人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第5回)
SIG-AM-05-07

複数文書の相対的特徴可視化による理解支援

A Mulit-Document Relative Characteristics Visualization for Understanding Support

薦田和弘 [1*]　大澤幸生 [1]

Kazuhiro Komoda[1], Yukio Ohsawa[1]

[1] 東京大学工学系研究科
[1]School of Engineering, the University of Tokyo

Abstract: We propose a visualization method showing relative characteristics of multiple documents which have different contexts. We provide users with an interactive graph correlating common words in multiple documents with characteristic words in each document, using co-occurrence and context similarity information. By allowing the users to discover underlying differences between superficially similar documents and to become conscious of them as new viewpoints, we intend to support understanding in terms of cooperative group activities or individual information collections. We examine our method using a document set with context information.

1　はじめに

文書からの知識発見は，学術界や産業界における様々な領域で重要な役割を果たしている．従来，文書内に明示的に出現する事実やその関係性を抽出することで，学術やビジネスにおいて実用上の問題解決が試みられてきた[1]．一方で，文書の著者（情報提供者）の側に立てば，事実関係を示すことに加えて，数ある類似文書との関連性を意識しながら自分の主張の特徴点を把握し，新たな知見を適切に位置づけることによって自分の意図を読者（情報の受け手）に理解してもらう必要がある．例えば，学術論文では，著者独自の問題意識に端を発する数々の試行錯誤をもとに，最新の研究成果に基づく主張が表現される．この主張を他の主張から差別化し，同時に他の主張との関連性を示す特徴点を明確にしてはじめて，学術分野において主張を適切に位置づけ，既存の知見と新たな知見の結合を促すことができる．

文書単独の分析によって得られない相対的な特徴や関連性に対して，2 つ以上の文書集合を組み合わせ，集合間の差に相当する概念を抽出する差異分析が有効である．既存の研究においては，集合間に明確で一方向的な対立関係が想定される場合，あるいは，評価対象とその属性等（評価視点）[2]の様に，集合同士をグループ化して区別する指標が明確な場合が対象とされた．

一方で，本稿では，ユーザの目的に応じて，着目する文書集合D_1と，比較対象としての文書集合D_2を指定する．D_1とD_2を完全に区別することは目的ではなく，両者の間の関連性も考慮したい．この時，D_1をD_2から差別化し，同時にD_2との関連性を示すキーワードの抽出と可視化を行う．例えば，ユーザが特定のテーマを掲げる学会への論文投稿を検討する際，「ユーザが投稿する論文」をD_1，「学会における関連論文集合」をD_2とすることで，ユーザの立場から，他の研究との差分を特徴的に示しつつ当該学会との関連性を失わないキーワードを重要だとみなして抽出することが可能である．また，そのようなキーワードと，D_1とD_2で共通で使用される単語の関係を可視化図で対話的に示すことができれば，投稿する論文において読者に対して強調すべき点を取捨選択できる等の利点がある．

本稿では，着目する文書集合D_1と比較対象D_2の間の相対的特徴を可視化する手法を提案する．本稿における貢献は以下の様にまとめられる．

・D_1内の文書d_kに特徴的なキーワードを抽出する際，D_1をD_2から差別化し，同時にD_2との関連性を示すような単語が重要と考え，単語の特徴量を計算する点．
・上記の特徴量を計算する際，対数尤度比に基づく尺度に，文脈類似度を導入したもの（3 節）を適用し，その有効性を検証する点．
・D_1, D_2の注目語，特徴語（3.3 節）を文共起情報によって関連づけ，両者を可視化する対話的環境を提供する点．

本論文の構成は以下の通りである．まず，2 節で関

* 連絡先：東京大学工学系研究科技術経営戦略学専攻
〒113-8656 東京都文京区本郷 7-3-1
E-Mail: kazuhirokomoda@gmail.com

連研究について述べる．3 節で提案手法を説明し，4 節で評価実験について述べ，5 節で可視化図の観察を行う．6 節で考察を行い，7 節で結論を述べ本稿をまとめる．

2 差異に着目した分析と可視化

文書集合の差異分析は，カテゴリ固有の問題特定や, 経験知の発見・共有につながる点で有益であり，様々な関連研究が行われてきた．文書集合間の差を求める手法として，事前に定義したカテゴリに基づき，各観点に属する概念を辞書として準備する方法[3]がある．例えば，製造業のコールセンターでは部品名，苦情，要望，質問といった観点とそれぞれに該当する表現を事前に集め辞書に登録し，分類された文書集合間の関係を概観する方法が用いられている．この場合，事前に固定された分析観点以外での柔軟な分析が行えないという課題がある．また特定の文書集合に特徴的な表現を抽出する方法[4]がある．これらの手法では，特定の文書集合を他の文書集合から差別化することができる単語を特徴的と判断するため，両者の間の関連性を考慮していない．

大平ら[5]は，議論参加者の共有知識を増やし，相互理解の構築を目指すという目的で，各参加者が作る「人P_iがオブジェクトO_jをI_kと考える」という組を平面に配置し差異を可視化した. この場合には, 個々人の意見の差異が簡潔な外部表現として得られなければならないという制約がある．

3 提案手法

複数文書の相対的特徴を反映する単語の抽出と可視化を行う提案手法の詳細を記述する．

3.1 事前準備

まず，ユーザが着目する文書集合D_1と，比較対象としての文書集合D_2を，重複のないように（$D_1 \cap D_2 = \phi$）用意する．例えば，ユーザが特定のテーマを掲げる学会への論文投稿を検討する際,「ユーザが投稿する論文」をD_1,「学会における関連論文集合」をD_2とすることで，ユーザの立場から，他の研究との差分を特徴的に示しつつ当該学会分野との関連性を失わないキーワードを抽出して可視化する状況を想定する．文書集合全体$D \equiv D_1 \cup D_2$に対して不要語・語尾の除去，品詞の選択（名詞・動詞・形容詞・副詞）を行う．

文書集合D, D_1, D_2で用いられる語彙の集合U, U_1, U_2について，共通部分$S = U_1 \cap U_2$に属する語（「共通語」と定義する）は，D_1とD_2において表層的に共有される主張に関連する単語を含み，一方で差集合$U_1 - S, U_2 - S$に属する語（「特有語」と定義する）は, D_1, D_2それぞれの特徴的な主張に関連する単語を含むと考えられるため，以下で「共通語」と「特有語」の一部の関係を可視化する．

3.2 単語の特徴量の計算

着目する文書集合D_1に含まれる文書d_kに対して，文書集合全体Dの語彙の各単語α_jの特徴量を計算する．今回は，単語の特徴量を決定する方法として，対数尤度比に基づく尺度LLR^+を提案する．これは，特定分野における単語の特徴度を測る尺度[6]を参考に，D_2において単語α_jが出現する際の文脈類似度c_simを新たに考慮したものである．

まず，D_1, D_2内の文書で出現する単語を（重複を許して）順に並べ，単語トークン系列$v_1, \cdots, v_{n_{D_1}}, v_{n_{D_1}+1}, \cdots, v_n$を作成する. $n = n_{D_1} + n_{D_2}$であり, n_{D_1}, n_{D_2}はD_1, D_2の単語トークン数である. 次に, 文書$d_k (k = 1, \cdots, |D_1|)$と単語$\alpha_j (j = 1, \cdots, |U|)$が与えられた時，ある単語トークン$v_i (i = 1, \cdots, n)$に対応する確率変数$W_j^i, T_k^i$の値をそれぞれ$w_j^i, t_k^i$とし，以下の様に定義する．

$$w_j^i = \begin{cases} 1 \ (v_i = \alpha_j) \\ 0 \ (otherwise). \end{cases}$$

$$t_k^i = \begin{cases} 1 \ (v_i \text{は} D_1 \text{内で出現}) \\ 1 \ (v_i \text{は} D_2 \text{内で出現}, v_i = \alpha_j, c\_sim(v_i, d_k) \geq \theta_{sim}) \\ 0 \ (otherwise). \end{cases}$$

$c_sim(v_i, d_k)$は，集合間の Jaccard 係数を用いて以下の様に定義する．

$$c\_sim(v_i, d_k) = \max_l \left(\text{Jaccard}\left(ctx(v_i), ctx(v'_l)\right) \right).$$

ただし，トークンv_iの文脈$ctx(v_i)$を，v_iが出現する文で使用される単語の集合（ただし単語v_i自身は除く）と定義し，$d_k \in D_1$においてv_iと同じ値を持つ単語トークン$v'_l (l = 1, \cdots, L)$の文脈を，$ctx(v'_l)$ $(l = 1,2, \cdots, L)$とする．θ_{sim}は類似度の閾値である．以上で定義された変数の関係を表 1 に示す.

この時，それぞれの単語トークンv_iが確率的に独立であると仮定すると，系列$\Upsilon_{jk} = \langle w_j^1, t_k^2 \rangle, \langle w_j^2, t_k^2 \rangle, \cdots, \langle w_j^n, t_k^n \rangle$の生起確率は以下で定義される．

$$Pr(\Upsilon_{jk}) = \prod_{i=1}^{n} Pr(W_j^i = w_j^i, T_k^i = t_k^i).$$

また，単語トークンv_iについて，確率変数W_j^i, T_k^iの値

表 1: 文書d_k ，単語α_jに関する確率変数W_j^i, T_k^i．?は，$c_sim(v_i, d_k) \geq \theta_{sim}$ならば1

	D_1	D_2
i	$1, \cdots, n_{D_1}$	$n_{D_1}+1, \cdots, n$
v_i	$v_1, \cdots, v_{n_{D_1}}$	$v_{n_{D_1}+1}, \cdots, v_n$
W_j^i	010100010101	0100010000000000
T_k^i	111111111111	0?000?0000000000

表 2: 単語トークンv_iについての集計結果

	$T_k^i = 1$	$T_k^i = 0$
$W_j^i = 1$	a	b
$W_j^i = 0$	c	d

で出現頻度を集計したものが表 2 である．ここで，$a+b+c+d=n$である．文書d_kにおける単語α_jがD_1をD_2から差別化し，同時にD_2との関連性を示すならば，W_j^iとT_k^iの依存度が高まるという意図の下，表 2 の集計を行った．

表 2 に基づき，仮説H_{indep}「確率変数W_j^iとT_k^iとは互いに独立である」に対し，次の対数尤度比$LLR_0^+(d_k, \alpha_j)$を考える．

$$LLR_0^+(d_k, \alpha_j) = \log \frac{Pr(\Upsilon_{jk})}{Pr(\Upsilon_{jk}|H_{indep})}$$

$$= \sum_{i=1}^{n} \log \frac{Pr(W_j^i = w_j^i, T_k^i = t_k^i)}{Pr(W_j^i = w_j^i, T_k^i = t_k^i | H_{indep})}.$$

上式において$Pr(W_j^i = w_j^i, T_k^i = t_k^i | H_{indep})$は$H_{indep}$が成立するとした場合の$Pr(W_j^i = w_j^i, T_k^i = t_k^i)$であり，その値は$Pr(W_j^i = w_j^i) Pr(T_k^i = t_k^i)$に等しい．表 2 に基づく推定により，以下の値が計算できる．

$$Pr(W_j^i = 1, T_k^i = 1) = \frac{a}{n}, Pr(W_j^i = 1, T_k^i = 0) = \frac{b}{n},$$

$$Pr(W_j^i = 0, T_k^i = 1) = \frac{c}{n}, Pr(W_j^i = 0, T_k^i = 0) = \frac{d}{n},$$

$$Pr(W_j^i = 1) = \frac{a+b}{n}, Pr(W_j^i = 0) = \frac{c+d}{n},$$

$$Pr(T_k^i = 1) = \frac{a+c}{n}, Pr(T_k^i = 0) = \frac{b+d}{n}.$$

$LLR_0^+(d_k, \alpha_j)$は，確率変数W_j^iとT_k^iの依存性の度合いが高いほど大きな値を取るため，以下の様に補正した$LLR^+(d_k, \alpha_j)$を用いることで，d_kにおいてD_1をD_2から差別化し，同時にD_2との関連性を示す単語α_jに対して高い特徴量を与えることができる．

$$LLR^+(d_k, \alpha_j) = sign(ad - bc) LLR_0^+(d_k, \alpha_j).$$

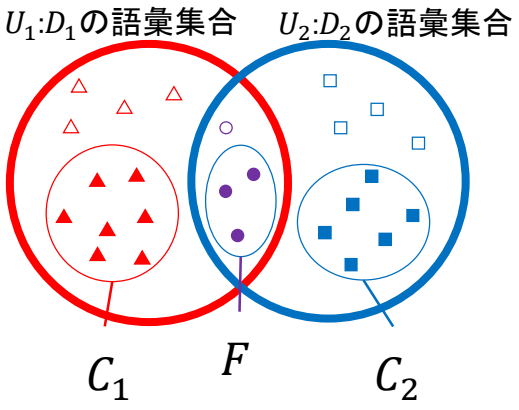

図 1: 「共通語」「特有語」と「注目語」「特徴語」の関係．赤い円は文書集合D_1の語彙集合U_1，青い円は文書集合D_2の語彙集合U_2を表す．共通語から選ばれた特徴量上位の単語が注目語（紫で塗られている），特有語から選ばれた特徴量上位の単語が特徴語（赤あるいは青で塗られている）である．

3.3 可視化する注目語と特徴語の選択

共通語と特有語の中から，可視化に使用する単語を選ぶ．まず，3.2 節により，D_1に含まれる各文書d_kに対して，各単語$\alpha_j \in U$の特徴量を計算する．次に，D_2に含まれる各文書d_kに対しても特徴量を計算するため，D_1とD_2を入れ替えて 3.2 節を再度行う．以上で，全文書が全単語の特徴量情報を持つため，以下用いる単語の特徴量は，各文書における特徴量の和とする．共通語から特徴量上位のものを指定個数選んだものを「注目語」，D_1, D_2の特有語から特徴量上位のものをそれぞれ指定個数選んだものを「特徴語」と定義する．以下，注目語集合をF，文書集合D_1, D_2の特徴語集合をそれぞれC_1, C_2とする．以上の関係を図 1 に示す．

3.4 注目語と特徴語の文共起情報

ある注目語$f \in F$が文書集合D_1において出現したものをfのトークン(token)と呼び，f'とする．f'が出現する同一文中に，特徴語$c_1 \in C_1$のトークンc_1'が存在すれば，$(f, c_1 n_1(f, c_1))$の組を作成する．$n_1(f, c_1)$の値は 3.3 節で求めた特徴語c_1の特徴量とする．注目語fと特徴語c_1に関するグラフを作成するため，f

とc_1の間のリンクの重みを以下で計算し，重みに応じた太さを持つリンクを描画する．

$$w(f,c_1)=\frac{|c_1|n_1(f,c_1)}{\sum_{\widehat{c_1}}|\widehat{c_1}|n_1(f,\widehat{c_1})}.$$

ただし，$|c_1|$は特徴語c_1のD_1におけるトークン数，$\sum_{\widehat{c_1}}|\widehat{c_1}|n_1(f,\widehat{c_1})$は$f$を含む全ての組の重みの総和である．以上を文書集合$D_2$，特徴語集合$C_2$についても同様に行い，合わせて描画することで，ある注目語$f\in F$について，注目語fと特徴語の文共起情報に基づくグラフが得られる．このグラフは，注目語が，それぞれの文書集合においてどのような特徴語と共に用いられるかを示す．文書集合間の共通語のみに着目した従来の要約・集約において看過されやすい，配慮すべき重要な論点の展開の違いに焦点を当てている．

4 実験

4.1 実験仕様

3.2 節の特徴量計算を評価するため，論文タイトルに含まれるべき単語を論文要旨から推測する評価実験を行った．論文のタイトルは，他の研究者が最初に注目する重要な部分であり，既存研究との差分を特徴的に示しつつ，分野における一般性や関連性を失わない簡潔な言語表現が求められる．このような言語表現は，論文要旨を，関連する他の論文要旨と比較した際に抽出できる場合が多いと考えた．今回は，2013 年度人工知能学会全国大会「自然言語」セッションに投稿された 33 論文要旨d_k $(k=1,\cdots,33)$を利用し，前処理後の要旨情報から「他の論文との差分を特徴的に示しつつ，セッションとの関連性を示す(*)」単語を 3.2 節の手法で抽出した．3.1 節において，ある 1 本の論文要旨d_kをD_1，残りの 32 本の論文要旨をD_2とすることで，d_kについて単語を特徴量によってソートしたリストを得て(retrieved)，対応する論文タイトルから(*)を満たす単語を人手で選択した(relevant)．θ_{sim}は経験的に 0.22 と設定した．なお，比較手法LLRは，t_k^iの定義式において右辺 2 行目を考慮しない．

評価指標として，$k=1,\cdots,33$の 33 論文の Mean Average Precision(MAP)を算出した．MAPは以下の式で定義される．

$$MAP=\frac{1}{|D|}\sum_{d_k\in D}\frac{1}{|\Omega_k|}\sum_{\alpha_j'\in\Omega_k}Precision\left(rank\left(d_k,\alpha_j'\right)\right).$$

ここで，Ω_kは論文要旨d_kのタイトルから人手で抽出した(*)を満たす語の集合（正解データ）であり，

表 3：実験結果（人手による(*)の選定）

手法	MAP
LLR^+	0.281
LLR	0.278

$rank(d_k,\alpha_j')$は文書$d_k\in D$において単語$\alpha_j'\in\Omega_k$が手法によって得た順位，$Precision\left(rank\left(d_k,\alpha_j'\right)\right)$はその順位までの出力結果における適合率である．この評価指標では，各文書d_kについて，手法による順位づけ結果の上位$|\Omega_k|$単語の中にΩ_k内の単語が全て含まれる場合が最も望ましいとされ，この時 Average Precision の値は 1 である．

4.2 実験結果

実験の結果を表 3 に示す．MAP値に関して，全体的な性能の向上が確認できた．実際，1 つの論文を除いて Average Precision 値が不変または増加している．以下，特徴的な個別事例の詳細を確認する．

表 4 は，LLR^+とLLRにおけるAP値の差が最大$(0.372-0.268=0.104)$となった事例である$(k=3)$．LLR^+では，文書の潜在トピックを抽出する手法である Latent Dirichlet Allocation(LDA)に関する単語をより上位で抽出でき，「自然言語」セッションへの関連性を示したと言える．また，LLRで上位に来ている tag や put という単語は，当該論文における具体的な処理に関する単語である．他の論文との関連性が低い単語の順位を下げることができたために，比較的良い結果が得られたと考えられる．

表 5 はLLR^+とLLRにおけるAP値の差がないが，値が最も高い(0.603)事例である$(k=19)$．非常に少数の文によって提案内容のみが的確に表現されており，かつ「完備束表現」という表現自体が「自然言語」セッションにおいては非常に特徴的な表現だったため，文脈類似度が設定した閾値θ_{sim}を上回りにくく，LLR^+とLLRで同じランキング結果となった．タイトルに含まれるべき単語が上位に抽出されており，LLRによって得られる Average Precision が最も高かった．

表 6 は，LLR^+とLLRにおけるAP値の差が最小$(0.250-0.293=-0.043)$となった事例である$(k=9)$．33 論文要旨の中で唯一 Average Precision が低下した．LLRでは$|\Omega_k|=4$のうち 3 単語が上位 4 単語に含まれているが，LLR^+ではΩ_kのうち 2 単語が低い順位を得ているため，僅かではあるが Average Precision が低下した．

表 4：$k = 3$における手法ごとの特徴量上位語を順に並べたリスト．下線は$\Omega_{k=3}$に含まれる単語．

$\Omega_{k=3}$	{supervis, alloc, dirichlet, latent, pseudo, label}
LLR^+	latent, document, llda, alloc, dirichlet, label, topic, lda, tag, put, estim, limit, pseudo, abil, exce, supervis
LLR	document, llda, label, latent, tag, put, topic, estim, limit, pseudo, abil, exce, supervis, applic, alloc, dirichlet

表 5：$k = 19$における手法ごとの特徴量上位語を順に並べたリスト．下線は$\Omega_{k=19}$に含まれる単語．

$\Omega_{k=19}$	{represent, languag, lattic, natur, data, structur, complet}
LLR^+, LLR	represent, complet, bag-of-word, add, data, structur, lattic, mathematic, defin, order, …

表 6：$k = 9$における手法ごとの特徴量上位語を順に並べたリスト．下線は$\Omega_{k=9}$に含まれる単語．

$\Omega_{k=9}$	{topic, domain, user, knowledge}
LLR^+	latent, user, knowledge, alloc, dirichlet, difficulti, domain, term, lda, captur, technic, topic
LLR	user, knowledge, difficulti, domain, term, captur, technic, depend, key, background, level, belong, adapt, focu, experiment, show, previou, person, lot, determin, inform, topic

5　可視化図の観察

公開報告書の様な文書は，組織内外の人間に広く読まれることを意図しているにも関わらず，分野の専門知識を前提として記述され，読み手にとっては複雑な構造を持つ傾向にある．専門家が，難解な報告書を非専門家である読み手に分かりやすく伝えることを考える際，非専門家にとってより身近な資料を用意して両者の位置づけを比較することで，有益な指針が得られると考えられる．日本原子力安全部会「福島第一原子力発電所事故に関するセミナー第 8 回」議事メモの一部（東北地方太平洋沖地震の概況及び福島第一原子力発電所の事故の概要）をD_1，「福島第一原発　原子力安全」というキーワードで検索して得られた Yahoo!ニュース記事数件をD_2とした時，3 節の処理を経た可視化図を図 2 として掲載する．特徴量計算によって選択された注目語，特徴語が上から順に並んでいる．

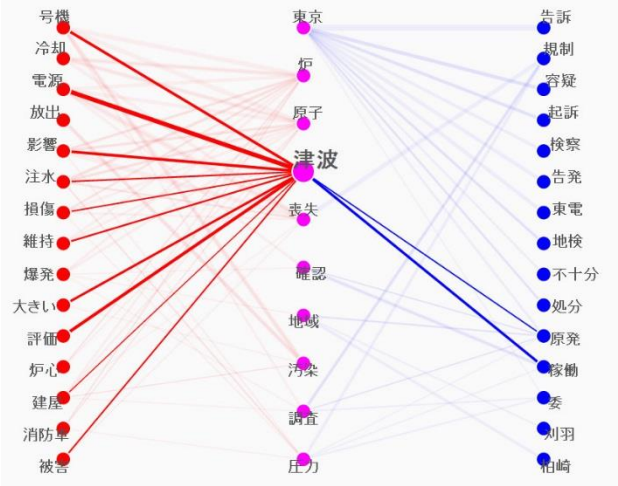

図 2:「津波」にマウスを載せてフォーカスする．

なお，図 2 は Web ブラウザ上での閲覧・操作を意図している．一度に表示されるリンクの数を重みの閾値によって変更でき，最も重要なリンクから優先して確認することができる．また，ノード（＝単語）の上にマウスを載せると，そのノードがハイライトされ，関連するリンクのみが表示されるため，特定の単語に容易に着目できる．

図 2 では，注目語，特徴語自身に加え，それぞれの単語が他のどのような単語と共起しているかが読み取れる．例えば，D_1に出現する特徴語の多くが「炉」や「原子」といった単語と共起しているが，一方でD_2においてそれらの単語は特徴語との共起は多くなく，リンクは「東京」に集中している．D_1とD_2を直接結びつけている「津波」に関しては，注目語であってもその使われ方が大きく異なることが推測できる．D_1を執筆する立場としては，D_2の特徴語との関連性が強い話題を強調して説明し，一方で，D_2を収集する立場としては，例えば，刑事告発に関する記事を減らして原子炉や汚染に関する記事を増やすなど，よりD_1との関連性が強くなるような関連情報を収集して再度可視化を行うような，チャンス発見[7]の二重らせんプロセスを行うことが可能である．

6　考察

4 節の論文タイトル課題において，提案手法LLR^+は比較手法LLRに対して，全体的な性能の向上が確認された．4 節の個別事例に留まらずより一般的に，提案手法と比較手法の両方における性能，そして提案手法に対する比較手法の性能を向上させるために考慮すべき要素を検討する．

課題全体の性能を向上させるため，課題に沿った適切な前処理を行うことが望ましい．まず，複合語・

類義語を考慮に入れる必要がある．具体的には”Latent Dirichlet Allocation”の様な単語列を複合語として扱い，さらに”LDA”の様な省略形と類義語であるという情報を与えることに相当する．また，現在は，名詞・動詞・形容詞・副詞を取り扱っている．これは，文中で強い意味内容を持ち他の単語との関連性が高い可能性のある単語を網羅するという意図の下行った処理であるので，今後，不要な単語を取り除くことが可能である．

提案手法を比較手法に対して向上させるためには，タイトルに含まれるべき単語の候補を絞ることが望ましい．論文のタイトルに含まれる単語は全てが同じ重要度を持つわけではなく，単語自体の機能，著者が命名において重視する観点に応じて，タイトルに含まれるべき単語は変化するため，単語の重み情報を取り入れた評価を行うことができる．実際，今回は，前処理では取り除かなかったが，「他の論文との差分を特徴的に示しつつ，セッションとの関連性を示す(*)」を満たさない単語を人手で除去した．これは(*)を満たす単語に1，満たさない単語に0の重みを与えることに相当する．今後は，タイトルに不適切な語の除去だけでなく重要語の強調も行い，また一人の評価者ではなく，論文の著者に依頼して得られた重み情報による実験を行うことが考えられる．

5 節の可視化図においては，まず全語彙を特徴量で順位づけし，共通語集合，特有語集合に属する単語から特徴量上位のものを選んで可視化図に出力し（注目語と特徴語），その後に注目語と特徴語の文共起情報を用いてリンクを結んだ．したがって，現状では単語の順位は全て，3 節で求めた特徴量に依存している．また，文書集合間の特徴語は，必ず図の中心にある注目語を媒介として結びついているように可視化されている．したがって，注目語については，文書集合間の特徴語（ピンクノード）を関連付ける度合いによって選定する方が望ましい可能性がある．あるいは，3 節で特徴的かつ関連性を示す単語を抽出し，その際に文脈類似度を導入したことを考えると，その結果を生かし，文書集合間の特徴語同士（赤ノードと青ノード）をより直接的に結びつけるような可視化によって，ユーザが意外な関連性に気づくことができる可能性がある．その実現のため，文脈類似度c_simの計算方法と閾値設定も課題である．今回は Jaccard 係数を用いて定義したが，cosine 類似度や自己相互情報量など別の方法との比較も検討したい．

7 結論

本稿では，着目する文書集合D_1と比較対象D_2の間の相対的特徴を可視化する手法を提案した．D_1に特徴的な単語を計算する際，対数尤度比に基づく尺度に文脈類似度を導入してD_2との関連性を示す単語を高く評価する点，注目語と特徴語の文共起情報に基づく対話的可視化環境を提供する点が特徴である．ある論文要旨に特徴的で，他の論文要旨集合との関連性を示すような単語を抽出して論文タイトルを再現する実験を行い，提案手法の有効性を示した．また，2 つの文書集合における「注目語」と「特徴語」の間の文共起情報を用いた対話的インタフェースを提供し，専門的な報告書と関連するニュース記事を題材に，観察結果を報告した．

参考文献

[1] Mack, R. and Hehenberger, M.: Text-based knowledge discovery: search and mining of life-sciences documents, Drug Discovery, Vol. 7, No. 11, pp. S89-S98, (2002)

[2] 乾孝司, 板谷悠人, 山本幹雄, 新里圭司, 平手勇宇, 山田薫: 意見集約における相対的特徴を考慮した評価視点の構造化, 自然言語処理, Vol. 20, No. 1, pp. 3-26, (2013)

[3] 那須川哲哉: コールセンターにおけるテキストマイニング, 人工知能学会学会誌, Vol. 16, No. 2, pp. 219-225, (2001)

[4] Hisamitsu, T. and Niwa, Y.: Topic-Word Selection Based on Combinatorial Probability, NLPRS, Vol. 1 (2001)

[5] 大平雅雄, 山本恭裕, 蔵川圭, 中小路久美代: EVIDII: 差異の可視化による相互理解支援システム, 情報処理学会, Vol. 41, No. 10, pp. 2814-2826, (2000)

[6] 内山将夫, 中條清美, 山本英子, 井佐原均: 英語教育のための分野特徴単語の選定尺度の比較, 自然言語処理, Vol. 11, No. 3, pp. 165-197, (2004)

[7] Ohsawa, Y. and McBurney, P.: Chance Discovery, Berlin-Heidelberg, Springer Verlag, (2003)