,y)$ に各々重み w_{term}，w_{pos}, w_{sc}, w_{bib} を適用した積和演算結果である．

この類似度を融合する方法を，事前に構築した索引を用いる探索法に適用することは，次の理由のため困難である．前述の「各特徴の重要性を探索実行時にユーザが自由に決定できる」という要件を満たすことが難しい．グラフ索引型類似探索法の場合，類似度を融合する度にグラフを構築することになるため，探索実行の即時性が失われる．

3.2.2 結果段階での融合

複数の特徴のそれぞれを用いて独立の類似探索を行った結果を融合し，最終的な類似探索結果を得るアプローチであり，reranking systems や search reranking とも呼ばれる [13, 14]．Search reranking は探索結果が得られた後に融合を行う為，

ユーザが探索実行時に自由に変更しても探索の即時性は失われない．一般的な search ranking は，個別の特徴と類似度を用いて得られた探索結果リスト（初期探索ランク）を，各リストのオブジェクトやそのオブジェクトに付与されたスコア [15] に基づいて，再ランク付けし 1 つの探索結果リストを作成する．また，最近が初期探索リストを基に各々からグラフを構築し，そのグラフを融合する方法も提案されている [3]．

初期探索ランクをグラフで融合する方法 [3] は，各特徴に対し異なる探索アルゴリズムを用いて探索を行う．グラフ索引型類似探索法は任意の特徴と類似度定義に適用可能であるという特長を有するため，各特徴に対する探索に適用できる．更に，結果をグラフで出力することも容易であるため，各初期探索ランクと結果のグラフとを同時に得ることができる．4 節で述べる提案法はこの方法を採用する．グラフの融合は，各特徴で用いた類似度定義とは別に新たに距離を定義し，各グラフの和集合を作成することで行われ，reranking は融合されたグラフに PageRank[16] を適用することで行う．

4. 提案法

提案法は，複数の特徴で各々類似探索した結果を融合する絵本類似探索法である．まず，オブジェクト集合は，絵本集合とし，それぞれの絵本から 2 節で述べた 4 種の特徴ベクトルを抽出する．絵本間の類似度尺度としては，特徴ベクトルのコサイン類似度を用いる．類似探索法としてはグラフ索引型類似探索を採用し，グラフ索引構築段階では絵本集合と 4 種の特徴ベクトルから 4 つのグラフ索引を構築する．探索段階では，クエリとして入力された絵本から 4 つの特徴ベクトルを抽出し，それぞれの特徴ベクトル用のグラフ索引を用いて，4 つの探索結果を得る．その後，[3] の方法で 4 つの探索結果を融合し，最終的な探索結果を出力する．

5. 実験

5.1 評価方法

2 節の絵本データベース中の絵本 835 冊それぞれをクエリとして，提案法で絵本データベースを対象に類似探索を行った場合，探索結果にどのような傾向があるかを調査することで提案法の性能を定性評価する．グラフ索引として k-DR グラフ（k=12）を用い，クエリを一冊入力するごとに 4 種の特徴ごとにクエリ自身を除くコサイン類似度上位 100 冊の絵本をそれぞれ探索した．次に [3] の方法で各特徴の探索結果を融合し，最終的な類似探索結果として各絵本ごとに上位 5 冊ずつの類似絵本を得た．

5.2 結果と考察

最終的に出力された 5 冊の類似絵本が，融合前の単独の特徴による探索結果の複数で上位にランクされているかどうかを調べた．表 1 に各絵本の最終的な類似絵本上位 5 冊について，4 種の特徴による単独の探索結果の内，上位 100 冊以内であった特徴数を示す．特徴 1 種あるいは 2 種で上位 100 冊以内にランクされた類似絵本の割合は合わせて 7 割に達する．一方，特徴 4 種すべてにおいて上位 100 冊以内であった類似絵本の割合は 9.1%と比較的少ない．

表 2 に，類似絵本が上位 100 冊以内であった特徴数ごとに，最終順位の割合を示す．これによれば，上位 100 冊以内であった特徴数が増えれば増えるほど，最終的に上位となる傾向があることが分かる．表 1 の結果と合わせて考えると，多数の特徴すべてについて類似している絵本は，存在する冊数は少ないが存在した場合にはより上位で選ばれやすいと考えられる．

表 1: 類似絵本が上位 100 冊以内の特徴数

	特徴 1 種	特徴 2 種	特徴 3 種	特徴 4 種
1 位	210	265	215	145
2 位	292	264	178	101
3 位	288	326	159	62
4 位	309	311	173	42
5 位	323	336	146	30
計	1422 (34.0%)	1502 (36.0%)	871 (20.9%)	380 (9.1%)

表 2: 上位 100 冊以内の特徴数と最終順位割合

	1 位	2 位	3 位	4 位	5 位
特徴 1 種	14.8%	20.6%	20.2%	21.8%	22.7%
特徴 2 種	17.6%	17.6%	21.8%	20.7%	22.3%
特徴 3 種	24.8%	20.4%	18.2%	19.8%	16.7%
特徴 4 種	37.8%	27.1%	16.4%	11.0%	08.8%

図 2: 「あなたがだいすき」リザ ベイカー（著），デイビッド マクフェイル（画），日当陽子（訳），フレーベル館，2011 年

図 3: 単一特徴の探索結果と融合後の結果

表 3: 表紙を引用した絵本の出典

タイトル	著者・挿絵画家等	出版社	出版年
あなたがだいすき	鈴木まもる	ポプラ社	2002
たいせつなあなたへ あなたがうまれるまでのこと	サンドラ・ポワロ＝シェリフ（著・画），おーなり由子（訳）	講談社	2010
だいすき！ぎゅっ	秋田喜代美（監修）	ベネッセコーポレーション	2010
ふわふわだあれ？	いりやまさとし	学研教育出版	2008
いのちのカプセル まゆ	新開孝	ポプラ社	2008
もったいないばあさん もりへいく	真珠まりこ	講談社	2011
ちびゴリラのちびちび	ルース・ボーンスタイン（著・画），いわたみみ（訳）	ほるぷ出版	1978
ずーっとずっとだいすきだよ	ハンス・ウィルヘルム（著・画），久山太市（訳）	評論社	1988
アンパンマンとげんきにあいさつ	やなせたかし（原作），トムス・エンタテインメント（作画）	フレーベル館	2005
しろかぶくんとアンパンマン	やなせたかし（原作），トムス・エンタテインメント（作画）	フレーベル館	2011
できるといいねはみがき	やなせたかし（著），東京ムービー（著）	フレーベル館	2000

具体的な例として図 2 に示す絵本をクエリとした際の単一特徴それぞれの探索結果と，融合後の探索結果上位 3 冊の表紙を図 3 を示す．また，図 3 中の表紙の出典を表 3 に示す．このクエリ絵本では，母から子へ度々「あなたがすき」という言葉が発せられる．そのため，形態素の原形の特徴のみで行った類似探索では,「あなた」「すき」が多く現れる絵本が上位にランクインした．この例では，形態素の原形で上位 1, 2, 4, 5 位にランクインした絵本が品詞大分類や意味クラスでも 100 位以内であったため，そのまま融合後の上位にランクインした．形態素の原形で 3 位にランクインした絵本は「あなた」のみが多く現れる本であり，他の種類の特徴単独の探索結果では順位が 100 位未満であったため，融合結果には残らなかった．融合後 3 位の絵本「だいすき！ぎゅっ」は，意味クラスでは 1 位であり，意味クラス《1300:**愛好**》が多く現れる．この意味クラスにはクエリ絵本に多く現れる「すき」が所属する．「だいすき！ぎゅっ」は原形による探索結果でも 59 位に入っているため，融合後にランクインしたものと思われる．一方，書誌情報で上位にランクインしたものは他の特徴の上位 100 冊以内に表れるものが 1 冊もなく，融合結果にも現れなかった．なお，融合結果の 4, 5 位はそれぞれ「あなたが生まれるまで（ジェニファー・デイビス（著），ローラ・コーネル（画），槙朝子（訳），小学館，1999 年）」「ちいさなあなたへ（アリスン・マギー（著），ピーター・レイノルズ（画），なかがわちひろ（訳），主婦の友社，2008 年）」であった．クエリとして与えた絵本と同じく親子の絆をテーマにした絵本が上位を占めたといえる．以上のように，この例では，複数の特徴それぞれによる探索結果を融合することで，どの単独の特徴のみでも得られなかった同一テーマの絵本 5 冊を得ることができた．

6. 終わりに

本稿では，テキストから抽出した複数種類の特徴を融合してグラフ索引型類似探索を行う方法について検討した．特徴の融合にあたり，探索前の類似度の段階で融合する方法と探索後の結果の段階で融合する方法とを比較した．前者は各特徴の重みの変更を行うにあたりグラフ索引の再構築が必要となるため，より容易に重み付けを行える後者を採用した．実際に絵本データベースについて，形態素の原形，品詞大分類，意味クラス，書誌情報の 4 種類の特徴を融合する類似探索の評価実験を行った．結果，複数の特徴において上位となる絵本は総数としては少ないものの，単一の特徴においてのみ最上位となる絵本よりも融合後の結果に反映される傾向があり，複数の特徴による複合的な評価に繋がる可能性を確認した．今後は特徴間の重み付けの方法に関する検討が必要である．また，TETDM[17] のようなテキストデータマイニングツールを利用し，テキストから多様な特徴を抽出して利用するとともに，挿絵の画像特徴といったテキスト由来以外の特徴の利用についても拡張したい．

参考文献

[1] Whitehurst, G. J., Falco, F. L., Lningan, C. J., Fischel, J. E., DeBaryshe, B.D., Valdez-Menchaca, M. C. and Caulfield, M.: Acceleratаing language development through picture book reading, Developmental Psychology, 24(4), pp. 552-559, (1988).

[2] Atrey, P. K., Hossain, M. A., El Saddik, A., Kankanhalli, M. S.: Multimodal fusion for multimedia analysis: a survey, Multimedia Systems, Vol. 16, pp. 345-379, (2010).

[3] Zhang, S., Yang, M., Cour, T., Yu, K., Metaxas, D.: Query Specific Rank Fusion for Image Retrieval, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol. 37, No. 4, pp. 803 - 815, (2015).

[4] 服部 正嗣，青山 一生: グラフ索引を用いた絵本の類似探索?特徴の融合と結果のグラフ可視化?，情報処理学会研究会 第 10 回ネットワーク生態学シンポジウム，(2013).

[5] 森信介，中田陽介，Neubig, G.，河原達也: 点予測による形態素解析，自然言語処理，Vol. 18, No. 4, pp. 367-381, (2011).

[6] 藤田 早苗，平 博順，小林 哲生，田中 貴秋: 絵本のテキストを対象とした形態素解析，自然言語処理，Vol. 21, No. 3, pp. 515-539, (2014).

[7] 池原 悟，宮崎 雅弘，白井 諭，横尾 昭男，中岩 浩巳，小倉 健太郎，大山 芳史，林 良彦: 日本語語彙大系，岩波書店，(1997).

[8] Aoyama, K., Saito, K., Yamada, T. and Ueda, N.: Fast similarity search in small-world networks, Complex Networks: Int. Workshop on Complex Networks, pages 185-196. Springer(2009).

[9] Watts, D. J., Strogatz, S. H.: Collective dynamics of 'small-world' networks, Nature, Vol. 393, No. 6684, pp. 409-10, (1998).

[10] Kleinberg, J. M.: Navigation in a small world, Nature, Vol. 406, No. 6798, pp. 845,(2000).

[11] Wang, J., Li, S.: Query-driven iterated neighborhood graph search for large scale indexing, Proc. ACM Int. Conf. Multimedia, pp. 179-188, (2012).

[12] Aoyama, K., Saito, K., Sawada, H. and Ueda, N.: Fast approximate similarity search based on degree-reduced neighborhood graphs, Int. conference on Knowledge Discovery and Data Mining, (2011).

[13] Mei, T., Rui, Y., Li, S., and Tian, Q.: Multimedial search reranking: A literature survey, ACM Computing Surveys, Vol. 46, No. 3, article 38, (2014).

[14] Wang, M., Li, H., Tao, D., Lu. K. and Wu. X: Multimodal graph-based reranking for Web image search, IEEE Trans. Image Process., Vol. 21, No. 11, pp. 4649-4661, (2012).

[15] Hazen, T. J., Shen, W. and White, C.: Query-by-example spoken term detection using phonetic posteriorgram templates, Int. Workshop on Acoustic Speech Recognition & Understanding, pp. 421-426, (2009).

[16] Langville, A. N. and Meyer, C. D.: Google's PageRank and Beyond: The science of search engine rankings, Princeton University Press, (2006).

[17] Total Environment for Text Data Mining (TETDM), http://tetdm.jp/, (2015-03-23).