D-020

# ファジィクラスタリングを用いた検索語拡張手法

## A query expansion technique based on fuzzy clustering

鈴木 裕規† 能登谷 淳一‡ 草苅 良至‡ 笠井 雅夫‡
Yuki Suzuki Junichi Notoya Yoshiyuki Kusakari Masao Kasai

## 1. はじめに

計算機システムの高性能化とネットワーク化によって，電子化された情報に計算機を利用してアクセスする機会が増えている[1]．それに伴って，求める情報を得るための効率的な検索に対する要求はますます高まっている[2]．

現在，Web 上の情報を検索するために，いくつかの検索語を入力することにより，それらの検索語を含む Web ページを検索するシステムによるサービスが多数供給されている．しかし，従来から指摘されているように，ユーザは情報検索の際にいつも求める情報に関連する適切な検索語を知っているとは限らない[2]．

情報検索分野では，ユーザの検索語の発見を手助けする検索語拡張の研究がなされてきた[3]．検索語拡張を行う機能を持った Web 検索システムは既に運用されているが，多くの検索語拡張システムは，拡張語候補の所属分野や個数をユーザが制御するための機能を提供していない．そこで，ユーザの必要に応じて拡張語の所属分野や個数を制御できるのならば，ユーザが求める情報の検索をより効率的に行えると考えた．

本研究では，ユーザが拡張語の所属分野や個数をある程度制御可能な「柔軟に利用できる」検索拡張語提示システムを提案する．

## 2. 単語ファジィクラスタリングに基づく検索語拡張手法

本研究で提案する検索拡張語提示システムは，事前に与えられた文書集合から抽出した単語を，分野を反映したファジィクラスタに分類し，ユーザから与えられた検索語と２つのパラメータ（検索拡張語導出パラメータ）によって検索語候補の選抜を行うことで検索語拡張を実現する．

提案システムは大きく２つの部分処理に分けることができる．１つ目は，事前に与えられた文書集合から抽出した単語をファジィ分割する部分処理．２つ目は，入力された検索語とパラメータを用いて前述の分割結果から拡張語候補を選抜する部分処理である．以降に各部分処理を簡単に説明する．

### 2.1 ファジィクラスタリングを用いた単語の分類

ファジィクラスタリングによって単語をファジィ分割すると図１のようになる．

単語 $t_i$ はクラスタ $G_k$ に帰属度 $u_{ik}$ によって所属することになる．このとき，$u_{ik}$ は閉区間[0,1]の値をとるため，後述する検索拡張語導出パラメータとの比較によってクラスタ（分野）数，単語（拡張語）数を容易に制御できる．

†秋田県立大学大学院　システム科学技術研究科
‡秋田県立大学　システム科学技術学部

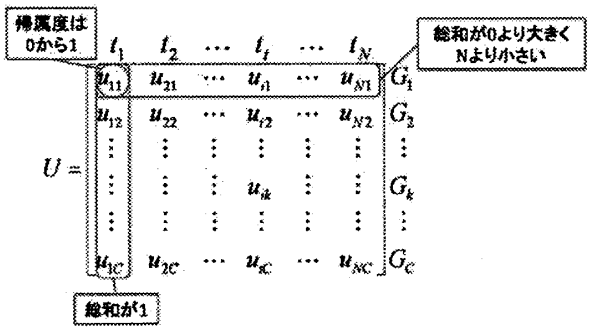


図 1. ファジィ分割（ここで，単語集合 $\mathbf{T}=\{t_1,t_2,...,t_i,...,t_N\}$，クラスタ集合 $\mathbf{G}=\{G_1,G_2,...,G_k,...,G_c\}$）

提案システムで利用する際には，ファジィクラスタリングを行った結果が分野ごとにファジィ分割されていることが望ましい．しかし，単語を分野ごとに分類するには，単語と分野の関係についての情報が必要になる．本研究では，分野情報の代替として単語の文書への出現頻度情報による分類を行った．出現頻度情報で分類を行うと，各クラスタは同じ分野に出現する傾向のある単語で構成されることになる．同じ分野の文書では単語の出現傾向が類似すると考えると，各クラスタは分野の近似と考えることができる．

### 2.2 検索拡張語導出パラメータによる拡張語候補の選抜

提案システムを利用するユーザは，検索拡張語導出パラメータを操作することにより，拡張語候補の所属分野数と個数を制御できる．以下に検索拡張語導出パラメータを利用した検索拡張語候補導出の手順を示す。

#### 2.2.1 クラスタ選抜用パラメータ

クラスタ選抜用パラメータにより検索語が所属するクラスタ集合 $\mathbf{G}=\{G_1,G_2,...,G_c\}$ から選抜クラスタ集合 $\mathbf{G}'=\{\mathbf{G}_k \mid u_{ik} \geq P_G\}$ を得る．クラスタ選抜用パラメータ $P_G$ は０から１の間の実数値としてユーザに要求され，しきい値として利用される（図２）．

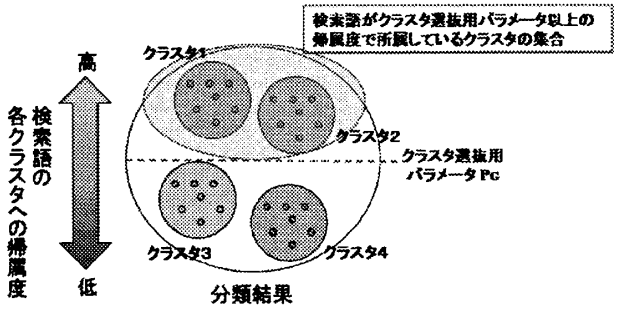


図 2. クラスタ選抜用パラメータ適用のイメージ図

#### 2.2.2 拡張語候補パラメータ

拡張語候補パラメータにより単語集合 $\mathbf{T}=\{t_1,t_2,...,t_N\}$ から拡張語候補集合 $\mathbf{T}'=\{t_j \mid u_{jk} \geq P_T, \mathbf{G}_k \in \mathbf{G}'\}$ を得る．拡張語候補パラメータ $P_T$ は０から１の間の実数値としてユーザに要求され，しきい値として利用される（図３）．

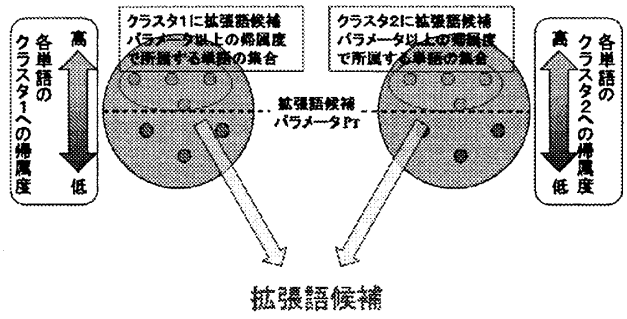


図 3. 拡張語候補パラメータ適用のイメージ図

## 3. 提案システムの検証

提案システムによる拡張語制御の性質に関する検証を以下のデータについて行った．

- 文書：「第 6 回 FIT フォーラム講演論文集」より 20 稿
- 全文書中の総単語数：1956
- クラスタ数：20
- 検索語：「アルゴリズム」，「ソフトウェア」

今回の検証では，単語の重み付けとして，文書検索の分野でよく知られている tf-idf 法を採用した．また，ファジィクラスタリング手法としては，Fuzzy c-means 法を利用した．

### 3.1 クラスタ選抜用パラメータによる選抜クラスタ数の変動

ここでは，クラスタ選抜用パラメータの変化による選抜クラスタ数の変動の検証を行った結果を示す．検証は拡張語候補パラメータを 0.9 に固定して行った．検索語「アルゴリズム」を与えたときのクラスタ選抜用パラメータの変化による選抜クラスタ数を図 4 に，検索語「ソフトウェア」を与えたときのクラスタ選抜用パラメータの変化による選抜クラスタ数を図 5 に示す．

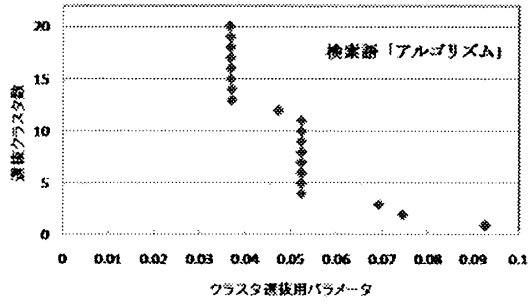


図 4. クラスタ選抜用パラメータによる選抜クラスタ数の制御①

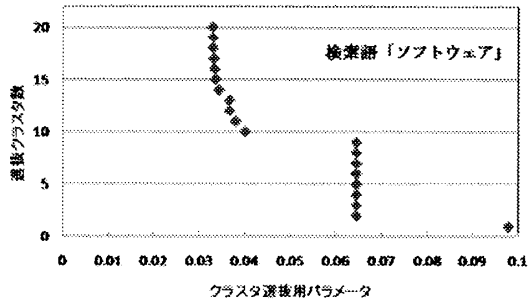


図 5. クラスタ選抜用パラメータによる選抜クラスタ数の制御②

図 4, 5 よりクラスタ選抜用パラメータを大きくしていくと，選抜クラスタ数が単調に減少していくことがわかる．

### 3.2 拡張語候補パラメータによる検索語候補数の変動

ここでは，拡張語候補パラメータの変化による拡張語候補数の変動の検証を行った結果を示す．検証はクラスタ選抜用パラメータを 0.09 に固定して行った．検索語「アルゴリズム」を与えたときの拡張語候補パラメータの変化による拡張語候補数を図 6 に，検索語「ソフトウェア」を与えたときの拡張語候補パラメータの変化による拡張語候補数を図 7 に示す．

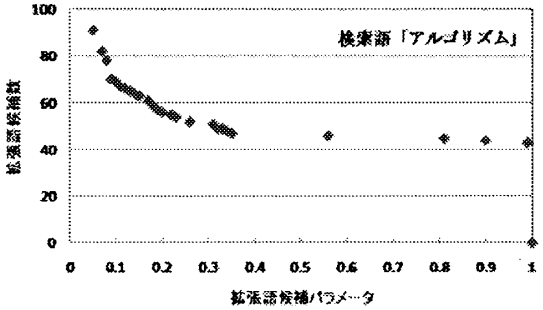


図 6. 拡張語候補パラメータによる拡張語候補数の制御①

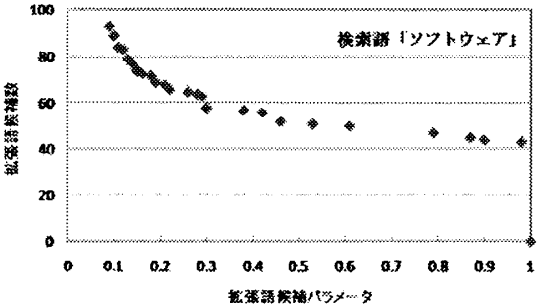


図 7. 拡張語候補パラメータによる拡張語候補数の制御②

図 6, 7 をみると拡張語候補パラメータを大きくしていくと，拡張語候補数が単調に減少していくことがわかる．

## 4. まとめ

本研究では，拡張語の所属分野や個数をユーザが制御可能な検索拡張語提示システムの提案を行った．

提案手法では，クラスタ選抜用パラメータと拡張語候補パラメータと呼ぶ 2 種類のパラメータを考えることで，拡張語の所属分野の広さと，拡張語の数の双方を制御可能とした．

実験による検証より，提案システムは検索拡張語導出パラメータを導入することで，ユーザがある程度拡張語の所属分野数と候補数を制御できることを示した．

## 参考文献

[1]北研二，津田和彦，獅々堀正幹，“情報検索アルゴリズム”，共立出版（2002）．

[2]堀幸雄，今井慈朗，中山堯，“ユーザの Web 閲覧履歴を用いた検索支援システム”，情報知識学会誌 Vol.17,No.2,pp.95-100（2007）．

[3]Ricardo Baeza-Yates，Berthier Ribeiro-Neto，“Modern Information Retrieval”，ADDISION WESLEY（1999）．