D-029

ベイズクラスタリングに基づく個人嗜好を考慮した検索語拡張手法

A Technique for Personalized Query Expansion based on Bayesian 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

インターネットの発展によって検索システムを用いた情報検索が日常的に行われるようになってきている．キーワードを利用した Web 検索のための支援手法として，ユーザによる検索語の発見を手助けする検索語拡張の研究が行われてきた[1]．現在，いくつかの Web 検索システム上で，Web 上の多量の文書集合中の語の共起確率や，多くの利用者による投入検索語の共起確率などの情報を利用した検索語拡張手法が利用されている．それらの，システム全体から得られる情報を利用する検索語拡張手法では，多量の情報を統計処理に利用可能である反面，個々のユーザに対する個人嗜好に応じた拡張語の提案が難しいという特徴がある．例として，ユーザが投入した検索語が多義語であった場合を考える．既存手法の多くでは，ユーザがどの意味で検索語を利用したかとは関係なく，検索語と同綴りの語と多く共起する語などが提案される．

これに対し，個々のユーザの閲覧文書集合や投入検索語の履歴から得られる情報を利用する手法が提案されている．例えば，安川らはユーザの嗜好を考慮した検索支援を行う手法を提案している[3]．システム中の多量の文書の中では，各語はそれぞれの文書が属する分野に依存して，多様な意味で利用されるのに対して，ユーザの履歴中では，各語はユーザの興味分野に応じた限定された意味で利用されていると考えられる．

個々のユーザの閲覧，検索履歴は，通常システム全体の文書集合と比較して小さいため，ベイズ統計の手法が有用である．安川らの手法[3]では，ベイズクラスタリングを利用して検索履歴から単語のクラスタリングを生成し，検索支援に利用する．

本研究では，従来のベイズクラスタリングとは異なり，ファジィ分割の帰属度を，ベイズ推論を用いて更新することにより，以下のような特徴を持つ検索拡張語提示システムを提案する．

- 拡張語の所属分野をユーザが制御可能
- 拡張語の個数をユーザが制御可能
- 少ない利用ログを用いてユーザの嗜好を反映した語を提案

2. 提案手法

2.1 ファジィクラスタリングを用いた検索語拡張

本研究では，著者らが先に提案した単語ファジィクラスタリングに基づく検索語拡張手法[4]で利用するための単語ファジィ分割を，ユーザの個人嗜好情報を反映して構築する手法を提案する．既存の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ことが[4]の研究の背景である．従来の検索語拡張システムでは，クリスプ集合に単語を分類するため，多義語は，その主要な出現分野での意味や用法に従っていずれか 1 つの分野（分割）にのみ所属することになる．また，従来のシステムでは，提案される拡張語の個数をユーザが変更することは難しい．そこで，[4]の研究では，ユーザが拡張語の個数と所属分野を制御できる検索語拡張システムを提案した．

単語のファジィクラスタリングに基づく検索語拡張手法[4]は，事前に与えられた文書集合から抽出した単語を，各単語の属する分野を反映するようなファジィクラスタに分類し，得られた結果に対して，ユーザから与えられた検索語と 2 つのパラメータを用いることで検索語拡張を実現する．

この手法は，事前に与えられた文書集合からファジィ分割を構成する前処理部分と，実際にユーザに拡張語を提示する拡張語選出処理部分の大きく 2 つの部分処理に分けることができる．

前処理部分は，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である．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 のような帰属度行列を得る．このようにして構成されたファジィ分割は各単語が主に利用される分野に関する情報を反映していると考える．つまり，各クラスタは分野の近似であると考える．ファジィ分割では，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 2 つのパラメータによってクラスタ（分野）数，単語（拡張語）数を制御することができる．

拡張語選出処理部分は，実際にユーザが検索語を入力した際に，それをもとに拡張語を提示する部分処理である．ここでは，入力された検索語に加え，2 つのパラメータを操作することで前述の分割結果から拡張語候補を選抜する．

今，ユーザによって検索語 t_i が与えられた場合を考える．まず，1 つ目のパラメータであるクラスタ選抜用パラメータにより検索語 t_i 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．この操作により，拡張語の分野を制御することができる．次に，2 つ目のパラメータの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．この操作によって，拡張語の個数を制御することができる．クラスタ選抜用パラメータ P_G，拡張語候補パラメータ P_T はともに 0 から 1 の間の実数値としてユーザが入力する値であり，しきい値として利用される．

[4]の手法では，事前に与えられる文書集合にユーザの利用履歴などを用いることでユーザ個人の嗜好を考慮した検索語拡張を行うことができる．しかし，ユーザの利用履歴

† 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

情報が十分に蓄積されていない状況では，ファジィ C-means 法などのファジィクラスタリング手法では，適切なクラスタを構成することは困難であるため，十分な嗜好の考慮ができない．そこで，本研究では，ベイズ推論を用いることで，ユーザの利用ログでクラスタの更新を行い，少量の利用ログからユーザの嗜好を考慮した検索語拡張を行う手法を提案する．

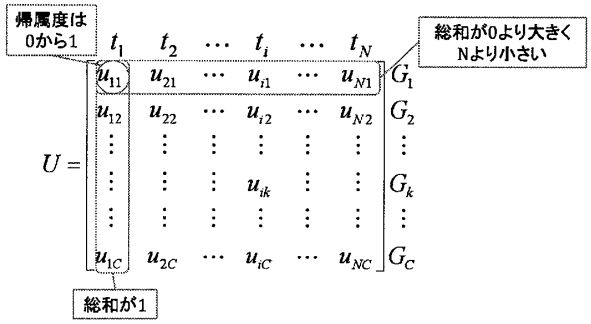

図 1. ファジィ分割（ここで，単語集合 $\mathbf{T}=\{t_1,t_2,\ldots,t_i,\ldots,t_N\}$，クラスタ集合 $\mathbf{G}=\{G_1,G_2,\ldots,G_k,\ldots,G_C\}$）

2.2 初期クラスタ生成

本研究では，初期クラスタとして事前にシステム全体のログから得られる情報を利用してクラスタリング[4]を行った結果を用いる．この結果はファジィ分割となり図 1 のようになる．このように初期クラスタを生成することにより，ユーザの利用履歴が全く得られていない状況であっても，検索システムは正常に動作する．

このとき，本提案手法でベイズ推論を利用するにあたり，「帰属度 u_{ik}」を「単語 t_i の所属するクラスタが G_k である確率」と考える．u_{ik} は単語 t_i がクラスタ G_k 所属する度合いであり，かつ 0 から 1 の値であるので確率分布と同様に扱うことができる．

2.3 ベイズ推論による帰属度の更新

本研究のベイズ推論による帰属度の更新は以下の手順で行われる．前述の方法により構成された初期クラスタ $\mathbf{G}$ が得られているものとする．ユーザが文書 d を閲覧した状況を想定し，d より得られる情報を $\mathbf{G}$ に反映する方法を考える．本研究では，文書はベクトル $d=(d[0],d[1],\cdots,d[i],\cdots,d[N])$ により表されるものとする．$d[i]$ は文書 d に単語 t_i が出現するときに 1，それ以外のとき 0 とする．ここでは，クラスタ G_k と単語 t_i（但し，$d[i]=1$）に着目し，u_{ik} が更新される例を示す．$d[i]=1$ である全単語に対し，式（1）により $P(t_i\in G_k)$ すなわち u_{ik} を更新することになる．単語 t_i について $P(d[i]=1|t_i\in G_k)$ すなわち式（2）により近似的に得られる値と u_{ik} との積により，更新後の u_{ik}' すなわち $P(t_i\in G_k|d[i]=1)$ を求める．この作業を文書の入力がある度に全てのクラスタについて行う．

$$P(t_i\in G_k|d[i]=1)=\frac{P(d[i]=1|t_i\in G_k)P(t_i\in G_k)}{\sum_{l=1}^{c}\left(P(d[i]=1|t_i\in G_l)P(t_i\in G_l)\right)} \quad (1)$$

$$P(d[i]=1|t_i\in G_k)=\frac{\sum_{i=1}^{N}u_{ik}d[i]}{\sum_{i=1}^{N}u_{ik}} \quad (2)$$

式（1）では，$P(t_i\in G_k)$ は単語 t_i の所属するクラスタが G_k である確率を表し，$P(t_i\in G_k|d[i]=1)$ は単語 t_i を含む文書 d が選ばれた後で，単語 t_i の所属するクラスタが G_k である確率を表す．$P(t_i\in G_k)$ が，文書 d が選ばれたという結果を得たことにより $P(t_i\in G_k|d[i]=1)$ に更新されたと言うことになる．$P(t_i\in G_k)$ を事前確率，$P(t_i\in G_k|d[i]=1)$ を事後確率と呼び，これらを確信度という．ファジィ分割に話を焼き直すと，事前確率 $P(t_i\in G_k)$ は更新前の帰属度 u_{ik} のことである．事後確率 $P(t_i\in G_k|d[i]=1)$ はユーザが文書 d を選んだことによって更新された帰属度 u_{ik}' を表す．確信度はユーザが文書 d を選んだという結果が得られるたびに，更新されていくことになる．更新の際，各単語の全クラスタへの帰属度の総和（図 1 の各列の和）が，1 になるように正規化を行う．

$P(d[i]=1|t_i\in G_k)$ は式（2）で表される．$P(d[i]=1|t_i\in G_k)$ は単語 t_i がクラスタ G_k に属するという事象が観測されたときに，文書 d に単語 t_i が含まれている確率で，尤度関数に相当する．単語がクラスタにクリスプに所属している場合を考えると，$P(d[i]=1|t_i\in G_k)$ は，クラスタ G_k に所属する単語 t_i が文書 d に出現する確率と考えることができる．本研究では，単語はクラスタにファジィに所属する（図 1）ため，$P(d[i]=1|t_i\in G_k)$ の値を近似式（2）により推定する．式（2）では，文書 d に出現するクラスタ G_k に所属する単語の個数は，各単語の G_k への帰属度と文書 d における出現（0 または 1）の積であると考え，クラスタ G_k に所属する単語 t_i が文書 d に出現する確率を推定する．

3. 今後の課題

今後は，ベイズ推論による帰属度の更新を実装し，先の研究結果と併せて評価実験を行う予定である．また，今回は，単語の文書への出現情報を出現するかしないか（0 または 1）で表現したが，単語の出現数を利用することにより，より詳細なクラスタ生成が可能であると考えられる．その場合に，ベイズ推論にどのような尤度関数が利用可能であるか考える必要がある．

参考文献

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

[2] 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）.

[3] 安川美智子，横尾英俊，“クエリログから獲得した関連語のクラスタリングに基づく Web 検索支援”，電子情報通信学会 Vol.J90-D, No.2, pp.269-280（2007）.

[4] 鈴木裕規，能登谷淳一，草苅良至，笠井雅夫，“ファジィクラスタリングを用いた検索語拡張手法”，第 7 回情報科学技術フォーラム，第 2 分冊，pp.89-90（2008）.

[5] Makoto Iwayama, Takenobu Tokunaga, “A Probabilistic Model for Text Categorization: Based on a Single Random Variable with Multiple Values”, Proc. Applied NLP, pp.162-167（1994）.

[6] Makoto Iwayama, Takenobu Tokunaga, “Hierarchical Bayesian Clustering for Automatic Text Classification”, Proc. IJCAI95, pp.1322-1327（1995）.