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

相澤 彰子　　影浦 峡

学術情報センター

{akiko, kyo}@rd.nacsis.ac.jp

1 はじめに

本稿では，少数のキーワードを手がかりにした学術的テキストの自動分類に関する検討結果を報告する．

テキスト文書の自動分類では十分な量のテキストデータの存在を想定する場合が多いが，たとえば，学術データベースに入力される検索文から対象分野を判定する等の応用においては，利用可能な特徴語の数は僅かである．従来よりテキスト文書を自動分類するためのアプローチとして，(1) テキスト中の文字列の出現頻度に基づく手法，(2) 予め定義されたシソーラスを用いる方法の2つが存在するが [1][2]，以下に述べる理由によって，特徴語の数が十分でない場合には，これら従来の手法をそのまま適用することは困難であると考えられる．

まず，(1) の出現頻度に基づく方法では，多数の特徴語から構成されるベクトル空間上に文書や文書カテゴリを配置し，その上での内積やコサイン尺度を類似度とする．一般に与えられた文書集合に対する有効な特徴語とそれらに対する適当な重み付けを定めることが重要であるとされており，tf*idf（term frequency * inverse document frequency）や相互情報量など種々の統計量による検討比較が行われている [3][4]．しかしこの場合に，類似度計算は文書とカテゴリ間での特徴語の共有を前提としているため，入力される語数が少なくなってしまうと，分類の性能が低下してしまうことが予想される．

一方 (2) のシソーラスに基づく方法では，予めシソーラス上で定義されたカテゴリと対応する語の体系を手がかりに，文書とカテゴリの関連度を判定する．しかし，学術文献のような専門性の高い分野においては，特徴語として重要な語が必ずしもシソーラス中に定義されているとは限らないという問題があり，特に，入力される語数が少ない場合には，すべての語が未知語であるといったケースが発生することから，分類の性能が低下してしまうことが予想される．

以上の背景に基づき本稿では，文献データベースから自動的に構築した用語関係のグラフ表現に基づくテキスト分類の可能性を検討する．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（非類似度）を定義し，これに基づきカテゴリと文書の関連度を計算する．

本稿で提案する手法において，用語グラフは一種のシソーラスの役割を果たすことになるが，対象となる文献データベースから自動的に構築する点が，従来のシソーラスに基づく手法と異なる特徴となっている．このような用語グラフを用いることにより，共起する特徴語が全くない文書の間でも距離の計算が可能になり，特に少数のキーワードしか利用できない場合において有効に分類が行えることを予備的な実験により示す．

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．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，ここ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である．

用語グラフ上で，任意の用語 A と B の間の距離を A から 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 A から用語 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1で，A と B の間には，経路「$A-D-B-E$」と経路「$A-C-B$」が存在し，その平均長は $(3+2)/2 = 2.5$ となる.「極大」とは，A と B の間に上記の2つの経路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，図1では，C と 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「$A-C-F-B$」のような迂回路は選択されないことを示す．

同様にして，用語集合 W_A と W_B の間の距離を「W_A

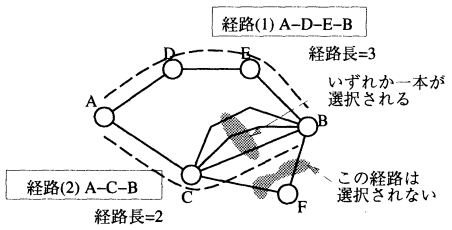

図 1: 平均経路長の計算例

に含まれるいずれかの用語から W_B に含まれるいずれかの用語にいたる極大の経路集合で，互いにリンクを共有せず，平均経路長が最短であるようなものの平均長」として定義する．経路が存在しない場合には距離は無限大となるが，実際の適用においてはこのようなケースはほとんど観察されていない．なお距離の計算には，グラフ理論で最大フロー容量を求めるために用いられる Dinic のアルゴリズムを適用している．

本稿の実験では，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，表 2 に示す 20 学会で発表された 311,463 件から和著者キーワードを取り出してリンク集合 L を作成した．この場合の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（すなわち異なり用語数）は 354,769 個，また簡単な正規化後に異なりリンク総数は 297,039 本となった．

以上のようにして作成した用語グラフは，ツリー構造によって概念間の階層的な関係を表現する通常のシソーラスとは異なり，グラフ的な構造によって用語間の距離を直接的に表現している．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ることから，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，もともと複合語を単位とすることから単語間の強度測定等による複合語の自動抽出 [5] が不要であること，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 テキスト分類実験の概要

実験に用いたテキスト分類問題は，学会発表データベースからランダムに選択した 100 個の文献を，表 2 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．

評価用データとして用いた 100 個の文献は，前節の用語グラフの作成には使用していないもので，それぞれ実際の発表学会が正解として与えられている．また参照用データとしては，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，平均 65,573 文献／カテゴリ）
(B) 小規模参照用データ（文献総数 973，平均 49 文献／カテゴリ）

の 2 種類を作成する．比較のため，頻度に基づくアプローチの中でも基本的な手法である tf*idf に基づく方法（以下 **TFIDF**），および本稿で提案する用語グラフに基づく方法（以下 **GRAPH**）のそれぞれについて，共通の参照用データからカテゴリの特徴を定め，評価用データを用いて分類の正解率を比較する．

TFIDF では，まず，学会ごとに，対応するすべての参照データの和文タイトル，和著者キーワード，和文抄録を取り出し，形態素解析ツール（CHASEN Ver1.5）を用いて特徴語と出現頻度情報の一覧を作成する．この場合に複合語の問題を配慮して，「連接する名詞は新たな特徴語とする」等のルールを適用する．たとえば「情報検索」という語からは，「情報」「検索」「情報検索」のすべてが特徴語として取り出されることになる．次に，各特徴語の重みを tf*idf で定め学会ごとに特徴語ベクトルを作成する．すなわち，学会 i における用語 j の出現頻度（i に含まれるすべての文献中での出現頻度の総和）を v_{ij}，用語 j が出現する学会の数を n_j とするとき，学会 i における用語 j の重み w_{ij} を次式で定める．

$$w_{ij} = \frac{v_{ij}}{\sum_k v_{kj}} \times \frac{1}{n_j} \qquad (1)$$

評価用データについては著者キーワードから取出した用語集合から特徴語ベクトルを生成し，コサイン尺度を用いてもっとも近い学会を選んで分類結果とする．

GRAPH では，参照用データの著者キーワード中の用語について，学会ごとの出現頻度を計算した上で，各特徴語の重みを式 (1) の tf*idf により定め，上位 N 語を学会の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の上位 N 語との間の距離を用語グラフ上で求め，もっとも近い学会を選んで分類結果とする．N の値としては $N = 20$ および $N = 1000$ の 2 通りを設定している．

さらに参考のため **TFIDF*** として，**TFIDF** と同様にカテゴリの特徴量を求めた上で，評価用データに

表 1: 実験に用いたテキスト分類手法

手法	参照用データの利用項目	カテゴリ特徴量	評価用データの利用項目	テキスト特徴量	カテゴリーテキスト関連度の定義
GRAPH	著者キーワード	tf*idf により選んだ上位 N語	著者キーワード	著者キーワード集合	用語グラフ上での用語集合間の距離
TFIDF	タイトル，著者キーワード，抄録	tf*idf により重み付けした全ての語	著者キーワード	tf*idf により重み付けした著者キーワード	特徴ベクトル間のコサイン尺度
TFIDF*	タイトル，著者キーワード，抄録	tf*idf により重み付けした全ての語	タイトル，著者キーワード，抄録	tf*idf により重み付けした全ての語	特徴ベクトル間のコサイン尺度

表 2: 実験で用いた参照データ

		大規模参照データ			小規模参照データ		
ID	学会名（文書カテゴリ）	文献数	著者キーワード数	自動抽出した索引語数	文献数	著者キーワード数	自動抽出した索引語数
0	電子情報通信学会	86,105	110,214	664,828	266	1,008	7,367
1	日本建築学会	55,679	66,626	364,038	175	753	5,220
2	情報処理学会	29,350	40,016	221,403	95	379	2,761
3	高分子学会	24,742	35,993	238,038	76	357	3,055
4	土木学会	23,843	32,944	167,913	74	280	2,284
5	電気学会	17,784	26,626	161,102	56	199	2,087
6	計測自動制御学会	23,904	13,129	120,936	40	159	1,456
7	土質工学会	7,879	5,920	64,286	23	93	912
8	日本セラミックス協会	7,590	11,632	87,776	21	93	1,087
9	日本薬学会	6,610	14,402	108,682	20	74	899
10	精密工学会	5,927	9,409	54,335	17	65	642
11	テレビジョン学会	4,753	11,027	61,552	17	78	857
12	システム制御情報学会	4,580	9,251	48,101	14	51	596
13	日本解剖学会	3,930	7,797	53,182	11	52	702
14	日本農芸化学会	3,835	10,065	74,707	15	62	923
15	日本植物生理学会	3,830	6,567	59,853	12	54	721
16	日本放射線技術学会	3,718	7,077	49,039	13	50	563
17	日本家政学会	2,915	7,026	51,401	11	51	615
18	日本応用動物昆虫学会	2,826	5,787	51,433	8	40	554
19	日本生態学会	2,438	5,935	46,169	9	41	552
合計（異なり数）		311,463	354,769	2,086,070	973	3,769	23,525

ついても和文タイトル，和著者キーワード，和文抄録を利用して同様に文献ごとの特徴語ベクトルを定め，コサイン尺度によりもっとも近い学会を選んで分類結果とする場合についても比較を行った．

表 1 に実験に用いた手法の概要を，表 2に各カテゴリに対する特徴量の要約をまとめる．

4 実験結果

表 2 の 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度が高いもの，「日本農芸化学会」と「電子情報通信学会」のように関連度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．

学会間の類似度計算は **TFIDF** および **GRAPH** のそれぞれの手法を用いて行ったが，両者に大きな違いはみられず，人間の直感ともよく合致する結果が得られたことから，前者で求めた類似度に対して UPGMA による機械的なクラスタリングを適用し 11 個の学会クラスを設定した．その結果を表 3に示す．

次に，大規模参照用データに対する結果を表 4にまとめる．ここで **TFIDF** および **GRAPH** における評価用データは，たとえば「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集合であり，平均 4.24 個の語から構成されている．また表の中の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している．この場合は 100 個のデータ中で，キーワードがすべて未知語であるものが 1 つ（＝「まちづくり, 新通市場, 参加者の拡大啓発,

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

表 4: 大規模参照用データに対する結果

	20 学会正解数	11 学会クラス正解数	有効データ数
GRAPH (N=20)	56	73	99
GRAPH (N=1000)	62	85	99
TFIDF	61	80	99
TFIDF*	81	93	100

「しかけ」」) 含まれたため, 有効データ数は 99 個となっている.

GRAPH では, 各カテゴリに対する特徴語の数が 20 個および 1000 個のそれぞれの場合について実験を行った. 表 4 から, 20 個のキーワードしか利用しない場合でもかなりよい正解率が得られ, 1000 個を用いる場合では, **TFIDF** よりも高い正解率が得られることがわかる. 一方, **TFIDF*** では, 1 文献あたり平均 54.5 個の語を索引語として抽出し, 正解率が他の場合よりも高いことから, 利用可能な特徴語の数が正解率に大きく影響していることがわかる.

なお一般にはテキスト学習の分野では, 特徴語の数が多くなると過学習の効果によって分類の精度が低下するため, 特徴語の選択および重みの「学習」が重要とされている. 今回の実験では, このような選択や学習の効果は調べていないが, **TFIDF** および **TFIDF*** において参照データ中に現れた 2,086,070 個の語すべてを特徴語としているにもかかわらず, 比較的よい正解率が得られていることは注目に値する.

最後に, 小規模参照用データに対する結果を表5にまとめる. 表5から明らかなように, 参照用データおよび評価用データの特徴語数がともに少なくなる場合には, **TFIDF** では判定に十分な情報が得られないため, 有効データが減少し正解率が低下する. 一方, **GRAPH** では, 正解率の低下はみられない. これは, 前出の実験と同じ大規模データから作成した用語グラフをシソーラスとして有効に利用しているためであると考えられる. **TFIDF*** では評価用データがテキストとして与えられるため依然として高い正解率を保っているが, 参照用データの数がさらに少なくなれば, **TFIDF** と同様の現象が生じることが予想される.

表 5: 小規模参照用データに対する結果

	20 学会正解数	11 学会クラス正解数	有効データ数
GRAPH (N=20)	60	81	99
GRAPH (N=1000)	67	86	99
TFIDF	37	49	68
TFIDF*	73	88	100

5 考察

自動レレバンスフィードバックでは, 与えられた検索語を含む文書集合からさらに重要度の高い語を取出して検索語とする操作が行われる. 本稿で用いた用語グラフ上でも, リンクが 1 文献中で共起する用語集合に対応していることを利用して,「与えられた検索語を始点のノードとして, 経路長 1 で到達可能なノード集合の中で, 到達可能経路数が多いもの」を取り出せば, 類似の操作が定義できる. このように「用語」と「用語間の共起関係」に基づくグラフ構造の上で, 多様な操作や類似度尺度を定義することが可能であり, 計算量の問題を考慮しつつ, 頻度情報に基づく統計的アプローチと組み合わせた適用が有効であると考えている.

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] 湯浅, 上田, 外川「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).

[2] 福本, 鈴木, 福本「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).

[3]Koller,D and Sahami, M. "Toward otimal feature selection," in Proc. ICML-96, 284-292 (1996).

[4]Koller, D. and Sahami, M. "Hierarchically classifying documents using very few words," in Proc. ICML-97, 170-178 (1997).

[5] 木村, 小舘, 朱, 浦野, 富永「文書頻度を用いた論文データベース中の文書の分類に関する検討」情報処理学会自然言語処理研究会 128-23 (1998).