科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２７　日現在

機関番号：　１５５０１
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５０１２８

研究課題名（和文）内容に基づいた文学作品の典拠和歌検索システムに関する研究

研究課題名（英文）A study on contents-based search system for source poems of Japanese literary works

研究代表者
中田　充（NAKATA, Mitsuru）

山口大学・教育学部・准教授

研究者番号：６０３０４４６６

交付決定額（研究期間全体）：（直接経費）　　2,800,000円、（間接経費）　　840,000円

研究成果の概要（和文）：本研究では，文学作品の典拠となる和歌を内容に基づいて検索する手法を提案した．提案した検索手法は，(1)指定された検索語，(2)検索語と完全に同じ意味を持つ類似語，(3)検索語と類似した意味を持つ類似語，(4)検索語から連想される連想語を含む和歌を検索可能とするものである．同義語と類似語は，主にEDR電子化辞書の概念体系に基づいて求められる．連想語は和歌に含まれる単語の共起頻度情報に基づいて求められる．さらに，万葉集巻1中の84首の和歌に含まれる509語の単語を対象として連想語を求める実験を行い，連想語を求める指標としてMI-Scoreを用いて連想語を求めることが妥当であるとの結論に至った．

研究成果の概要（英文）：In this research, we have proposed a search function for source poems of Japanese literary works based on contents of the poems. Our proposed function makes it possible to obtain source poems which include (1) specified keywords, (2) equivalent terms, (3) synonyms, and (4) associative words. An equivalent term is a word that has the completely same meaning of the keyword. A synonym is a word that has the almost same meaning as the keyword. These are obtained according to the conceptual hierarchy in the EDR electronic dictionary, while associative words are obtained based on the co-occurrence information of words included in poems. Moreover, we have conducted the experiment which obtains associative relationship between 509 terms contained in 84 poems in the volume 1 of Manyoshu, the oldest anthology of Japanese poems. The experiments have shown that associative words should be obtained according to the criterion based on MI-Score.

研究分野：　総合領域

科研費の分科・細目：　情報学・図書館情報学・人文社会情報学

キーワード：　文学情報　典拠和歌　同義語　類義語　連想語　共起情報　万葉集検索

１．研究開始当初の背景

「月日は百代の過客にして，行かふ年も又旅人也」という『奥の細道』の文章は，「都をば霞とともにたちしかど秋風ぞ吹く白河の関」という和歌を「空間の移動である旅を時間の推移に置き換えた」と解釈した上で，「繰り返す時間を旅人」として表わそうとしている．このように，近世散文の中には，それ以前の和歌の内容を言い換えた表現が多く含まれている．その典拠である和歌を参照することは，より深い散文の理解に必要不可欠なものである．しかし，現在のところ，典拠となる和歌は表記に直接現れる語句に基づいてしか探ることが出来ていない．語彙や和歌の内容に基づいた典拠の発見が出来れば，文学史の流れのみならず，広く我が国全体の文化・伝統の由来や変遷をより鮮明なものにすることができる．

２．研究の目的

文学作品は，必ず前時代の作品の一部を典拠として発展させているという特徴がある．それは文化の変遷や伝統が次世代に継承されていく様相を示している．従ってその具体的な特徴を探ることは文学史の究明だけでなく，文化の変遷過程や伝統の継承を明確にし，歴史や民俗の分野にも大きな意義のあるものである．それには文章の意味による典拠を探ることが不可欠である．これには研究者が広く古典籍を読解し多大な時間をかけて探求するしか方法がないというのが現状である．語彙検索による典拠探査は近年容易になってきたが，意味による典拠検索は困難であり未発見の部分も多い．

そこで，文章の意味から典拠を探るシステムを開発し，古典籍における未発見の影響関係を明らかにしていくことを本研究の目的とする．なお，本研究では，対象とする和歌を我が国最初の歌集であり，最も多くの文学作品に影響を与えたと考えられる『万葉集』に掲載された和歌に限定する．

３．研究の方法

具体的な研究の方法は以下の通りである．

(1)本研究を進めるにあたり，代表者らがすでに開発・公開している万葉集和歌検索システムの改良から取りかかる．これまでの検索システムには，和歌の検索速度や和歌データのメンテナンス性などに問題があり，これらを解決する必要がある．

(2)従来の和歌検索手法では，ユーザーが指定したキーワード（以降，検索語）と和歌の文字列との間の単純な文字列マッチングにより，ユーザーが所望する和歌を検索する．これに対して，本研究で提案するシステムでは，より柔軟に内容に基づいて和歌を検索するために，検索語の同義語，類似語，連想語を求めて，検索語以外にこれらの単語を含む和歌を検索する．そのために，まず，検索語の同義語と類似語を求める手法を提案する．ここで，同義語とは検索語と全く同じ意味を持つ単語であり，類似語とは検索語と似た意味を持つが全く同じ意味を持つとは言えない単語である．

(3)検索語から連想される単語である連想語を求める．ある単語が指定された検索語と同じ和歌に同時に用いられている頻度が高い場合，２つの語は連想関係にあると考える．つまり，連想語は，２つの単語の共起頻度情報を用いて求める．

(4)(3)における単語の共起頻度情報を求めるために，和歌の内容を表すキーワードを抽出する仕組みを提案・実現する．

４．研究成果

23 年度は，和歌検索の速度を向上させる，より複雑な検索条件の指定を可能とする，和歌データのメンテナンスなどの更新がより簡単に行えるようにする，などの観点に基づいて，過去に構築していた万葉集和歌検索システムを，データベース管理システム SQLite3 とプログラミング言語 Java を用いて再構築した．これは，万葉集に掲載されている和歌を対称とした典拠和歌検索システムを実現するためには，まず，従来の万葉集検索システムの問題点を解決しておくことが必要であるとの判断からである．

その際に，従来では表現できなかった漢字に対応するために，文字コードを SJIS から UTF-8 に変更した．この結果，従来は 120 文字ほどの表現できていなかった漢字のうち，70 文字程度が表現可能となった．さらに，UTF-8 でも表示できない漢字については，アルファベットなどの記号で置き換えてデータベースに格納しているが，それらの漢字の画像も含めた形で検索結果を表示する機能も追加した．

加えて，現代仮名遣いと歴史的仮名遣いを自動的に変換して検索する機能を実現した．これまでは，和歌に含まれる歴史的仮名遣いを忠実にキーワードとして指定する必要があったが，これによりその手間を省略することが可能となった．また，ユーザーの利便性を考慮して，検索条件に指定されたキーワードを検索結果中に強調表示する機能，検索結果内での文字列検索などの機能も追加した．このように，従来の検索システムの大きな問題点を解決した．

さらに，次のステップとして，検索条件に指定されたキーワードの同義語からも和歌を検索する機能を実現した．あるキーワードの同義語は，概念体系辞書と日本語単語辞書を DB に格納したものを検索することで求めている．ただし，同義語を求める処理に時間が掛かるなどの問題があり，次年度に向けての課題となった．

24 年度では，万葉集に収録されている和歌を内容に基づいて検索する機能の基本設計を行い，それを採用した検索システムの試作版を作成した．さらに，国文学研究においては，複数の手書きの写本で伝えられている作品を対象とすることが多く，検索結果の提示の際には，写本のテキストと画像も結果に含めた形で提示してほしいという要求があることがわかった．しかし，負担の大きさ故に，多くの作品について全ての写本がテキスト化されておらず，微妙に内容が異なる写本を簡単に参照することが難しい．そのため，手書き日本語で書かれた写本の内容を容易にテキスト化する手法が必要とされている．そこで，代表者らがこれまでに実現してきた日本語手書き文字認識技術を，文学作品に適用し，その認識精度を評価して，問題の解決法を検討した．

まず，独立行政法人情報通信研究機構が提供している EDR 電子化辞書を用いて，単語の意味を表す概念のつながりを表す DAG(Directed Acyclic Graph)である「概念体系」(図 1)と各概念に属する単語の一覧である「単語表」（図 2）を，データベース管理システムを用いて実現した．概念cは，$c = (cid, cs, cmt, IB_c, IN_c)$と定義される．ここで，$cid$は概念の識別子であり，$cs$は概念見出しと呼ばれる「その概念を代表する単語」の単語見出しである．また，cmtは，その概念が持つ意味を表す解説文であり，IB_cとIN_cは，それぞれ，概念cの親概念と子概念の集合である．単語wは，$w = (wid, ws, cid)$と定義される．widは単語の識別子，wsは単語見出し（単語を構成する文字列）である．これらには，現在，約 40 万の概念と約 60 万の単語が格納されている．

次に，同じ概念に属する単語同士を同じ意味を持つ「同義語」，同じ概念ではないが似た意味を持つ（概念体系にて近い距離にある）概念に属する単語同士を「類似語」と定義し，概念体系と単語表から，和歌検索時のキーワード（検索語）の同義語と類似語を求める仕組みを設計し実現した．例えば，図 2 中では，単語“日本”,“ジャパン”,“日本国”は同義語であり，概念“日本の旧称”に属する単語“大和”と“秋津島”は，単語“日本”の類似語である．単語xの類似語は，単語xが属する概念とその下位概念，ならびに，単語xが属する概念の類似概念とその下位概念に属する単語である．ここで，類似概念とは，類似した意味を持つ概念であり，ルート概念以外に共通の上位概念をもつ任意の概念c_1, c_2について，以下のいずれかが成立するとき，c_1とc_2は類似概念であるいう．

- 概念c_1に属する単語の単語見出しと同じ単語見出しを持つ単語を概念c_2が含む．
- 概念c_1に属する単語の単語見出しを概念c_2の概念説明（または概念見出し）が含む．

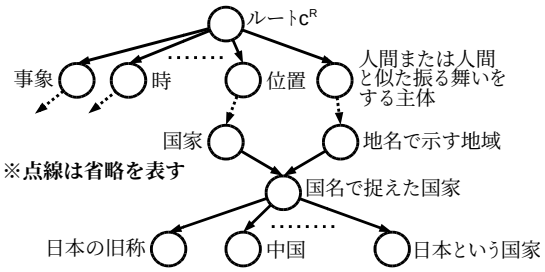

図 1：概念体系

概念
(3bcdec, “日本”, “日本という国家”, {444a40}, {})
(3bca94, “ジパング”, “日本の旧称” , {444a40}, {})
(444a40, “”, “国名で捉えた国家”, {30f772, 444a5f, {3bcdec, 3bca94, …})
(30f772, “”, “国家”, {…}, {444a40, …})
(444a5f, “”, “地名で示す地域”, {}, {444a40, …})
単語
(JWD0373071,“日本”,3bcdec)，(JWD0373072,“日本国”,3bcdec)，(JWD0575082,“ジャパン”,3bcdec)，(JWD0373060,“大和”,3bca94)，(JWD0373051,“秋津島”,3bca94)

図 2：概念と単語の例

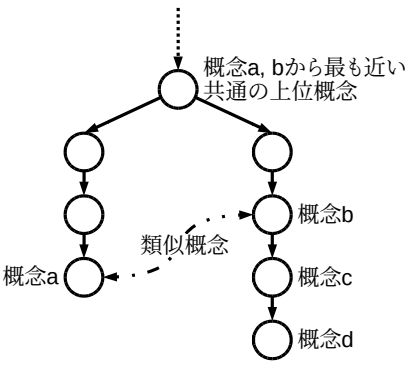

図 3：概念距離

さらに，検索時にかかる時間の短縮のために，全ての単語の類似語をあらかじめ求め，「概念距離」の近い類似語のみを検索の対象とする．概念距離とは，ある単語xとその類似語yの意味の相違を表す尺度である．単語xが属する概念c_xと類似語yが属する概念c_yの関係に応じて，以下の 2 種類の概念距離がある．

【概念距離 1】$\mathrm{dis1}(c_x, c_y)$
概念c_yが概念c_xの類似概念であり，c_xとc_yの最近傍の共通の上位概念を概念c_zとする．$c_x \neq c_z$かつ$c_y \neq c_z$のとき，概念c_xからみた概念c_yまでの概念距離 1：$\mathrm{dis1}(c_x, c_y)$は，概念c_xからc_zまでの距離（辺数）である．$c_x = c_z$または$c_y = c_z$のとき，$\mathrm{dis1}(c_x, c_y) = \mathrm{dis1}(c_y, c_x) = 0$である．

【概念距離 2】$\mathrm{dis2}(c_x, c_y)$
概念c_yが概念c_xの下位概念である（$c_x = c_z$）とき，概念c_xからみた概念c_yまでの概念距離 2: $\mathrm{dis2}(c_x, c_y)$は，概念c_xからc_yまでの距離である．

図 3 は概念体系の一部分を示している．概念aとbが類似概念であるとき，dis1(a, b) = 3，dis1(b, a) = 2である．また概念cがbの下位概念であるので，dis1(b, c) = dis1(c, b) = 0，dis2(b, c) = 1となる（同様に，dis2(b, d) = 2）．

「概念体系」，「単語表」，「概念距離」に基づいて，検索語のみならず，その同義語と類似語も用いて和歌を検索するシステムを，プログラミング言語 Java を用いて実装した．また，万葉集の和歌は漢字のみで表記されているため，筆者らがこれまでに提案してきた手書き文字認識技術が古典文学作品中の漢字にどれくらい有効であるかを認識実験により評価し，その問題点を解決する手法を提案した．

最終年度である 25 年度では，類似語を求める手法の改善と，連想語を求める手法の提案を行った．これまでの類似語を求める手法では，検索語と全く似ていない意味の単語が類似語として得られたり，逆に，類似語が全く得られないことが少なくなかった．これらの問題を解決するため，単語の共起性を利用した類似語を求める新たな手法を考察した．同時に，単語の共起性に基づいて連想語を求める手法を提案した．

一首の和歌において，ある単語がほかの単語と同時に使用されているとき，これらの単語の間には共起性があるとする．共起性が高い単語同士は，複数の和歌において同時に使用される頻度が高いということであり，これらの単語同士を，万葉集における連想語とする．そのために，万葉集に含まれる単語同士の共起性を表現する共起行列（表 1）を作成し，単語同士の共起頻度や共起ベクトルを比較することで，連想語と類似語を求めることとした．提案した一連の手法は以下の通りである．

手順 1：万葉集の和歌の単語から構成される共起行列を作成する．万葉集の訓読文から名詞，動詞等の単語のみを抽出する．抽出した単語の集合を H と K(H=K)とする．H 中の単語を本文語，K 中の単語を共起語と呼ぶ．任意の本語と共起語とが 1 首中に共起する頻度をカウントし，各行が本文語，各列が共起語に対応するような共起行列を作成する．

手順 2：共起行列の行ベクトルは，対応する本文語の共起パターンを表しており，この行ベクトルを共起ベクトルと呼ぶ．共起ベクトルから零ベクトルである行を削除する．

手順 3：共起ベクトルを比較して類似語を探す．共起行列において，ある 2 単語に対応する共起ベクトルが近ければ，共起パターンが似ている．したがって，この 2 つの単語を意味的に近い，つまり類似語であるとする．なお，共起ベクトルの比較においての基準として，共起ベクトルのコサイン類似度を採用した．

手順 4：共起頻度を比較して連想語を探す．ある本文語とある単語が本文中で共起している回数が多いとき，それらの単語は，連想できる可能性が高い．つまり，共起ベクトルにおいて，本文語と共起語の共起頻度が高いとき，その本文語と共起語はお互いに連想語であるとする．

表 1：共起行列（一部抜粋）

	雲	草枕	大君	大和	天皇	都	宮処	国
雲		0	1	1	0	0	0	0
草枕	0		2	0	0	1	0	0
大君	1	2		1	0	2	1	3
大和	1	0	1		1	0	0	4
天皇	0	0	0	1		0	0	1
都	0	1	2	0	0		0	1
宮処	0	0	1	0	0	0		1
国	0	0	3	4	1	1	1	

さらに，万葉集の巻頭から 20 首の和歌の訓読文について手動で取り出した単語のみを対象として共起行列を作成し，提案した手法によって得られる類似語と連想語について考察した．その結果，共起ベクトルのコサイン類似度が高い単語同士はよく似た意味をもつ傾向が高いことが分かったが，これまでに提案してきた EDR 電子化辞書を用いた類似語を求める手法と比べて特に優位に結果が得られたとまでは言えなかった．

また，本文語に対して共起頻度の高い共起語を探すことで，連想語を見つけることができるが，共起頻度の高さを判定する適切な基準を考える必要があることが判明した．さらに，万葉集にのみ含まれる単語（古語など）や現代とは意味が異なる単語などが検索語の同義語や類似語として得られないという問題（問題 1）や，共起語と本文語が同じ単語の集合であり，同じ意味を持つ複数の共起語が存在するため，本文語が同じ意味を持つ複数の共起語と共起するときの共起頻度が別々にカウントされることになり，適切な共起ベクトルが得られないという問題（問題 2）も明らかになった．

次いで，上記の問題点 1 と 2 を解決するために，以下のような改善法を提案した．

【手法 1：概念体系の拡張】

EDR 電子化辞書に収録されている単語や概念は，現代日本語が中心であり，万葉集には含まれるが，現代は使われなくなっている日本語（古語）などはあまり含まれていない．また，万葉集の単語には，現代日本語と同じ表記であるが，意味が異なるものも少なくない．そこで，万葉集に含まれる単語を抽出し，万葉集におけるその意味に相応しい概念に属するように概念体系に組み入れることにより，問題 1 の解決を図る．例えば，「君が

代も我が代も知るや岩代の岡の草根をいざ結びてな」という和歌から，単語「代」を抽出する．この和歌において，「代」は「年齢」という意味である．したがって，「年齢」という概念に属させたように「代」という単語を概念体系に組み入れる．

【手法 2：共起行列の列の統合及び分割】

共起行列の列には，同じ意味を持つ単語が複数存在する．表 1 の例では，共起語「天皇」は「大君」など他にも様々な共起語と同じ意味をもつ単語である．つまり，本文語「国」は，天皇という意味の単語と 4 回共起しているが，共起回数が複数の列（大君と天皇）に分散されている．このような，同じ意味を持つ共起語を統合し，共起回数を合算する必要がある．そのために，共起語の意味属性を考慮した共起行列（表 2）を作成する．この共起行列では，列は 1 つ意味を表す共起概念となる．したがって，意味が同じ共起語は一つの列に統合され，共起語が複数の意味を持つ場合は複数の列が作成される．

表 2:共起語の意味情報を元に作成した共起行列（一部）

	雲	草枕	天皇	大和	都	昔の行政区分	国家	故郷
雲	-	0	1	1	0	0	0	0
草枕	0	-	2	0	1	0	0	0
大君	1	2	-	1	3	2	2	0
大和	1	0	2	-	0	1	2	1
天皇	0	0	-	1	0	1	0	0
都	0	1	2	0	-	1	0	0
宮処	0	0	1	0	-	1	0	0
国	0	0	4	4	2	-	-	-

※「-」は，共起概念に本文語が属していることを表す．

この手法に対して，万葉集巻一(84 首)の和歌の訓読文を対象に単語（509 語）を対象とした実験を行い，どのような単語同士が連想語の関係があるかを調べた．その際に，連想語を求める指標として，ダイス係数，T-Score，MI-Score の 3 つの値について調査し，万葉集和歌検索においては，MI-Score が妥当な連想語を求める指標であるといえるとの結論に至った．

本研究の成果の一部を，平成 25 年度(第 64 回)電気・情報関連学会中国支部大会にて発表したところ，情報処理学会中国支部優秀論文発表賞を受賞した．さらに，平成 26 年 3 月に情報処理学会全国大会にて発表したところ，情報処理学会第 76 回全国大会学生奨励賞を受賞した．

今後の課題として以下のものが挙げられる．

(1) 万葉集全 20 巻約 4500 首の全ての和歌に対して，提案手法を適用して連想語を求めること．
(2) 提案手法によって求めた連想語を用いた検索システムを実現すること．
(3) 「３．研究の方法」の(4)で述べた，和歌の内容を表すキーワードを抽出する仕組みを提案して実現すること．

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

① 岡田雄揮，鍵村好美，中田充，葛崎偉，吉村誠，「単語の意味情報ならびに共起情報を用いた万葉集和歌検索手法の提案」，情報処理学会第 76 回全国大会講演論文集，巻 4，pp.4-803 - 4-804，2014，査読無し．
② 岡田雄揮，鍵村好美，中田充，葛崎偉，吉村誠，「単語の共起性に基づいた万葉集和歌検索機能の考察」，平成 25 年度(第 64 回)電気・情報関連学会中国支部大会論文集，巻 1，pp.183-184，2013，査読無し．
③ 岡田雄揮 中田充 葛崎偉 吉村誠，「概念階層を用いた万葉集和歌検索機能の考案」，情報処理学会第 75 回全国大会講演論文集，巻 4，pp.4-837 - 4-838，2013，査読無し．
④ 岡田雄揮，中田充，葛崎偉，吉村誠，「万葉集和歌検索シス テムの改良に関する研究」，平成 24 年度(第 63 回) 電気・情報関連学会中国支部大会論文集，巻 1，pp.452-453，2012，査読無し．
⑤ Ryo Arakawa，Mitsuru Nakata，Qi-Wei Ge，Makoto Yoshimura，「An Improved Character Clipping Method for Consecutive Handwritten Character Recognition by Feature Graph」，Proc. of ITC-CSCC 2012 (DVD)，pg.E-T2-03，2012，査読有り．

〔学会発表〕（計 5 件）

① 岡田雄揮，中田充，「単語の意味情報ならびに共起情報を用いた万葉集和歌検索手法の提案」，情報処理学会 第 76 回全国大会，2014 年 3 月 11 日，東京電気大学（東京都足立区）．
② 岡田雄揮，中田充，「単語の共起性に基づいた万葉集和歌検索機能の考察」，平成 25 年度(第 64 回) 電気・情報関連学会中国支部大会,2013 年 10 月 19 日,岡山大学(岡山市)．
③ 岡田雄揮 中田充，「概念階層を用いた万葉集和歌検索機能の考案」，情報処理学会第 75 全国大会，2013 年 3 月 8 日，東北大学（仙台市）．
④ 岡田雄揮，中田充，葛崎偉，吉村誠，「万葉集和歌検索シス テムの改良に関する研究」，平成 24 年度(第 63 回) 電気・情報関連学会中国支部大会,2012 年 10 月 20 日，島根大学（松江市）．
⑤ Ryo Arakawa，Mitsuru Nakata，Qi-Wei Ge，Makoto Yoshimura，「An Improved Character Clipping Method for Consecutive Handwritten Character Recognition by Feature Graph」，The 27th International Technical Conference on Circuits/Syste

ms, Computers and Communications, 2012 年 7 月 17 日, 札幌コンベンションセンター(札幌市).

６．研究組織

(1)研究代表者

中田 充（NAKATA, Mitsuru）
山口大学・教育学部・准教授
研究者番号：60304466

(2)研究分担者

吉村 誠（YOSHIMURA, Makoto）
山口大学・教育学部・教授
研究者番号：30263750

葛 崎偉（Qi-Wei, Ge）
山口大学・教育学部・教授
研究者番号：70141116

(3)研究協力者

岡田雄揮（OKADA, Yuki）
H26 年 3 月 山口大学大学院教育学研究科修了. 同 4 月 NEC フィールディング株式会社入社（H23 年度より研究協力者）

鍵村好美（KAGIMURA, Yoshimi）
H26 年 3 月 山口大学教育学部卒業. 同 4 月 NEC システムテクノロジー株式会社入社（H25 年度より研究協力者）