D-043

単語と解説文の対応関係を用いた短歌の内容検索手法

千葉　將貴[†]　　鈴木　優[‡]　　川越　恭二[‡]

[†] 立命館大学大学院　理工学研究科　　[‡] 立命館大学　情報理工学部

1　はじめに

本研究では，短歌をその情景やイメージから検索を行う手法について提案する．現在，多くの情報検索技術が提案されているが，その前提として検索対象文書に数多くの単語が含まれていることが上げられている．なぜなら，検索対象文書に含まれている単語の数が少ない場合，単語の種類数も少ないと考えられ，その検索対象文書を検索するために必要な問合せが少なくなるという問題点があるためである．また，短歌や俳句はその少ない単語量でより多くの情報を伝達しようとする文書形態であるため，短歌を検索する際には既存の情報検索技術だけでは十分ではないといえる．

短歌には，その意味を解説するための解説文が付随していることが多い．つまり，解説文はその短歌のメタデータの一種であると解釈することができる．解説文に含まれる単語の数は既存の情報検索技術を適用する上で必要十分であると考えられる．ところが，全ての短歌に解説文が付随しているわけではなく，解説文の無い短歌も存在する．このような短歌を検索するためには，短歌に含まれる単語自身からその短歌のメタデータを自動生成する必要がある．

そこで本研究では，短歌に含まれる単語とその短歌の解説文に含まれる単語との相関関係を導出することによって，解説文の無い短歌を検索するための手法について提案を行う．提案手法では，短歌に含まれる単語と解説文に含まれる単語の共起頻度を計算し，共起頻度が高いと考えられる単語の組を導出する．これらの単語の組から，解説文の無い短歌に対して，その短歌に含まれる単語と共起頻度の高い解説文の単語を計算し，メタデータとして保持する．それらの短歌を検索する際に，提案手法によって計算されたメタデータを利用することによって，より精度の高い短歌の検索を行うことが可能となる．

2　提案手法

2.1　共起頻度の導出

提案手法では形態素解析を利用して，短歌に含まれる単語とその短歌の解説文に含まれる単語との共起頻度を計算する．次に，全ての共起単語の組から共起頻度が高いと考えられる共起単語を導出する．流れを以下に説明する．

A Content-based Japanese Poem Retrieval Method based on Relationships between Terms and Explanation of Poems
Masataka CHIBA, Yu SUZUKI and Kyoji KAWAGOE
[†]Graduate School of Science and Engineering, Ritsumeikan University
[‡]Faculty of Information Science and Engineering, Ritsumeikan University

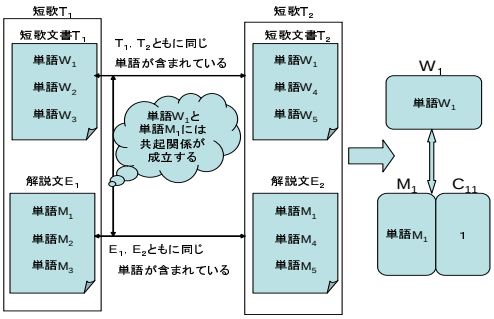

図 1: 共起関係の導出例

短歌とその解説文の集合を $T_i(i{=}1,\ 2,\ \cdots,\ I)$ とし，短歌文書を S_i，解説文を E_i とする．ここで，T_i には S_i と E_i の二つが含まれる $(T_i \subseteq (S_i,E_i))$．短歌に含まれる単語を $W_j(j{=}1,\ 2,\ \cdots,\ J)$ とすると，W_j は S_i に属する $(S_i \subseteq W_j)$．解説文に含まれる単語を $M_k(k{=}1,\ 2,\ \cdots K)$ とすると，M_k は E_i に属する $(E_i \in M_k)$．ここで，S_i に W_j が属し，$S_{i'}$ に W_j が属しており，E_i に M_k が属し，$E_{i'}$ に M_k が属した場合，W_j と M_k には共起関係が成立すると定義する．i' は短歌を区別するための変数である．共起関係が成立する際の条件式は (1) 式のようになる．

$$(S_i \in W_j \wedge S_{i'} \in W_j) \wedge (E_i \in M_k \wedge E_{i'} \in M_k) \tag{1}$$

このとき，短歌に含まれる単語と解説文に含まれる単語の共起関係を満たす共起頻度を C_{jk} とし，共起関係が成立した数を C_{jk} に格納する．W_j は M_k と C_{jk} からなる単語の組で構成される．

$$W_j = (M_k, C_{jk}) \tag{2}$$

短歌に含まれる単語と解説文に含まれる単語において，共起関係を導出する具体例を図 1 に示す．

図 1 において，T_1，T_2 があり，T_1 には S_1 と E_1 の 2 つが属しており，T_2 には S_2 と E_2 の 2 つが属している．S_1 には，W_1 と W_2 と W_3 の 3 つが含まれており，S_2 には，W_1 と W_4 と W_5 の 3 つが含まれている．また，E_1 には，M_1 と M_2 と M_3 の 3 つが含まれており，E_2 には，M_1 と M_4 と M_5 の 3 つが含まれている．ここで，短歌間で同一の短歌に含まれる単語 W_1 と短歌間で同一の解説文に含まれる単語 M_1 には共起関係が成立する．このとき，W_1 と M_1 を単語の組とし，共起頻度 M_{11} に 1 を格納する．同様に，T_i における全ての要素の対応関係を考慮し，短歌に含まれる単語と解説文に含まれる単語の対応関係を導出する．

2.2 解説文に含まれる単語に対する重み付け

2.1 節で導出した単語の組は，解説文に含まれる単語 M_k が短歌に含まれる単語 W_j における短歌に含まれる内容を保持していることを示す．例えば，短歌に含まれる単語として春，解説文に含まれる単語として桜という単語の組が導出された場合，桜は春という単語における短歌の内容を保持する．そして，共起頻度 C_{jk} が高ければ高いほど，M_k が W_j において短歌に含まれる内容であることを強く示す．ここで，W_j における M_k の大域的頻度を考慮することによって，M_k が W_j において短歌に含まれる内容を示す度合いを明確にすることができる．本研究では，TF-IDF 法により重み付けを行う．

2.1 節で導出した単語の組において，短歌に含まれる単語の総数を o とする．また，短歌に含まれる単語 W_j と解説文に含まれる単語 M_k の共起頻度を C_{jk}, 重みを h_{jk} とする．ここで，対象となる解説文に含まれる単語 $M_{k'}$ と共起関係が成立した短歌に含まれる単語の総数を j' とする．j' は短歌に含まれる単語を区別するための変数であり，k' は解説文に含まれる単語を区別するための変数である．このとき，W_j における M_k の重み h_{jk} は式 (3) のように表す．

$$h_{jk} = C_{jk} \cdot \log \frac{o}{j'} \tag{3}$$

抽出された重み h_{jk} を印象度として定義し，2.1 節で導出した単語の組は h_{jk} から再構成される．

$$W_j = (M_k, h_{jk}) \tag{4}$$

2.3 解説文の無い短歌に対するメタデータの付与

本節では，2.2 節で算出された単語の組を基に解説文の無い短歌に対して，メタデータを付与するための手法を説明する．

2.2 節で算出された単語の組において，短歌に含まれる単語 W_j は，解説文に含まれる単語 M_k と印象度 h_{jk} から構成されている．ここで，構成された単語の組における W_j と解説文の無い短歌に含まれる全ての単語においてマッチングを行い，マッチングすれば解説文の無い短歌に M_k と h_{jk} をメタデータとして付与する．メタデータを付与する際，M_k が重複した場合は，重複した h_{jk} を加算し，加算した印象度をメタデータとして保持する．図 2 にメタデータ付与の具体例を示す．

図 2 において，解説文の無い短歌文書に単語が含まれており，2.2 節で算出された単語の組の W_1 と W_4 がマッチングした場合，W_1 と W_4 における単語の組を解説文の無い短歌にメタデータとして付与する．このとき，単語 M_1 が重複するため，重複した M_1 の印象度 h_{11} と印象度 h_{41} を加算し，加算した印象度をメタデータとして保持する．同様に，解説文の無い短歌に含まれる全ての単語と 2.2 節で算出された単語の組を考慮し，メタデータとして保持する．

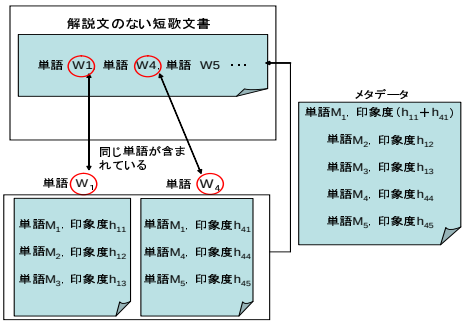

図 2: メタデータ付与の具体例

2.4 内容検索の実現

本節では，内容検索を実現するために解説文の無い短歌に対しての検索順位の決定をどのように行うかを説明する．

利用者は短歌に対して抱いている情景やイメージを問合せとする．ここで，解説文の無い短歌を検索する際に，利用者の問合せと 2.3 節で解説文の無い短歌に保持されているメタデータ M_k とでマッチングを行う．マッチングを行った際，利用者の問合せと同じメタデータ M_k が付与されている短歌を検索結果として出力する．このとき，解説文の無い短歌にメタデータとして保持されている印象度が大きい順に検索結果に反映することで，利用者が意図する短歌を容易に導き出すことが可能となる．

3 まとめと今後の課題

本稿では，短歌に含まれる単語とその短歌の解説文に含まれる単語との相関関係を導出することによって，解説文の無い短歌を検索するための手法について提案した．本手法により，解説文が付随されている短歌と同様に解説文の無い短歌を検索することが可能となる．

今後は，再現率と適合率を算出するための実証実験を行うことで，本手法の有用性を実証する予定である．また，本稿で解説文の無い短歌を検索するための手法を提案したが，利用者のイメージや背景といった利用者の感性は考慮していない．そこで，利用者の感性を考慮した検索を実現するために，利用者の感性を用いた従来の検索システム [1, 2] を考慮し，短歌を対象とした利用者の感性による検索を実現するための手法を提案する予定である．

参考文献

[1] 苅谷花子, 倉林修一, 清木康. “ 味覚印象を対象としたメタデータ生成方式と印象検索方式の実現 ”. 情報処理学会研究報告, pp. 145–152, 2004.

[2] 佐藤真一, 堀江晴彦, 山内正, other. “ 感性データベースシステムとその多次元インタフェース ”. 情報処理学会研究報告, データベースシステム 124-11, 情報学基礎 62-11, pp. 81–88, 2001.