社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (10)
2009－ＮＬ－189 (10)
2009／1／23

グラフを用いたバイオ医療専門用語の類義語獲得

鈴木 郁美　　原 一夫　　新保 仁　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
{ikumi-s, kazuo-h, shimbo, matsu}@is.naist.jp

コーパスから抽出した文脈情報により作成する専門用語グラフに対し，グラフを辿ることで節点間の類似度を計算する手法を適用し，類義語獲得に応用した．雑誌「蛋白質・核酸・酵素」をコーパスとして用いた実験で，コーパスでの出現頻度が少ない専門用語をクエリとして与えた場合，ラプラシアン拡散カーネル行列を用いた手法が比較的高い精度を示した．この結果は，専門性の高いレアな用語を既存のシソーラスに登録する場面において，ラプラシアン行列ベースの手法の有効性を示唆するものである．

Synonym Acquisition for Biomedical Thesaurus Expansion: a Graph-based Approach

Ikumi Suzuki　　Kazuo Hara　　Masashi Shimbo　　Yuji Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma Nara 630-0192 Japan
{ikumi-s, kazuo-h, shimbo, matsu}@is.naist.jp

We apply graph-based methods to problems of biomedical synonym acquisition. Given a graph of biomedical terms constructed from a corpus, the methods calculate term similarities by traversing the graph to capture shared features between nodes. An experimental study shows that, for query terms appearing less than three times in the corpus, the Laplacian diffusion kernel gives better accuracy than the methods based on the adjacency matrix.

1 はじめに

近年，バイオテクノロジーの目覚ましい進展にともない，分子生物学分野における新たな知見の報告とそれに基づく新規医療の開発等が，世界中の研究者により急ピッチで進められている．一方，分野（e.g., 基礎医学，臨床医学，薬学，生物学，解剖学，etc.）ごとに高度に専門化されるそれぞれの知見について，バイオサイエンス全体での位置付けを把握するためには，分野横断的なデータベースの構築（たとえば，米国国立医学図書館による MEDLINE [1]，文部科学省による統合データベースプロジェクト[2] など）とともに，各分野に登場する専門用語間の同義関係（あるいは，上位、下位関係）を記述するシソーラス辞書（e.g., 米国国立医学図書館による MeSH [3]，京都大学の金子周司氏らによるライフサイエンス辞書[4]）が役に立つ．とくに，シソーラス辞書は，専門外の人にとって馴染みのない専門用語が頻出する各分野の研究報告を理解する上で欠かせない．

本稿の目的は，最新の技術・知見の成果報告等に関連して日々新しく登場する専門用語について，既存のシソーラス辞書への新規登録（i.e. シソーラス拡張）を支援するシステムの設計である．具体的には，大規模なバイオ文献テキスト（以下，コーパス）を用い，まず，(1) コーパス中に現れる専門用語を節点とするグラフを作成する．次に，(2) グラフ構造を用いて専門用語間の類似度を算出する．本稿では，ここに隣接行列ベースの手法とラプラシア

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://lifesciencedb.mext.go.jp/
[3] http://www.nlm.nih.gov/mesh/meshhome.html
[4] http://lsd.pharm.kyoto-u.ac.jp/ja/

ン行列ベースの手法を適用する．そして，(3) 新規登録対象の専門用語（クエリ）に対し，類似度が高い順番に登録済の専門用語をランク付けし，提示する（i.e. 類義語検索または類義語獲得）．シソーラス辞書の編集者は，ランキング結果を参考にして，新しい専門用語をシソーラス辞書にマッピングすることができる（たとえば，ランク1位の専門用語の近傍に配置する），と我々は考えている．

類義語獲得のこれまでの研究は，コーパスから文脈を抽出して語の特徴ベクトルを作り，コサイン類似度により語と語の類似度を測る方法が主である [13, 14, 12]．しかし，この方法は，共通する文脈が存在しないとき，類似度を常にゼロにする性質がある．とくに，コーパスでの出現頻度が少ない語（たとえば専門性の高い用語）は，コーパスから十分な情報の文脈を抽出できない事態が想定される．このとき，本来は類似しているはずの語とも文脈を共有しにくくなるため，類似度が低く計算されてしまう可能性がある．

このことを解決するために，本稿では語を節点とするグラフを作成し，グラフを辿って節点から節点へ至る経路を考慮して語の類似度を計算する方法を用いる．グラフ上の経路を考慮に入れるこの方法は，商品推薦システム構築，語義曖昧性解消などでその有効性が確かめられており [3, 7]，類義語獲得の精度向上にも寄与することが期待できる．

なお，本稿では，コーパスからの専門用語抽出については言及しない（すなわち，何をもって専門用語とするかについては議論しない）．ここでの焦点は，クエリとして与えられる専門用語に対し，類義の専門用語をランキング形式で提示するシステムを設計することである．

本稿の構成は以下の通りである．次節で類義語獲得の先行研究を紹介したのち，グラフを用いた専門用語間類似度の算出方法について，3節で具体的に述べる．4節で，コーパスとして「蛋白質・核酸・酵素」[5]，シソーラスとしてライフサイエンス辞書を使用して行った，バイオ医療専門用語の類義語獲得実験について述べる．実験結果の検証は，専門性の高いレアな用語を既存のシソーラスに登録する場面を想定し，コーパスでの出現頻度が少ない専門用語をクエリとする場合について，主に行う．最後に今後の方向について議論する．

[5]http://www.kyoritsu-pub.co.jp/pne/

2 類義語獲得の関連研究

類義語獲得の関連研究には，寺田ら [13]，萩原ら [14]，王ら [12] がある．これらの研究は，前後文脈が似ている語は意味が類似しやすいという仮定のもと，コーパスから文脈情報を抽出し，それらの頻度を要素とする特徴ベクトルを各語について作成し，コサイン類似度により語の類似度を測る部分で，共通している．寺田らは，文脈情報として前後に隣接する語を用い，window 幅を 2 とし，特徴ベクトルの要素を log(頻度+1) とする場合に最も高い精度を得ている[6]．萩原らは，依存関係を文脈として用い，直接の依存関係だけでなく 2 つの依存関係の合成まで特徴ベクトルに加えることが有効であると報告している．また，王らは，クエリ $term_q$ に対して得られるランキング結果を，そこにランクインしている語 $term_r$ をクエリにして得られるランキングにおける $term_q$ のランクを用いて，リランキングする手法を提案している．

Hagiwara [4] は，語に対してではなく，語のペアに対して特徴ベクトルを作成し，語のペアに関して，類似している／していない，の 2 値分類問題を SVM で解き，SVM の返す値（分離平面からの距離）による類義語ランキングを試みている．

Blondel et al. [1] は，(新聞記事や学術雑誌ではなく）辞書をコーパスとして用い，見出し語からその語釈文内に現れる語へのリンクを張ることにより語のグラフを作成し，Kleinberg の HITS [6] を一般化した方法を類義語獲得に適用している．

3 方法

3.1 専門用語グラフの作成

コーパス中に N 種類の専門用語が出現するとして，まず，各専門用語 $term_n$ $(n = 1, \ldots, N)$ に対して特徴ベクトル v_n を作成する．特徴ベクトル作成のための文脈情報としては，寺田ら [13]，萩原ら [14]，王ら [12] にならい，コーパスから抽出する隣接語，依存関係（係り受け）を用いる．さらに，医学用語の分類には接尾辞が有効との報告 [11] があり，接尾辞も使用する．

専門用語グラフは，専門用語を節点とし，節点 i と節点 j を結ぶ枝の重みを $a_{ij} = v_i \cdot v_j$（内積）として与え，作成する．ここで，a_{ij} を (i, j)-要素してもつ行列 A は，グラフの**隣接行列**と呼ばれるものである．

[6]4 節の実験でもこの設定を用いる．

3.2 グラフを用いた専門用語間類似度の算出

3.2.1 隣接行列ベースの手法

グラフの隣接行列は，節点間の類似度としてそのまま用いることができる．実際，$term_n$ の特徴ベクトル v_n を長さ 1 に規格化した u_n を用いて隣接行列 Adj を作ると，節点 i と節点 j のコサイン類似度は，$u_i \cdot u_j$ であり，Adj の (i,j)-要素となる．しかし，この場合，直接リンクのない節点間の類似度は常にゼロとなり，たとえ他の節点を介して間接的につながりがあったとしても，そのことは類似度計算に反映しない．

そこで，グラフの枝を辿ることで節点間の類似度を計算する方法が提案されている．そこでは，まず，節点 i から 2 本の枝（セルフ・ループを含む）を経由して節点 j に至る場合の，節点 i と節点 j の類似スコアを定義する．経路の数は，その途上で通過する（1 つの）節点 k の種類の数に等しく，N である．節点 k （$k = 1, \ldots, N$）を通る経路のスコアを，隣接行列 A （の要素）を用いて $a_{ik} \times a_{kj}$ と定義すると，N 個の経路に関するスコアの合計は，$\sum_{k=1}^{N} a_{ik} \times a_{kj}$ であり，行列 A^2 の (i,j)-要素となる．これをもって，2 本の枝を経由する場合の節点 i と節点 j の類似スコアと定義する．同様にして，M 本の枝を経由する場合の節点 i と節点 j の類似スコアを定義でき，行列 A^M の (i,j)-要素に対応する．以上の類似スコアを（パラメタ β による重みを付けて）足し合わせた**フォンノイマン拡散カーネル行列 (VN):**

$$\begin{aligned} VN &= A(I - \beta A)^{-1} \\ &= A + \beta A^2 + \beta^2 A^3 + \cdots \end{aligned}$$

の (i,j)-要素を，節点 i と節点 j の類似度として使用することが提案されている [5]（I は単位行列）[7].

3.2.2 ラプラシアン行列ベースの手法

以上は隣接行列を用いる類似度計算であるが，他方，ラプラシアン行列

$$L = D - A$$

を用いる方法が提案されている [2]．ここで，D は，その対角要素が A の各行の要素の和となるような，対角行列である．ラプラシアン行列の各要素の符号を反転したもの（$-L$）は，非対角要素は隣接行列 A と一致するが，対角要素は隣接行列とは異なりマイナスの値となる．このことは，ラプラシアン行列を用いて経路のスコアを定義する場合，セルフ・ループを通る経路はスコアが小さくなることを意味する．とくに，多数の枝をもつ節点（ここで扱う専門用語グラフにおいては，広範に使用される専門用語，たとえば，“細胞”，“酵素” などに対応する）については，セルフ・ループのスコアは大きくディスカウントされることになる．したがって，それらの節点との類似度は，隣接行列を用いてグラフを辿る場合よりも概して小さい値となることが見込まれ，よって，一般的な意味をもつ専門用語が類義語ランキングの上位に入るのを防ぐ効果が期待できる．

ラプラシアン行列による類似度計算方法にはバリエーションがあり，よく使用されるものは以下の通りである [3].

正則化ラプラシアンカーネル行列 (RL):

$$\begin{aligned} RL &= (I + \beta L)^{-1} \\ &= I + \beta(-L) + \beta^2(-L)^2 + \beta^3(-L)^3 + \cdots \end{aligned}$$

ラプラシアン拡散カーネル行列 (DF):

$$\begin{aligned} DF &= \exp(-\beta L) \\ &= I + \beta(-L) + \frac{\beta^2}{2!}(-L)^2 + \frac{\beta^3}{3!}(-L)^3 + \cdots \end{aligned}$$

通勤時間カーネル行列 (CM):

$$CM = L^{+} \text{ (pseudo-inverse)}$$

これらの (i,j)-要素を，節点 i と節点 j の類似度として使用することができる．なお，L の代わりに $\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ を用いると性能が高くなるとの報告がある [9]．実際，次節の実験では精度が向上し，以下の実験の DF および CM についての結果は，$\tilde{L}$ を用いたものである．

4 実験

4.1 実験に使用するコーパスとシソーラス辞書

生命科学分野の総合レビュー誌である「蛋白質・核酸・酵素」（以下，PNE）の 1 年分（文の数は 13230）をコーパスとして使用し，シソーラス辞書としてはライフサイエンス辞書（以下，LSD シソーラス）を用いる．LSD シソーラスには，解剖，生物名，病名，物質名を表す 17888 種類の専門用語が登録されており，シソーラスツリーの各ノードに付与されている ID は，1 つの専門用語に対応する．ただし，

[7] 4 節の実験では，隣接行列として A ではなく Adj を用いたフォンノイマン拡散カーネル行列を使用する．

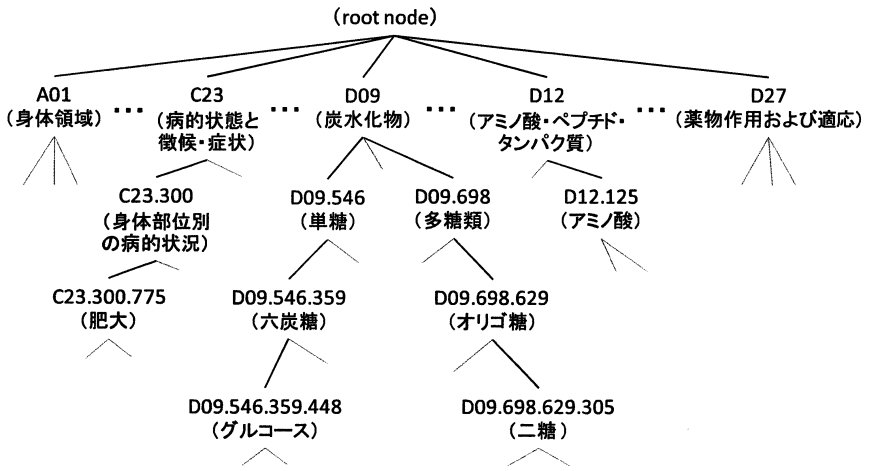

図1:ライフサイエンス辞書(LSD)のシソーラス・ツリーのイメージ. 第1階層(A01, A02, ・・・, D27)のノード数は64, ノードの平均の深さは5.1階層である. 各ノードに付与されるIDは1つの専門用語に対応する. ただし, 1つの専門用語に複数のIDが付与されることがある.

多義性のある専門用語には，複数の ID が付与される．LSD シソーラスのツリーのイメージを図 1 に示す．

PNE からの専門用語および文脈情報（隣接語，係り受け）の切り出しおよび抽出は，ChaSen [8]，CaboCha [9] を用いて行う．専門用語は，LSD シソーラスに登録されている専門用語のリストを ChaSen に与えることで，PNE から切り出す．文脈情報は，助詞，助動詞を除く単語（の原型）を単位としてテキストから抽出する．今回の実験で使用した PNE の 1 年分に登場する専門用語の種類の数は 1164 であり（トークンとしての登場回数はトータルで 17934 回），抽出した隣接語，係り受け，接尾辞の種類の数は，それぞれ 13216, 8229, 710 である．

4.2 評価方法

クエリとして与える専門用語を $term_q$ とし，これに対してシステムが出力するランキング上位 K の専門用語を $term_r$ $(r = 1, \ldots, K)$ とする．Q, R_r をそれぞれ，$term_q$，$term_r$ に対応するシソーラス ID の集合とし，$term_q$ と $term_r$ の LSD シソーラス・ツリーにおける類似度 $Sim(term_q, term_r)$ を次のように定義する [14]:

$$Sim(term_q, term_r) = \max_{id_q \in Q, id_r \in R_r} s(id_q, id_r).$$

[8] http://chasen-legacy.sourceforge.jp/
[9] http://www.chasen.org/ taku/software/cabocha/

表1:システムが出力するランキングに対する評価値(RankAveSim)の計算例: クエリ($term_q$) = "二糖" (Q= {D09.698.629.305})の場合, 以下のテーブルから, RankAveSim(1) = 0.86, RankAveSim(5) = 0.47となる.

ランク r	出力 $term_r$	シソーラスIDの集合 Rr	クエリとの類似度 $Sim(term_q, term_r)$	重み 1/r
1	オリゴ糖	{D09.698.629}	0.86	1
2	単糖	{D09.546}	0.33	0.50
3	アミノ酸	{D12.125}	0.00	0.33
4	肥大	{C23.300.775}	0.00	0.25
5	グルコース	{D09.546.359.448}	0.25	0.20

ここで，シソーラス・ツリーにおける id_q，id_r の深さをそれぞれ d_q，d_r とし，id_q，id_r の共通祖先ノードの深さの最大値を d_{share} として，

$$s(id_q, id_r) = 2 \times \frac{d_{share}}{d_q + d_r}$$

である．$s(id_q, id_r)$ は 0 以上 1 以下の数値となり，id_q と id_r がシソーラス・ツリー上で近くに位置するほど大きい値をとる．

システムの評価は，$Sim(term_q, term_r)$ の順位 r による重み $(\frac{1}{r})$ 付き平均として，

$$RankAveSim(K) = \frac{\sum_{r=1}^{K} \frac{1}{r} Sim(term_q, term_r)}{\sum_{r=1}^{K} \frac{1}{r}}$$

を用いる．計算例を表 1 に示す．

表2: クエリ="二糖"に対するランキング結果. 隣接行列(Adj), フォンノイマン拡散カーネル行列(VN), ラプラシアン拡散カーネル行列(DF)および通勤時間カーネル行列(CM)の比較. カッコ内はクエリとの類似度を表す.

ランク	ランキング作成に使用する行列			
	Adj	VN	DF	CM
1	スフィンゴ糖脂質 (0.25)	スフィンゴ糖脂質 (0.25)	オリゴ糖 (0.86)	単糖 (0.33)
2	チオグルコシド (0.25)	トリニトロベンゼンスルホン酸 (0.00)	単糖 (0.33)	オリゴ糖 (0.86)
3	トリニトロベンゼンスルホン酸 (0.00)	チオグルコシド (0.25)	アミノ酸 (0.00)	シュ糖 (0.89)
4	オリゴ糖 (0.86)	オリゴ糖 (0.86)	肥大 (0.00)	血糖 (0.22)
5	単糖 (0.33)	単糖 (0.33)	グルコース (0.25)	チオグルコシド (0.25)

表3: クエリ="アミン"に対するランキング結果.

ランク	ランキング作成に使用する行列			
	Adj	VN	DF	CM
1	卵白アルブミン (0.00)	卵白アルブミン (0.00)	ポリアミン (0.80)	メチルアミン (0.80)
2	クロラミン (0.33)	クロラミン (0.33)	スペルミン (0.67)	ヘミン (0.00)
3	ヘミン (0.00)	ヒドロキシルアミン (0.80)	ヒストン (0.00)	グルタミン (0.00)
4	ヒドロキシルアミン (0.80)	ヘミン (0.00)	ヘミン (0.00)	卵白アルブミン (0.00)
5	アセチルガラクトサミン (0.00)	アセチルガラクトサミン (0.00)	アセチルグルコサミン (0.00)	スペルミン (0.67)

4.3 結果と考察

PNE に登場する 1164 専門用語のうち，ただ一つの LSD シソーラス ID が付与されている 583 をクエリとして与え，そのそれぞれについて隣接行列 (Adj)，フォンノイマン拡散カーネル行列 (VN)，ラプラシアン拡散カーネル行列 (DF) および通勤時間カーネル行列 (CM) により類義専門用語ランキングを出力させ，結果を比較した[10]．多義性のないクエリのみを評価対象とする理由は，本稿の焦点は語義曖昧性解消ではなく，類義語獲得であることによる．結果の一例として，"二糖"，"アミン" をクエリとする場合のランキング結果を表 2，表 3 に示す．なお，正則化ラプラシアンカーネル行列 (RL) は，ラプラシアン拡散カーネル行列 (DF) と結果がほぼ同等であったため，比較の表では省略している．

使用する行列によるランキング出力傾向の違いを俯瞰するために，表 4，表 5 にクエリと出力結果の類似度 (RankAveSim(1) および RankAveSim(5)) の平均値を示す．ここで，PNE での登場回数 (2 回以下／3 回以上) によってクエリをグループ分け

[10]行列 Adj による結果は，コサイン類似度を用いた場合に相当する．VN と DF についてはパラメタのチューニングを行い，どちらについても $\beta = 0.001$ を用いた．

表4: ランキング精度の比較;クエリとランク1位の類似度 (RankAveSim(1)) の平均値.

クエリ		ランキング作成に使用する行列			
PNEでの登場回数	その種類の数	Adj	VN	DF	CM
2回以下	253	0.228	0.231	**0.255**	0.224
3回以上	330	**0.324**	0.318	0.257	0.289

表5: ランキング精度の比較;クエリとランク1~5位の類似度 (RankAveSim(5)) の平均値.

クエリ		ランキング作成に使用する行列			
PNEでの登場回数	その種類の数	Adj	VN	DF	CM
2回以下	253	0.213	0.214	**0.225**	0.187
3回以上	330	**0.265**	0.262	0.214	0.250

しているのは，登場回数が少ないクエリは，PNE から直接抽出できる文脈情報が乏しいはずであり，グラフを辿って間接的に文脈を獲得し類似度を計算する手法の良さが顕著になるとの予想による．登場回数 2 回以下のクエリに関して，Adj よりも VN，DF の精度が上回ったという結果は，ある程度，この予想を裏付けるものとなっている．

さらに，(頻度 2 以下のクエリでは) 隣接行列ベースの VN よりもラプラシアン行列ベースの DF が高性能を示したことは，(多くの専門用語とリンクをもつため，グラフを辿れば行き着きやすい) 一般的な意味をもつ専門用語に対し，それらを通る経路のスコアを小さく抑えることにより，一般的な専門用語をランキング結果の上位にランクしにくくするという，ラプラシアン行列ベースの手法の性質を反映したものである，と考えることもできる．

とはいえ，今回の実験結果は，ラプラシアン行列ベースの手法の強みを十分に示すものにはなっていない．実際，精度の差は大きくなく，登場回数 2 回以下のクエリであっても，DF が Adj を上回るものが大勢を占めるとはいえない結果となっている (表 6，表 7)．

精度向上に向けた今後の課題はいくつか考えられるが、その一つは，ラプラシアンベースのカーネル行列の最適化である．Zhu et al. [10] が議論しているような，(DF，CM，RL 等に限定しないという意味で) やや広いラプラシアンベースのカーネル行列のクラスを考え，訓練データに対し精度が最大となるものを使用する方法である．また，(ラプラシア

表6: クエリとランク1位の類似度(RankAveSim (1))の大小関係による，隣接行列(Adj)とラプラシアン拡散カーネル行列(DF)の比較.

クエリ		RankAveSim (1) の大小関係		
PNEでの登場回数	その種類の数	Adj > DF となるクエリ数	Adj = DF となるクエリ数	Adj < DF となるクエリ数
2回以下	253	35	168	50
3回以上	330	88	185	57

表7: クエリとランク1~5位の類似度(RankAveSim (5))の大小関係による，隣接行列(Adj)とラプラシアン拡散カーネル行列(DF)の比較.

クエリ		RankAveSim (5) の大小関係		
PNEでの登場回数	その種類の数	Adj > DF となるクエリ数	Adj = DF となるクエリ数	Adj < DF となるクエリ数
2回以下	253	81	79	93
3回以上	330	166	70	94

ン行列ベースの手法の元にもなっている）専門用語グラフの隣接行列の作成方法に立ち返り，グラフの節点間の距離を，訓練データを用いる学習により定めるという方向も考えられる [8].

5 おわりに

本稿では、コーパスから文脈情報を抽出して専門用語グラフを作成し，グラフ構造を用いて専門用語間の類似度を算出することを行い，類義語獲得に応用した．雑誌「蛋白質・核酸・酵素」をコーパスとし，そこに登場する専門用語をライフサイエンス辞書のシソーラスにマッピングすることを目的とする実験において，コーパスでの出現頻度が少ない専門用語をクエリとして与えた場合，ラプラシアン拡散カーネル行列を用いた手法が高い精度を示した．この結果は，専門性の高いレアな用語を既存のシソーラスに登録する場面において，ラプラシアン行列ベースの手法の有効性を示唆するものである．訓練データの学習による類似度行列の最適化が今後の課題である．

謝辞

本研究は，文部科学省 統合データベースプロジェクトから支援を受けて行われた．京都大学の金子周司氏には，ライフサイエンス辞書のシソーラスを提供していただいた．記して深く感謝する．

参考文献

[1] Blondel, V. D., Gajardo, A. I., Heymans, M., Paul and Dooren, V.: A measure of similarity between graph vertices: Applications to synonym extraction and web searching, *SIAM Review*, Vol. 46, pp. 647–666 (2004).

[2] Chung, F. R. K.: *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92) (Cbms Regional Conference Series in Mathematics)*, American Mathematical Society (1997).

[3] Fouss, F., Yen, L., Pirotte, A. and Saerens, M.: An Experimental Investigation of Graph Kernels on a Collaborative Recommendation Task, *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, Washington, DC, USA, IEEE Computer Society, pp. 863–868 (2006).

[4] Hagiwara, M.: A Supervised Learning Approach to Automatic Synonym Identification Based on Distributional Features, *Proceedings of the ACL-08: HLT Student Research Workshop*, Columbus, Ohio, Association for Computational Linguistics, pp. 1–6 (2008).

[5] Kandola, J., Shawe-Taylor, J. and Cristianini, N.: Learning Semantic Similarity, *Advances in Neural Information Processing Systems 15* (S. Becker, S. T. and Obermayer, K.(eds.)), MIT Press, Cambridge, MA, pp. 657–664 (2003).

[6] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol. 46, No. 5, pp. 604–632 (1999).

[7] Komachi, M., Kudo, T., Shimbo, M. and Matsumoto, Y.: Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, Association for Computational Linguistics, pp. 1010–1019 (2008).

[8] Shimizu, N., Hagiwara, M., Ogawa, Y., Toyama, K. and Nakagawa, H.: Metric Learning for Synonym Acquisition, *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, Coling 2008 Organizing Committee, pp. 793–800 (2008).

[9] Zhou, D., Bousquet, O., Lal, T. N., Weston, J. and Schölkopf, B.: Learning with Local and Global Consistency, *Advances in Neural Information Processing Systems 16* (Thrun, S., Saul, L. and Schölkopf, B.(eds.)), MIT Press, Cambridge, MA (2004).

[10] Zhu, X., Kandola, J., Ghahramani, Z. and Lafferty, J.: Nonparametric Transforms of Graph Kernels for Semi-Supervised Learning, *Advances in Neural Information Processing Systems 17* (Saul, L. K., Weiss, Y. and Bottou, L.(eds.)), MIT Press, Cambridge, MA, pp. 1641–1648 (2005).

[11] 山田寛康, 新保仁, 松本裕治: 文脈情報を用いた医学用語分類, 情報処理学会研究報告, 知能と複雑系研究会報告 (ICS-128), pp. 122–127 (2002).

[12] 王玉馨, 清水伸幸, 吉田稔, 中川裕志: 単語類似度ネットワークを通じた自動同義語獲得, 情報処理学会研究報告, 自然言語処理研究会報告 (NL-185), pp. 7–14 (2008).

[13] 寺田昭, 吉田稔, 中川裕志: 文脈情報による同義語辞書作成支援ツール, 情報処理学会研究報告, 自然言語処理研究会報告 (NL-176), pp. 87–94 (2006).

[14] 萩原正人, 小川泰弘, 外山勝彦: 類義語自動獲得における間接依存関係の有効性, 言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」論文集, pp. 43–46 (2007).