自然言語処理 131－8
音声言語情報処理 26－8
（1999．5．28）

学術文献の著者キーワードに基づく専門用語間の関連度計算とその応用

相澤 彰子，影浦 峡
{akiko, kyo}@rd.nacsis.ac.jp
学術情報センター

本稿では，学術文献に与えられる著者キーワード情報を利用して，専門用語の間の関連度を計算する方法について検討する．具体的には，まず学術文献データベース中の著者キーワード項目に注目して，同一文献に与えられたキーワードどうしを共起関係リンクで結び大規模な用語グラフを作成する．次にこのグラフ上での平均パス長を用いて任意の2つの用語間の距離を定義し，距離が近いものほど関連度が高いとみなす．本稿ではまた，実際に大規模な学術文献データベースを用いて用語グラフを作成し，これをテキストの自動分類問題に適用することによって，用語グラフに基づく関連度のもっともらしさを評価し，直接共起や出現文脈ベクトルを用いる場合と比較する．

Akiko AIZAWA and Kyo KAGEURA
National Center for Science Information Systems

In this paper, we report a method for calculating association score between technical terms utilizing statistical information obtained from the keyword lists assigned to academic papers by the authors. The proposed method first generates a global 'term graph' utilizing co-occurrencies of the keywords in the list. Then, an association score of a selected term pair is calculated based on the average path length between the terms on the graph. This paper also shows experimental results where a large-scale term graph actually extracted from NACSIS Academic Database is applied to text classification task. It is shown that the proposed association measure works effectively with such keyword corpus while other existing methods such as mutual information or context vector suffer from data sparseness problem.

1 はじめに

学術文献に与えられる著者キーワードは，その分野の最先端で研究を行う著者自身の手により選択された専門性の高い情報である．本稿では，大量の学術文献から収集された著者キーワードが，精度の高い情報源として利用できることに注目して，著者キーワード情報に基づく専門用語のシソーラス自動構築の可能性を探る．

著者キーワードをコーパスとして用いることの利点として，著者がキーワードとしてあげる語はそれぞれが独立した意味を持つ複合語であり，一般辞書には含まれない専門性の高い語例を豊富に含むこと，著者キーワードリストは著者自身の手により意識的に作成されており，テキスト領域中で同時に出現する語を機械的に取り出す場合と比較してはるかにノイズの少ない共起情報が得られることなどがあげられる．一方問題点として，1つの文献について高々数個

のキーワードが挙げられるのみであることから，データがスパースであることがあげられる．本稿ではこのようなスパース性に対処するための手段として，グラフ的な尺度を用いた関連度計算の可能性を検討する．このために，まず，学術文献データベース中の著者キーワード項目に注目して，同一文献に与えられたキーワードどうしを共起関係リンクで結ぶことにより大規模な用語グラフを作成する．次にこのグラフ上での平均パス長を用いて任意の2つの用語間の距離を定義する．さらに本稿で，テキストの自動分類問題への適用を通して，このような用語上での距離計算のもっともらしさを評価する．

以下，**2.** で，共起情報に基づくコーパス手法について簡単に概観したのち，著者キーワードのコーパスとしての性質を調べて議論する．**3.** では，用語グラフの定義と関連度の計算法をまとめる．**4.** では，実際に大規模な学術文献データベースを用いたテキスト分類実験の概要を述べる．**5.** では実験結果をまとめ，用語グラフを用いる場合と直接共起や出現文脈ベクトルを用いる場合の性能を比較する．最後に **6.** で考察を行う．

2 共起情報に基づく関連度計算

シソーラスの自動構築においては，コーパス中での共起情報を利用して語と語の関連度を定量的に評価する試みが従来より幅広く行われている．これらの手法を，統計量として用いる共起関係の種類に注目して分類すると，以下の3通りの場合が想定できる．

(1) 1次的共起に基づく方法

語と語が実際にテキスト中で共起する頻度から，相互情報量などの共起スコアを用いて語間の結び付きの強さを計算する [2] [3] .

一般に，各語を，その語が出現したテキストによって特徴づけて特徴ベクトルを構成して類似度を計算する場合には，一次的な共起情報を利用しているとみなせる [12] . また，複合語の切り出しを目的とする場合には，高い頻度で現れる単語列は語彙的な結束性が強いという仮定のもとで，概念的なまとまりを持つ単語列を抽出する [4][5][6][7] . さらに，対訳コーパスを対象として，高い頻度で共起する言語の異なる語対は互いに類似しているという仮定のもとに，対訳関係を自動抽出する場合もある [8][9][10] [11].

共起頻度を得るテキストの単位としては，文書全体を用いる場合，セグメント化あるいはアラインメントされた領域を用いる場合，形態素情報等を利用して特定の言語構造だけに注目する場合などがある．

(2) 2次的共起に基づく方法[1]

使われる文脈すなわち近接する語が似ている語は類似すると仮定し，語を，その語と共起する語の頻度ベクトル（以下，出現文脈ベクトルと呼ぶ）を用いて特徴付け，これに基づき語間の類似性を評価する．

一般に，各語を，その語がテキスト中で共起する語で特徴づけて特徴ベクトルを構成する場合には，このような2次的共起を考慮しているものとみなせる [14] [15] . 対訳抽出においても，単言語コーパス内での出現文脈ベクトルを利用する手法が提案されている [18][20]. また，2次的共起に基づく特徴ベクトルに対してさらに LSI を適用する場合もある [13] [19] [20].

共起頻度を得るテキストの単位としては，(1) と同様に，文書全体を用いる場合，セグメント化された領域を用いる場合などがある．

(3) より高次の共起に基づく方法

上記の2手法に対して本稿で検討するのは，3次以上の高次な共起情報も利用しようという拡張である．類似の方法は従来，人手により構築された辞書やシソーラスに基づく関連度計算法として用いられ，(1)(2) のコーパス手法と対比させて議論されることが多かった [15] [23] . たとえば，EDR 辞書 [21][14]，LDOCE[22]，Collins

[1] 一次的共起（first-order cooccurrence），2次的共起（second-order cooccurrence）という呼び方は [13] による．

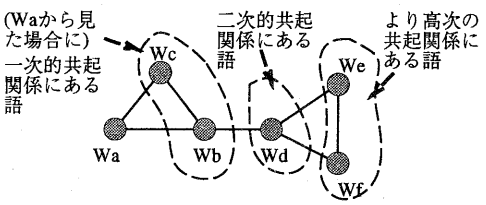
English Dictionary[23], WordNet[24] などを用いた例がある．また，対訳辞書に適用した例としては [25] などがある．

図 1に上記 3 種類の共起関係を対比させた例を示す．(a) の表では，3 つのテキスト領域において出現した語を黒丸で示してあり，(b) のグラフは，(a) で共起する語どうしをリンクで結んだものである．(b) において，同一のリンクで結ばれる関係が一次的共起（以下「直接共起」と呼ぶ）に，中間的なノードを 1 つ介して結ばれる関係が 2 次的共起，それ以外が高次の共起関係に対応する．本稿では以下，語から語にいたる最短経路のリンクの数を「共起次数」と呼ぶことにする．たとえば図 1において，語対 W_a と W_e は，$(W_a - W_c)$, $(W_c - W_d)$, $(W_d - W_e)$ の 3 本のリンクを経由することから共起次数は 3 である．

	W_a	W_b	W_c	W_d	W_e	W_f
$Text_1$	●	●	●	–	–	–
$Text_2$	–	●	–	●	–	–
$Text_3$	–	–	–	●	●	●

（ただし●が出現した語，–が出現しなかった語）

(a) テキストと出現語の関係



(b) 対応するグラフ表現

図 1: 1 次的，2 次的，より高次の共起関係

3 著者キーワードからの用語グラフの生成

3.1 キーワードコーパスの特徴

表 1 に，異なる 2 つのコーパスのスパース性の違いを，平均共起次数を用いて評価した例を示す．具体的には，学会発表データベース [26] から取り出した情報処理および人工知能分野の文献 27,389 件について，

(A) 和文表題，和文著者キーワード，和文抄録よりなる全文（以下「抄録全文コーパス」）

(B) 和文著者キーワード（以下「キーワードコーパス」）

のいずれかをテキスト領域とする場合を対象として統計量を調べた．抄録全文コーパスでは，テキストを形態素解析ツール（CHASEN Ver1.5 [27]）を用いて解析したのちに簡単な複合語処理を行い，抽出されたすべての出現語について共起次数を調べた．一方，キーワードコーパスについては，著者があげた個々のキーワードを 1 つの複合語とみなして，これらの間の共起次数を調べた．

表中の数値は，それぞれ，出現語の異なり総数 (W)，任意の語についてテキスト中で直接共起する語の平均数 (N)，出現語数に対する直接共起語の比率 ($r = \frac{N}{W}$)，到達可能な任意の 2 つの語の間の平均共起次数 (D)，到達不能である語対の比率 (u) である．D および u については，ランダムに 100,000 対の語をサンプリングして調べた．抄録全文コーパスとキーワードコーパスでは r の値が 1 桁異なり，これを受けて D や u の値も大きく異なることがわかる．

表 2には，上記の 100,000 語対に関する共起次数の分布を示してある．抄録全文コーパスでは，約 98%の語対が 2 次的共起の関係にあるのに対して，キーワードコーパスでは，2 次以下の共起で捉えられる語対は僅か 2%となっている．すなわち，抄録全文コーパスでは，語間の結び付きの度合を評価する際に 2 次以下の共起を考慮すれば十分であるのに対して，キーワードコーパスではスパース性の問題から，より高次の共起の考慮が必要であることがわかる．

表 1: 共起次数に基づくコーパスのスパース性の比較

	抄録全文コーパス	キーワードコーパス
出現語数 (W)	240,630 語	37,888 語
平均直接共起語数 (N)	378 語	11.6 語
平均直接共起語比率 (r)	0.16%	0.031%
平均共起次数 (D)	2.0	3.8
到達不能比率 (u)	0%	9.7%

表 2: コーパスによる共起次数の経験分布の比較

共起次数	抄録全文コーパス	キーワードコーパス
1	62	26
2	97,937	2,102
3	2,001	31,586
4	0	44,496
5	0	10,690
6	0	1,246
7	0	109
8	0	26
9	0	4
∞	0	9,715

3.2 用語グラフと用語間距離の定義

用語グラフ G を，n_i個の語を含む語のリスト $L_i = (w_{i1}, w_{i2}, \cdots, w_{in_i})$ の集合 L を用いて，$G = (W, L)$ で与える [28]．ただし W は L に含まれるすべての語の集合であり，G 上で Wはノード，L はリンクに対応する．通常のグラフは1本のリンクに対して端点となるノード2つを定義するが，本稿のでは任意個の端点を定義するようリンクの定義を拡張している．これは後述するように，平均パス長の計算において同じリンクを2度たどらないという制約条件を明示的に表現するための便宜的なものである．

用語グラフ上で，任意の用語 A と B の間の距離を W_a から W_b にいたる経路の長さ（すなわち経由するリンク数）を用いて，「用語 W_aから用語 W_bに到達する極大の経路集合で，互いにリンクを共有せず，その平均経路長が最小であるようなものの平均長」と定義する．たとえば図2で，W_a と W_b の間には，経路「$W_a - W_d - W_b - W_e$」と経路「$W_a - W_c - W_b$」が存在し，その平均長は $(3+2)/2 = 2.5$ となる．「極大」とは，W_a と W_b の間にその経路集合以外の経路が存在しないことを表す．「互いにリンクを共有しない」という制約条件は，同じリンクは一度しかたどらないことに対応し，共起次数が低いものをより優先する一種の正規化の役割を果たしている．たとえば図2において，W_c と W_b の間に存在する3本のリンクのいずれか1本のみが経路に含まれ，また「$W_a - W_c - W_f - W_b$」のような迂回路は選択されないことを示す．

距離の計算においては，グラフ理論で最小カット容量を求めるために用いられる Dinic のアルゴリズム [29] を参考に，用語 W_a から用語 W_b にいたる最短経路木を作成し，経路長の短い経路から順に数え上げている．上記の制約条件を設定したことから，得られる解は近似的なものであるが，計算量はノード数 n，リンク数 m に対して $o(n^2m)$ で押えられる．

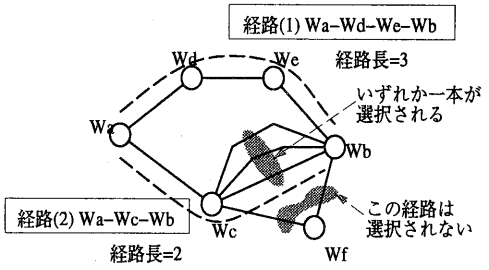


図 2: 平均経路長の計算例

用語グラフは単純に，1つの文献に対応するリンクの集合で定義され，文献データベースの更新にあわせて随時更新することが可能である．また，著者キーワードに注目することの利点として，多くの学術文献は和英両方のキーワードを持つことからそのままの形で多言語に対応できることがあげられる．

3.3 用語グラフの規模

学術情報センターの学会発表データベースに登録されている文献のうち，311,463 件から和文著者キーワードを取り出してリンク集合 L を作成したところ，文献あたりの平均キーワード数は 4.4，生成した用語グラフのノード数（すなわち W）は 354,769 個となった．その他の統計量については，$N = 14.4$, $r = 0.0041\%$, $D = 4.2$, $u = 7.1\%$となり，表 1のキーワードコーパスとほぼ同様の結果が得られた．

4 テキスト分類実験の概要

4.1 テキスト分類問題

語の出現頻度に基づくテキストの自動分類では一般に，多数の特徴語から構成されるベクトル空間上にテキストやカテゴリを配置し，内積やコサイン尺度を手がかりに，与えられたテキストにもっとも近いカテゴリを選択する．十分な量のテキストが利用可能である場合には，テキスト学習の分野における研究で示されているように，有効な特徴語の選択とそれらに対する適当な重み付けの学習によって分類の性能を高めることができる [30] .

一方，カテゴリを人手で登録する場合や検索文から対象分野を判定する場合などでは，利用可能な特徴語の数がわずかであることから，テキストとカテゴリ間での特徴語の共有を前提とする上記の方法では有効な計算が行えない．このような場合にシソーラスを併用すると，直接共起しない特徴語の間でも関連度の計算が可能になる [31][15][28]．すなわち，分類対象となるテキストの特徴語集合を $V = (v_1, \cdots, v_m)$，カテゴリ i の特徴語集合を $W_i = (w_{i1}, \cdots, w_{in_i})$，シソーラス上で定義される V と W_i の関連度 $sim(V, W_i)$ として，$sim(V, W_{i^*}) = \max_i sim(V, W_i)$ なるカテゴリ i^* を選択すればよい．ここで $sim(V, W)$ の計算においては，すべての (v_i, w_i) について関連度を計算するのではなく，V, W それぞれについて集計した統計量を用いることが一般的であり，本稿においてもこの方法で関連度計算を行っている．

4.2 テキスト分類実験の概要

以下の実験では，著者キーワードを用いて V および W をそれぞれ定めた上で，$sim(V, W)$ の計算法として，

(1) 相互情報量を用いる場合 (**MI**),

(2) 出現文脈ベクトルを用いる場合 (**CV**),

(3) 用語グラフ上での平均パス長を用いる場合 (**GRAPH**),

の 3 通りを想定してテキスト分類の性能を比較する．

実験に用いたテキスト分類問題は,「絶縁紙, 粘弾性」「電力系統, コヒレンシ, 動的等価縮約, 安定度」などの著者キーワードを手がかりに，文献を表 3 に示す 20 学会（カテゴリ）のいずれかに分類するものである．以下，カテゴリの特徴抽出に用いる文献集合を「参照用データ」，分類したい文献を「評価用データ」と呼び区別する．参照用データとしては，前節で用語グラフの作成に使用した文献の中から，

(A) 大規模参照用データ（文献総数 311,463 文献，平均 65,573 文献／カテゴリ）

(B) 小規模参照用データ（文献総数 1,000 文献，50 文献／カテゴリ）

の 2 種類を設定した．評価用データとしては，用語グラフの作成には使用していない文献の中から，(A) に対しては学会を問わずランダムに，(B) については各学会 5 文献ずつランダムに，それぞれ 100 個の文献を選んで用いた．評価用データの文献あたり平均キーワード数は (A) の場合 4.3 語，(B) の場合 4.5 語であった．

テキストの分類では，まず，各学会カテゴリごとに，参照用データ中の出現頻度を tf*idf で重みづけした値が上位である N 語を特徴語とし

て選択する．次に，評価用データの著者キーワード集合と各学会カテゴリのと特徴語の間の距離を，**MI, CV, GRAPH** いずれかの方法で求め，もっとも近い学会を選んで選択カテゴリとする．最後に，評価用データが実際に発表された学会名を正解カテゴリとして，結果の判定を行う．図3 に以上の処理手順の概要をまとめる．

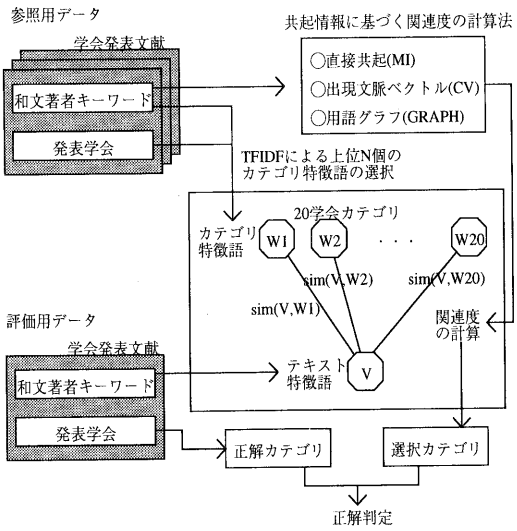


図 3: テキスト分類実験の手順

4.3 学会クラスの設定

実験に用いた 20 学会の中には，「土質工学会」と「土木学会」のように比較的関連性が高いもの，「日本農芸化学会」と「電子情報通信学会」のように関連性が低いものが存在する．評価にあたってこのような学会間の関連度を考慮するため，似かよった学会をまとめた学会クラスを設定し，学会判別の正解率とあわせて学会クラス判別の正解率も調べることにした．

具体的には，用語グラフ上での関連度に基づき UPGMA（算術平均を用いた対グループ法）による機械的なクラスタリングを適用し 11 個の学会クラスを求めた．その結果を表 3に示す．人間の直感ともよく合致するクラスが得られており，このことからも用語グラフ上での距離計算の有効性が確認できる．

表 3: 学会カテゴリのクラスタリング結果

CLASS 0	電子情報通信学会, 情報処理学会, テレビジョン学会
CLASS 1	日本建築学会, 土木学会, 土質工学会
CLASS 2	高分子学会, 日本セラミックス協会
CLASS 3	電気学会
CLASS 4	計測自動制御学会, システム制御情報学会
CLASS 5	日本薬学会, 日本農芸化学会, 日本植物生理学会
CLASS 6	精密工学会
CLASS 7	日本解剖学会
CLASS 8	日本放射線技術学会
CLASS 9	日本家政学会
CLASS 10	日本応用動物昆虫学会, 日本生態学会

5 実験結果

テキスト分類実験の結果を表 4にまとめる．表の中の「平均有効語数」は，参照データ中に出現しない未知語を除いた文献あたりの平均キーワード数を示す.「有効データ数」は，類似度計算が可能であった評価用データの数を示す．具体的には，文献中であげられているキーワードがすべて未知語であったり，すべてのカテゴリのいずれとも共起関係がなかったりした場合に分類不能な無効データとなる．また表では，**MI, CV, GRAPH** を比較するため，20 学会，11 学会クラスの両者について，(1)100 個の評価用データのうちの正解数，および (2) 関連度に基づき 20 学会をランク付けした場合の正解学会の平均順位を示した．N の値としては $N = 20$ および $N = 1000$ または 180 の 2 通りを設定した．

表の結果から明らかなように，大規模参照用データについて，$N = 20$ および 1000 のいずれの場合についても，**GRAPH** が最もよい結果を示している．また **CV** と **MI** を比較すると前者の方が性能が高いことから，高次の共起情報

の利用がテキスト分類問題において有効に働いていることがわかる．

また小規模参照用データついては，カテゴリの学習に使用した参照用データの数がカテゴリあたり50文献と少なく，必ずしもカテゴリ全体の頻出語ではない語が特徴語として選ばれている．この場合に，**MI**については，ほとんど分類のための手がかりは得られず，ランダムに学会を選んだ場合と同等の性能しか得られていない．一方，**CV**および**GRAPH**は依然として比較的高い分類性能を保っており，$N = 180$とした場合の**GRAPH**が一番よい性能を示している．

さらに上記の2実験において，正解数や正解学会ランキングに対するカテゴリ特徴語数の影響を見ると，$N = 1000$または180の場合に，**MI**および**CV**では性能が低下するのに対して，**GRAPH**では逆に性能の向上が見られるがわかる．これは，**MI**および**CV**では十分な統計情報が得られないため，テキスト特徴語とカテゴリ特徴語との共通語の有無だけが主要な手がかりとして用いられるためであると考えられる．

6 考察

本稿では，高次の共起情報を利用した用語間の関連度計算とそのテキスト分類問題への適用について述べ，コーパスのスパース性が高い場合には，3次以上の高次の共起情報の利用が有効であることを示した．本稿で扱ったのは，学術文献の著者キーワードというコーパスとしては特殊なものであるが，通常のテキストコーパスを用いる場合でも，たとえば言語構造や辞書情報に基づき信頼度の高い情報だけを取り出す場合などでは，同様なスパース性の問題が生じることが予想され，類似の手法が有効であると考えられる．

謝辞

本研究は学術振興会の未来開拓学術研究推進事業による「高度分散情報資源活用のためのユービキタス情報システムに関する研究」のもとで行われた．

参考文献

[1] Akiko AIZAWA and Kyo KAGEURA: "An Approach to the Automatic Generation of Multilingual Keyword Clusters," in Proc. of the First Workshop on Computational Terminology, p.8-14 (1998).

[2] G. Grefenstette: "Explorations in Automatic Thesaurus Discovery," Kluwer Academic Publishers (1994).

[3] Ludovic Lebart, André Salmen and Lisette Berry: "Exploring Textual Data," Kluwer Academic Publishers (1998).

[4] Church, K. W., & Hanks, P.: "Word association norms, mutual information, and lexicography." *Computational Linguistics*, 16(1) p.22-29 (1990). .

[5] Dunning, T.: "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics*, 19(1), p.61-74 (1993).

[6] Smadja, F.: "Retrieving collocations from text: Xtract." *Computational Linguistics*, 19(1), p.143-177 (1993).

[7] Su, K-Y., Wu, M-W., & Chang, J-S.:"A corpus-based approach to automatic compound extraction," *Proceedings of 32nd ACL*, p.27-30 (1994).

[8] Gale, W. and Church, K. W.:"Identifying Word Correspondences in Parallel Texts," *Proceedings of DARPA Speech and Natural Language Workshop*, p.152-157 (1991).

[9] Melamed, I. D.:"A Word-to-Word Model of Translational Equivalence," 35th Conference of ACL (ACL'97) (1997).

[10] Smadja, F. and McKeown, K. R., and Hatzivassiloglou, V.: "Translating Collocations for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, 22(1), p.1-38 (1996).

[11] 北村美穂子，松本裕治「対訳コーパスを利用した対訳表現の自動抽出」情報処理学会論文誌，vol.38, no.4, p.727-736 (1997).

[12] C.J.Crouch: "An Approach to the Automatic Construction of Global Thesauri," Information Processing and Management, Vol.26, No.5, p.629-640 (1990).

[13] Hinrich Schütze and Jan O. Pedersen: "A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval," Information Processing & Management, Vol.33, No.3, p.307-318 (1997).

表 4: テキスト分類実験の結果

		20 学会正解数	20 学会ランク	11 学会クラス正解数	11 学会クラスランク	平均有効語数	有効データ数
大規模参照用データ	**GRAPH** (N=20)	63	2.03	84	1.33	3.7	99
	GRAPH (N=1000)	67	1.65	85	1.24	3.7	99
	MI (N=20)	28	5.17	38	2.67	3.7	94
	MI (N=1000)	29	5.16	39	2.76	3.7	98
	CV (N=20)	52	2.53	76	1.67	3.7	99
	CV (N=1000)	49	2.17	69	1.65	3.7	99
小規模参照用データ	**GRAPH** (N=20)	48	3.61	65	2.18	3.7	97
	GRAPH (N=180)	52	2.31	71	1.58	3.7	97
	MI (N=20)	5	9.88	15	5.08	3.7	91
	MI (N=180)	4	10.37	8	5.47	3.7	94
	CV (N=20)	49	3.02	67	1.80	3.7	97
	CV (N=180)	39	2.95	61	1.85	3.7	97

[14] 湯浅夏樹，上田徹，外川文雄「大量文書データ中の単語共起を利用した文書分類」情報処理学会論文誌, Vol.36, No.8, p.1819-1827 (1995).

[15] 福本文代，鈴木良弥，福本淳一「辞書の語義文を用いた文書の自動分類」情報処理学会論文誌,ol.37, No.10, p.1789-1799 (1996).

[16] Finch, S. P.:"Finding Structures in Language," PhD Thesis, Univ. of Edingurgh, 216p (1993).

[17] Hughes, J.:"Automatically Acquiring a Classification of Words," PhD Thesis, Univ of Leeds, 226p (1994).

[18] 田中久美子，岩崎英哉「非対訳コーパスを用いた訳語関係の抽出」情報処理学会自然言語処理研究会報告，110-13, p.87-95 (1995).

[19] Hinrich Schütze: "Automatic Word Sense Discrimination," Computational Linguistics, Vol.24, No.1, p.97-124 (1998).

[20] Genichiro Kikui: "Term-list Translation using Mono-lingual Word Co-occurrence Vectors," in Proceedings of COLING-ACL'98, p.670-674 (1998).

[21] 崔進，小松英二，安原宏「EDR 電子化辞書を用いた単語類似度計算法」情報処理学会自然言語処理研究会報告，93-1, p.1-6 (1993).

[22] Hideki Kozima and Teiji Furugori : "Similarity between words computed by spreading activation on an English dictionary, " in Proc. of EACL-93, p.232-239 (1993).

[23] Niwa Yoshiki and Nitta Yoshihiko: "Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries," in the Proceedings of COLING-94, p.304-309 (1994).

[24] Jay J. Jiang and David W.Conrath: "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in Proc. of International Conference Research on Computational Linguistics (1997).

[25] Takenobu Tokunaga and Hozumi Tanaka: "The Automatic Extraction of Conceptual Items from Bilingual Dictionaries," in Proc. of PRICAI'90, p.304-309 (1990).

[26] NACSIS:"Introduction to the National Center for Science Information Systems," NACSIS (1997).

[27] 松本裕治他「日本語形態素解析システム CHASEN Vresion 1.5 使用説明書」, NAIST Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学院大学 (1997).

[28] 相澤彰子, 影浦 峡:「グラフ的類似度尺度による学術文献の自動分類に関する検討」言語処理学会第 5 回年次大会, p.217-220 (1999).

[29] R.E. タルジャン著, 岩野和生訳「データ構造とネットワークアルゴリズム」マグロウヒル (1989).

[30] Yiming Yang and Jan O.Pedersen: "A Comparative Study on Feature Selection in Text Categorization," in Machine Learning, Proc. of the 14th International Conference, p.412-420 (1997).

[31] 湯浅夏樹，外川文雄「概念識別子の頻度分布を利用した文書分類」情報処理学会情報学基礎研究会，39-5, p.33-40 (1995).