自 然 言 語 処 理 152－23
（２００２．１１．１３）

ベクトル空間モデルを用いた参考文献の同定

堀部 史郎† 新保 仁† 松本 裕治†

ある文献が他のどの文献を参照しているか、という文献の参照情報は、その文献の参考文献一覧の各一文（参考文）が指し示している文献を、文献データベースの中から同定することで獲得できる。この同定を、参考文と文献データの単なる文字列の完全一致判定で行なうことはできない。データに表記の多様性や誤りが存在するためである。
本稿では、参照情報の自動獲得の第一歩として、重み付きベクトル空間モデルを用いた対象データの絞り込み手法を提案した。参考文の指す文献がデータベースに存在するものへの絞り込みと、それに対する文献データ候補の絞り込みの双方で、本手法の有効性を示した。

Identification of Citations Using Vector Space Models

Shiro Horibe† Masashi Shimbo† Yuji Matsumoto†

Databases for scientific papers are widely used. Citation indices for such databases are invaluable in the retrieval of related papers. Methods to determine the relatioship between papers automatically are in demand; however, a naive method such as using exact string matches makes errors because of the various ways references can be formatted. In this paper, we propose a new indexing method based on weighted vector space models in order to determine which papers correspond to which citation entries. Thus, given a citation entry from a reference list, the method outputs candidate papers from the database which the entry may be referring to. Furthermore, the weighting helps to eliminate from the reference list papers that are not available in the database.

1 はじめに

近年、文献の参照・被参照情報の利用が進んでいる。文献の参照・被参照情報とは、ある文献が他のどの文献を参照しているか、または他のどの文献から参照されているか、という情報のことである。この情報は、文献検索システムの出力として、あるいは、文献の価値の算出にも使われている。

一般に、参照情報は文献の末尾にある参考文献一覧から得られる。人手でその情報を大量に獲得するのは現実的ではないため、自動獲得が望まれるが、そのためには電子化された参考文献一覧が必要である。そして、ある文献がどの文献を参照しているかがわかるということは、参考文献一覧の各一文（以降では参考文と呼ぶ）の指し示す文献を、題目・著者名・掲載誌名等の書誌情報からなる文献データベースの中から同定できることを意味する。この同定を、参考文と文献データの単なる文字列の完全一致判定で行なうことはできない。それには二つの理由がある。一つは、参考文の各項目（題目、著者名、掲載誌等）の書かれる順番に多様性があり、さらには、著者の名前と名字の順番など、各項目の中にも多様性があるからである。もう一つは、参考文と文献データベースの双方に誤りが含まれたり、省略表現が用いられたりするからである。誤りの原因には引用者によるものと、対象とするデータがOCR読み取りのデータであれば、OCRによるものもある。

本研究は、著者らが所属する奈良先端大の電子図書館に既に蓄積されているような、OCRで読みとられた文献データを対象とし、文献の参照・被参照情報の自動獲得を目的とした。そして、上述したような問題点から以下のような対応方針を取る。参考文と文献データの文字列の完全一致による同定ではなく、参考文と文献データの文字列の間に近さの尺度を導入し、その尺度を用いて同定する。また、参考文と文献データの数が非常に多いことから、軽い処理で近さを計算する必要がある。

本稿では、文献の参照・被参照情報の自動獲得の第一歩として、対象データの絞り込み手法を提案する。ベクトル空間モデルを用いて、文字列をベクトルに写像し、ベクトルの近さで文字列の近さを表す。その写像の際に、文献データが項目にわかれている点と特定項目に省略表現やデータの重なりが多い点に着目して、特定項目への重み付けを行う。

†奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
{shiro-ho,shimbo,matsu}@is.aist-nara.ac.jp

本稿の構成は以下の通りである。次の章で、関連研究を紹介し、その後、ベクトル空間への写像の方法と重み付けの手法を述べる。そして、その手法の有効性を検証するための予備実験とその結果を報告する。

2 背景

WWW 上にある文献から、参照・被参照情報を自動獲得している研究はいくつかある。

Citeseer [4] は、英語の論文を対象とした大規模な論文検索システムである。Lawrence ら [3] は、参考文の集合を同一の文献を指す参考文どうしにクラスタリングする、というタスクにおいていくつかの手法を比較している。編集距離の一種である LikeIt[1] より、単語の unigram と bigram を素性とする二値ベクトル表現を用いて、共通する単語（の組）の数と共通しない単語（の組）の数の比を利用する手法の方が、良い結果が得られると報告している。Citeseer のシステムにおいては、さらに、参考文の各項目をヒューリスティックな手法で特定し、その各項目毎に単語の二値ベクトルを用いた手法でクラスタリングしており、実験として、1158 件の参考文をクラスタリングした結果、5 パーセントのクラスにのみ誤りが含まれていたと報告している。そして、そのようにして得られた各クラスを用いて参照・被参照関係を獲得している。

Cora[6] も、英語の論文を対象とた論文検索システムである。基本的に Citeseer と同じ手法を用いて参照・被参照関係を獲得しているが、参考文の各項目をヒューリスティックな手法ではなく、HMMs を用いて特定するという点で異なる。実験として、200 件の参考文の 4479 単語に、項目のラベルを付けるというタスクにおいて、90 パーセント以上の精度が得られたと報告している。

難波ら [9] は日英の文献を対象として、参照・被参照関係を獲得している。参考文どうしをクラスタリングしたり、項目毎に分割したりせずに、個々の参考文と文献データから、文字の 6-gram を二文字ずつずらして取り出し、これらがある閾値以上一致するかどうかで同定している。

OCR の読みとりデータから、参照・被参照関係を獲得しているデータベースも存在する。ISI 社の SCI(Science Citation Index) は、自然科学分野を対象とした文献データベースである。また、同社の JCR(Journal Citation Reports) では、学術雑誌の比較、評価を行なっている。ある雑誌の掲載文献への引用が多ければ多いほど、その雑誌は権威がある、という考えに基づき参照・被参照関係の利用がなされている。

ISI 社は人手も用いて参照・被参照関係を獲得しているが、その自動化の試みとして、Hitchcock[2] らの Open Journal project がある。ここでは、参考文の各項目がすでに特定されたデータを対象としており、まず参考文の著者名と掲載年の項目を用いて一致する文献データを絞り込んでいる。それから、題目の項目に誤りへの対応としていくつかの前処理を行なった後、共通しない単語の数を閾値に用いて同定している。

また、三平ら [8] は、日本語を対象とし、参考文と文献データの同定に注目している。参考文の一部の項目をヒューリスティックな手法を用いて特定し、それらの項目については一致する文献データに候補を絞り込んでいる。それから、候補となった文献データの題目と参考文から文字 bigram を取り出し、一致する文字組の数と参考文の文字組の数との比を算出し、その値で文献データ候補をクラスタリングして同定している。実験として、59 件の参考文を同定するというタスクにおいて、58 件の参考文を正しく同定できたと報告している。

WWW 上の文献を対象としているシステムには、網羅的な文献の収集が困難だという問題点がある。その点、OCR 読み取りを対象とした場合、データベースに加える文献は恣意的に決められるため、電子化されていない古い文献を加えることや、ある特定の雑誌に関しては全て網羅するといったことも可能である。また、上記の研究の多くについては、文献や著者の同定処理に甘さがあり、これは SCI にも言えることだが、たとえば、同じ姓名を持った人の分類は考慮されていない、などの問題がある。

3 項目への重み付け

本手法は、ベクトル空間モデルを用い、参考文のベクトルとその参考文が指す文献データのベクトルが近くなるような写像を施すことを目指す。

ベクトルの素性には文字や単語や subsequence[5] などの素性が考えられる。subsequence は近さを計算する処理が重いことから、データの絞り込みに用いる素性としては適さない。本手法では文字あるいは単語を素性として用いる。

3.1 文献データへの重み付け

後の実験で示すように素性に文字や単語をそのまま使うと、必ずしも参考文とそれが指し示す文献データのベクトルが最も近くなるとは限らない。

たとえば、一方の掲載誌名は省略表記で、もう一方の掲載誌名は非常に長い正式表記で表された場合を考える。同一のものを指していても、表記が異なることによって、文字列としてみると大きな違い

が生じ、結果として、それぞれが離れたベクトルに写像されてしまう。

また、書誌情報の項目（題目、著者名、掲載誌等）の中には、ほとんど同じ内容を持つ文献データが多数存在する項目がある。たとえば掲載誌項目は、同じ雑誌に載っている文献は全て同じ掲載誌名を持つ。もし、参考文が指し示す文献データの掲載誌名が長いと、参考文の近くには、それが指し示す文献データだけでなく、それ以外の掲載誌名が同一な文献データもそれぞれ写像されてしまう。

本手法では、以上のことを考慮して、掲載誌項目に重みを付ける。

以下に文献データへの重みの付け方を示す。

1. 文献データを各項目毎にベクトルに写像する
2. 掲載誌項目のベクトル ($\vec{J}$) の大きさだけを α 倍 ($0 \leq \alpha \leq 1$) し、それを重み付き掲載誌ベクトルとする
3. 重み付き掲載誌ベクトルとそれ以外の項目のベクトルをそれぞれを足し合わせて、重み付き文献データベクトル ($\vec{D'}$) をつくる

したがって、文献データベクトルを $\vec{D}$ とすると、$\vec{D'}$ は以下の式から求まる。

$$\begin{aligned}\vec{D'} &= \text{（文献データから掲載誌を除いたベクトル）} \\ &\quad + \text{（重み付き掲載誌ベクトル）} \\ &= (\vec{D} - \vec{J}) + (\alpha\vec{J})\end{aligned}$$

3.2 参考文への重み付け

しかし、上記の文献データへの重み付け方法には問題がある。参考文は項目にわかれていないため、上記の方法では掲載誌項目に重みを付けることができない。その結果、重みをつける前に既に表記が完全一致していた参考文と文献データのベクトルが、文献データのみへの重み付けによって、離れたベクトルへと写像されてしまう。

そこで、以下に文献データと同じような重みを参考文に付ける方法を示す。

1. 参考文をベクトルに写像する。
2. 参考文ベクトル ($\vec{C}$) を文献データの掲載誌ベクトル ($\vec{J}$) に正射影し、これを参考文の掲載誌成分ベクトル ($\vec{C_J}$) とする。

   $\vec{C}$ と $\vec{J}$ のなす角を θ とすると、

   $\vec{C_J} = \frac{|\vec{C}|\cos\theta}{|\vec{J}|}\vec{J}$
3. 参考文の掲載誌成分ベクトルの大きさを α' 倍する。ただし α' は、α が文献データの掲載誌ベクトルにつけた重みであるのとは異なり、掲載誌ベクトルに正射影された成分への重み付けである。

   そこで、文献データベクトル ($\vec{D}$) を文献データの掲載誌ベクトルに正射影したベクトル ($\vec{D_J}$) が文献データの掲載誌ベクトルを α 倍することで何倍になるかを求め、これを α' とする。

   $\vec{D}$ と $\vec{J}$ のなす角を ϕ とすると、

   $\vec{D_J} = \frac{|\vec{D}|\cos\phi}{|\vec{J}|}\vec{J}$

   $\alpha' = \frac{|(\vec{D_J} - \vec{J}) + (\alpha\vec{J})|}{|\vec{D_J}|}$
4. 参考文ベクトルから参考文の掲載誌成分ベクトル ($\vec{C_J}$) を引いてから重み付き参考文掲載誌ベクトル ($\alpha'\vec{C_J}$) を参考文ベクトルに足し、重み付き参考文ベクトル ($\vec{C'}$) をつくる。

したがって、$\vec{C'}$ は以下の式から求まる。

$$\begin{aligned}\vec{C'} &= \text{（参考文から掲載誌成分を除いたベクトル）} \\ &\quad + \text{（参考文の重み付き掲載誌成分ベクトル）} \\ &= (\vec{C} - \vec{C_J}) + (\alpha'\vec{C_J})\end{aligned}$$

こうすることで、掲載誌の項目に重みを付け、かつ、文献データと参考文の両方に同じ重みをつけることができる。

4 実験

参考文に対する文献データの絞り込みという観点から、以上の手法の有効性を検証するために実験を行なった。

4.1 実験条件

本実験は、著者らの研究室（奈良先端大自然言語処理学講座）で管理している文献データと参考文を対象とした。

文献データには、自然言語処理、人工知能、情報検索、認知科学の関連文献の国際会議予稿集や論文誌から英語のもの、18601 件を用いた。文献データは、著者名、題目、掲載誌、ページ、掲載年、の各項目毎にわかれていて、それぞれ人手にて入力されている。

参考文には、1997 年から 2001 年発行の自然言語処理関連の論文誌、会議予稿集から得られた 1707

本の英語文献の参考文献一覧より、人手で各文に区切りを入れて得られた 30855 件から、無作為に選んだ 2000 件を用いた。参考文は OCR で読みとってテキスト化しているため、読みとり誤りが存在し、項目毎にもわかれてもいない。また、参考文の指し示す文献が文献データベース中に存在するものは、人手で判定した結果、2000 件中 650 件であった。

近さを計算するにあたって前処理を行なった。参考文と文献データの大文字の使い方は統一されていないことを考慮して、全ての文字を小文字化した。また、参考文は OCR 読み取りにより得られたため、文字列に改行が含まれている。そこで、改行と改行により単語がわかれた場合に生じるハイフンを除去した。ハイフンの除去には、繋げた文字列が辞書に存在するかどうかで判定した。

ベクトルの素性には単語 unigram と単語 bigram と文字4 -gram の三通りを行なった。単語は、英数字以外の文字とスペースで区切った。

また、ベクトルの各素性のスケールには、頻度を x とすると、

$$Damp(x) = \begin{cases} 1 & (|x| \geq 1 \text{ のとき}) \\ x & (\text{それ以外のとき}) \end{cases}$$

とした場合と、

$$Log(x) = \begin{cases} 1 + log_e(x) & (|x| \geq 1 \text{ のとき}) \\ x & (\text{それ以外のとき}) \end{cases}$$

とした場合を用いた [7]。

重みを付ける項目は掲載誌項目である。重み付けは、文献データのみ、文献データと参考文の両方の二通りに対して、0 から 1 の範囲で行なった。掲載誌の重みが 1 とは、重みをつけずに掲載誌項目をそのまま用いることを意味する。

ベクトルの近さの尺度には cosine 類似度を利用した。

4.2 実験結果と評価

実験は二つの観点から行なった。

実験 1 ある参考文に対して文献データの正解候補を絞り込む性能を評価するための実験である。

正解が存在する参考文 650 件を対象とし、最も cosine 類似度の高い文献データが、その参考文が指し示す正解文献データであれば正解とし、そうでなければ不正解とした。不正解件数もさることながら最低位にも注目する。最低位とは、文献データを cosine 類似度が大きい順に並べた際に、正解文献データの順位が最も低かった事例の順位を表す。この最低位が小さければ小さいほど、文献データの正解候補を狭い範囲にまで絞り込むことができることを意味する。

表 1 の各条件は、重みを付けなかった時と、最低位が最も小さくなる重みの時の不正解数と最低位を表す。どの条件においても、重みを付けることで最低位は小さくなっている。さらに、文献データと参考文の両方への重み付けは、文献データのみへの重み付けより、同程度かさらに小さくなっている。特に、素性が単語 unigram でスケールが $Damp$ の場合は、文献データのみへの重み付け、文献データと参考文の両方への重み付けの両方で、最低位を 3 番までに収めている。

図 1 は、最低位が特に小さかった条件（素性が単語 unigram でスケールが $Damp$）における、掲載誌への重みと最低位の関係を表す。全ての範囲で重みをつけない時より最低位は小さくなっている。重みを小さくし過ぎると、最低位はやや大きくなる。また、参考文と文献データの両方への重み付けの方が、文献データのみへの重み付けより、重みの変化に対して広い範囲で小さい最低位が維持されている。

実験 2 参考文の集合を、文献データ中に正解が存在しそうな参考文の集合に絞り込む性能を評価するための実験である。

文献データ中に正解が存在する参考文 650 件と、正解が存在しない参考文 1350 件を合わせた 2000 件を対象とした。

参考文と各文献データを比較し、cosine 類似度の最高値を求める。その最高値が、ある cosine 類似度の閾値を超えれば正解が存在すると判定し、超えなければ正解は存在しないと判定した。この判定に対して、再現率と適合率を以下のように定義する。

$$\text{再現率} = \frac{TP}{TP + FP}$$

$$\text{適合率} = \frac{TP}{TP + FN}$$

ただし、TP、FP、FN、TN はそれぞれ以下の表で与えられる。

	データ中に正解がある 650 件	データ中に正解がない 1350 件
正解ありと判定	TP	FN
正解なしと判定	FP	TN

今回目標とした絞り込み手法という観点からは、高い再現率を達成していることが最も望ましい。表 2 は、各条件に対して、再現率が 1、かつ、最も高い適合率を示す重みを選んで、その時の適合率、不正解数、最低位を示した。

素性	スケール	条件	重み	不正解	最低位
単語 unigram	Damp	–	1.0	3	150
		D	0.1	7	**3**
		CD	0.5	**2**	**3**
	Log	–	1.0	7	605
		D	0.3	9	9
		CD	0.6	5	5
単語 bigram	Damp	–	1.0	17	9720
		D	0.2	16	1462
		CD	0.5	15	957
	Log	–	1.0	21	10390
		D	0.2	15	2017
		CD	0.5	16	1781
文字 4-gram	Damp	–	1.0	18	290
		D	0.1	8	45
		CD	0.4	10	23
	Log	–	1.0	20	781
		D	0.2	8	209
		CD	0.5	12	94

表 1: 実験 1 の結果。条件は、–は重みをつけない、D は文献データのみ、CD は参考文と文献データの両方に重みをつけたことを示す。

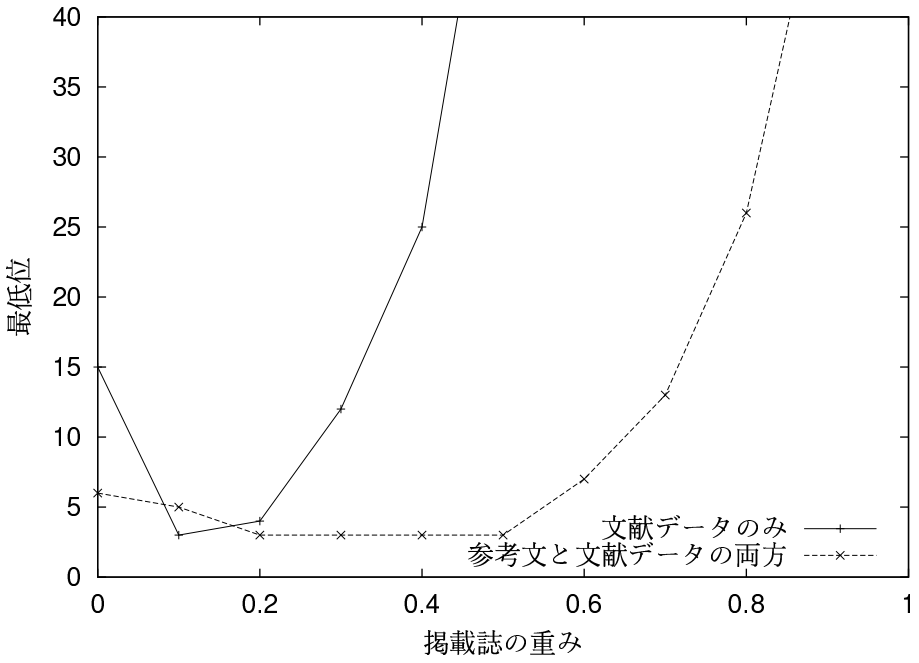

図 1: 実験 1 の不正解事例の最低位。素性は単語 unigram、スケールは $Damp$ を用いた。

素性	スケール	条件	重み	適合率	不正解	最低位
単語 unigram	Damp	–	1.0	0.581	3	150
		D	0.6	**0.730**	5	74
		CD	0.7	**0.690**	3	13
	Log	–	1.0	0.505	7	605
		D	0.5	0.660	7	59
		CD	0.7	0.573	6	16
単語 bigram	Damp	–	1.0	0.484	17	9720
		D	0.9	0.515	15	9468
		CD	0.9	0.493	16	9307
	Log	–	1.0	0.502	21	10390
		D	0.8	0.524	20	10003
		CD	0.9	0.513	18	9941
文字 4-gram	Damp	–	1.0	0.605	18	290
		D	0.4	0.660	7	139
		CD	0.9	0.543	14	141
	Log	–	1.0	0.507	20	781
		D	0.3	0.617	8	266
		CD	0.8	0.565	16	357

表 2: 実験 2 の結果。条件は、–は重みをつけない、D は文献データのみ、CD は参考文と文献データの両方に重みをつけたことを示す。

ほぼ全ての条件で、重みを付けることにより適合率は高くなった。再現率を 1 にする時の適合率を求めたため、最高でも 73 パーセントであった。参考文と文献データの両方に重みを付けるより、文献データのみに重みを付けた方が、適合率はより高くなった。

図 2 は、再現率を 1 にした時の適合率が特に高かった条件（素性が単語 unigram でスケールが $Damp$）における、再現率と適合率の関係を表す。ほとんどの範囲で、参考文と文献データの両方に重みを付けるより、文献データのみに重みを付けた方が、より高い適合率を示した。重みの付け方による適合率の違いは、再現率が 1 に近付いたところで最も顕著になった。

図 3 は、再現率を 1 にした時の適合率が特に高かった条件（素性が単語 unigram でスケールが $Damp$）における、掲載誌への重みと適合率の関係を表す。重みが約 0.6、0.7 の時、最も適合率が高くなった。しかし、重みを小さくし過ぎると重みをつけない場合より適合率は低くなった。また、適合率を最も高くする重みのピークと実験 1 で示した最低位を最も小さくする重みのピークは異なった。このことから参考のために実験 1 と同様、不正解数、最低位をあわせて表に含めてある。

4.3 考察

本実験は OCR 読み取りデータを対象としたため、単語を素性の基準に選ぶと、ある単語の文字に一ヶ所でも誤りが含まれた時に、その単語全体が別の単語と認識されてしまい、ベクトルの近さに悪影響を与えることが予想された。そこで、文字 4-gram を基準に取った実験も行なったが、結果的に単語 unigram の方が性能が良かった。これは、OCR の読み取り精度が思ったより良かったためだと考えられる。

また、単語 bigram は単語 unigram より性能が悪かったが、その理由として著者名の表記の問題がある。名字と名前の順番、名前のイニシャル表記、ミドルネームの省略などがあるため、名字は一致したとしても、二単語の繋がりとしては一致しないことが多いからである。

スケールは Log より Damp の方が全体的に性能が良かった。参考文のような短い文字列を対象としてベクトルに写像する時は、Damp の方が良いと思われる。文字列が短い、つまり素性の数が少ないため、大きな値を持つ素性の一致による影響が、素性の数が多い時に比べて大きくなるからである。

実験 1 で、最低位を大きく引き下げていた参考文には、それが指し示す文献データの掲載誌の表記が省略表記されているという特徴があった。文献データの表記を修正することでも、最低位を改善することはできると思われる。しかし、小規模なデータ

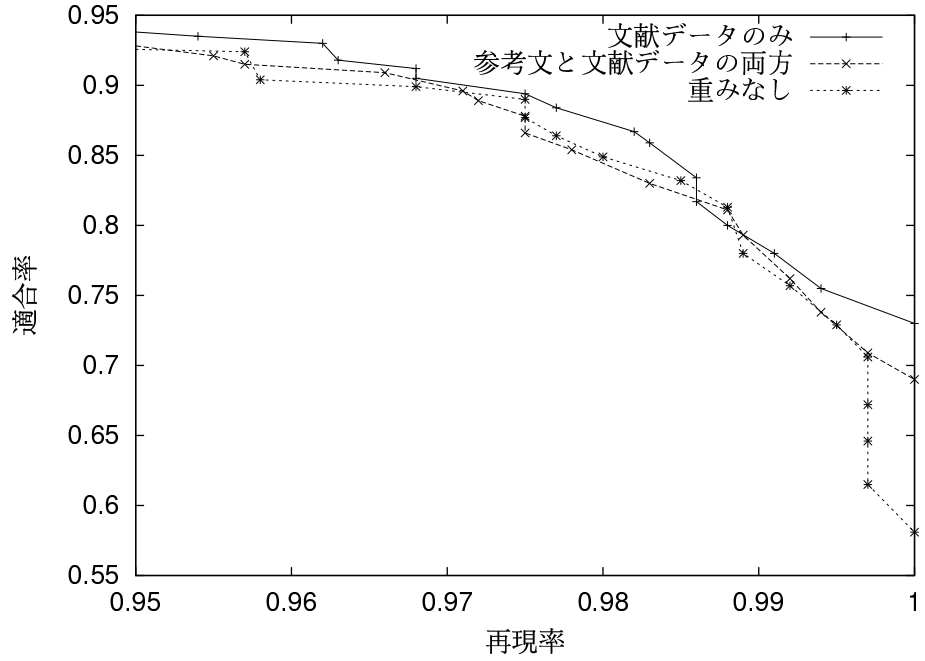

図 2: 実験 2 の再現率に対する適合率。素性は単語 unigram、スケールは $Damp$ を用いた。文献データのみへの掲載誌の重みは 0.6、参考文と文献データの両方への掲載誌の重みは 0.7 である。

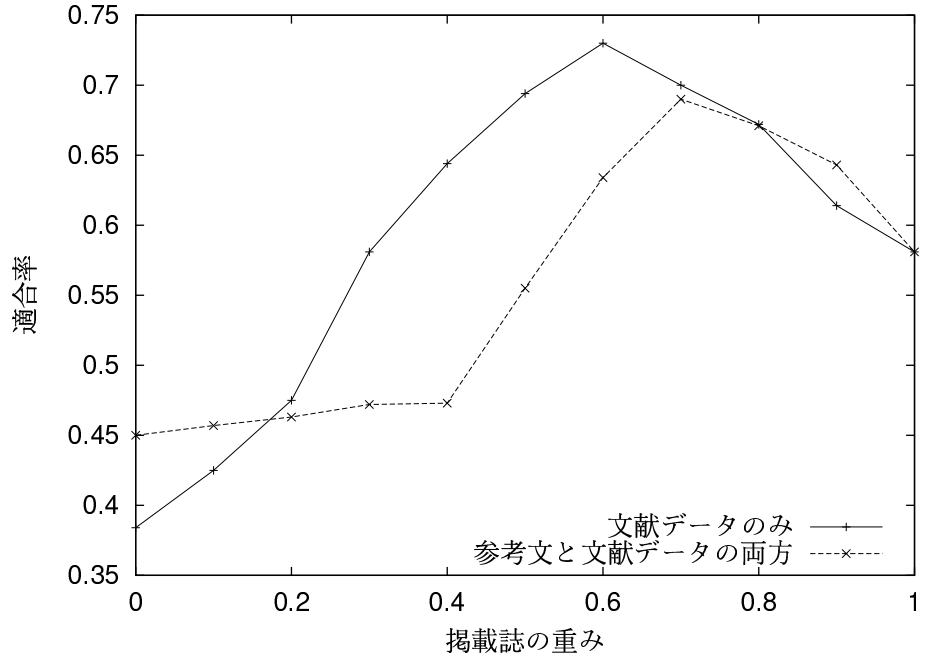

図 3: 実験 2 の適合率の最大値（再現率 1 の時）。素性は単語 unigram、スケールは $Damp$ を用いた。

ベースにおいてはそれで構わないが、大規模なデータベースでは修正のコストも高くなる。このことからも、掲載誌項目に小さな重みを付けるだけで最低位を低く抑えることができる本手法は有用である。

実験 2 で、適合率を引き下げていた参考文の掲載誌の表記は、省略して書かれている場合が多かった。文献データの掲載誌への重み付けは、その表記の違いの影響を小さくすることができる。つまり、参考文とそれが指す文献データ間の cosine 類似度を高くすることができ、結果として適合率をあげることができた。一方、参考文の中には、それが指し示す文献データと、掲載誌の表記は一致しているが、他の項目の表記が異なるものも存在する。掲載誌への重みを下げることは、そのような参考文とそれが指し示す文献データとの cosine 類似度を、大きく下げる。そのため、重みを下げ過ぎると適合率は下がってしまった。したがって、適合率をあげるためには重みを調節することが重要である。

また、実験 2 では、参考文が文献データ中に正解を持つかどうかの判定に、全ての文献データと比較して、最大値となった cosine 類似度を用いた。しかし、判定の基準は他にも考えられる。例えば、最大値とその次との cosine 類似度の差を用いる方法がある。これは、正解文献データの cosine 類似度は絶対的に大きいだけでなく、他の正解ではない文献データと比べて、相対的にも大きいことが期待できるからである。

5 まとめ

ある参考文に対して文献データの正解候補を絞り込むタスクと、参考文の集合を、文献データベース中に正解が存在しそうな参考文の集合に絞り込むタスクの両方において、掲載誌の項目に重みを付けることが有効であることを示した。

今回の実験では、英語文献のみを対象とし、日本語文献は扱わなかった。これは、OCR による日本語参考文の読み取り精度がきわめて悪かった[1]ためである。日本語文献の取り扱いの際には、このような精度の悪化や、それに起因する単語分割の困難さを考慮すると、文字 n-gram も有効であると期待できる。

今後は、絞り込まれた正解候補文献データに対して、より詳細な同定の処理を行なうことと、第 4 節の実験 2 で用いた手法を発展させ、参考文の指し示す文献が文献データベース中に存在するかしないかの判定を考えている。

[1]誤りが特に多かったのが、英語の参考文献が混在している場合だった。

謝辞 本研究の実施にあたって、筆者の所属する研究室の諸氏には常日頃から多くのコメントと助力を頂いた。ここで深い感謝の念を表したい。

参考文献

[1] S. R. Buss and P. N. Yianilos. A bipartite matching approach to approximate string comparison and search. Technical report, NEC Research Institute, 1995.

[2] S. Hitchcock, L. Carr, S. Harris, J. M. N. Hey, and W. Hall. Citation linking: Improving access to online journals. In R. B. Allen and E. Rasmussen eds., *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pp. 115–122, Philadelphia,PA,USA, 1997.

[3] S. Lawrence, K. Bollacker, and C. L. Giles. Autonomous citation matching. In O. Etzioni ed., *Proceedings of the Third International Conference on Autonomous Agents*, New York, 1999. ACM Press.

[4] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[5] H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C. C. Watkins. Text classification using string kernels. In *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, pp. 563–569, 2000.

[6] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *AAAI Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[7] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, second edition, 1999.

[8] 三平善郎, 山本喜一. 引用文献の同定. コンピュータソフトウェア, Vol.14, No.1, pp. 35–39, 1997.

[9] 難波英嗣, 奥村学. 多言語論文データベースを用いたサーベイ論文検出ーサーベイ論文自動作成の実現に向けてー. 言語処理学会第 8 回年次大会発表論文集, pp. 531–534, 2002.