小特集「国際会議で見つけたオススメ論文」

# The 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2006)

大原 剛三 大阪大学産業科学研究所
Kouzou Ohara The Institute of Scientific and Industrial Research, Osaka University.
ohara@ar.sanken.osaka-u.ac.jp

**Keywords:** information retrieval, relevance feedback, clustering.

## 1. 論文の背景

今日，膨大な情報がインターネット上に蓄積されているが，検索サイトを利用しても目的とする情報になかなか辿り着けないという話をよく耳にする．このような問題を解消するための手法はWorld Wide Web（WWW）の普及当初から数多く研究されており，その代表的なアプローチとして適合性フィードバック（relevance feedback）がある[徳永 99]．適合性フィードバックを用いたWeb検索の多くは，現在の検索質問によって得られた検索結果を利用者に提示し，利用者から与えられる各Webページ，もしくは各Webページの索引語の検索要求に対する適合性に基づき，より良い検索結果が得られるよう検索質問を拡張する．

しかしながら，一般にWebページの作成者は同一のWebページ中に異なったサブトピックに関する複数の文書を混在させる．そのため，サブトピックを意識せずWebページ全体を一つの単位として扱う既存手法は検索質問の拡張に失敗する場合がある．紹介する論文[Yoo 06]では，この問題を解消するために，検索結果のWeb文書をサブトピックごとにクラスタリングし，クラスタを代表する索引語を用いてより適切な検索拡張を実現することを目的としている．

## 2. 提案手法の概要

本論文では，Webページ作成者の意図する話題構造はレイアウト構造に反映されており，それに従いWebページを分割することでサブトピックの文脈をより特徴づける索引語を抽出することができる，という考えに基づき，(1) レイアウトに基づいたWebページの分割，(2) 文脈を考慮した索引語の選択，(3) Webページのクラスタリングについてそれぞれ3.1，3.2，3.3節で独自の手法を提案している．以下，それらの内容について補足と紹介者の意見を交えながら概説する．

(1) レイアウトに基づいたWebページの分割

本論文では，Webページの部分文書（セグメント）間の話題のつながりを表現するsegment-topic-paths（1節で定義）により文脈を規定する．segment-topic-pathsを得るために，ここでは，スクリプトコードなどの不要要素の削除，HTMLタグ間の階層構造などの解析を通して，セグメント間の関係を木構造で表現するGeneric Nested-Structure Patterns（GNSPs）（Figure 1）をWebページから抽出する．木構造の根節点はWebページ全体の文書に，葉節点は最小のセグメントに対応し，各節点は階層的にラベル付けされる．例えば，Webページ全体を表す根節点は{1}，そのページが三つの下位セグメントをもつ場合，それらは{1.1}，{1.2}，{1.3}，さらにそれらのうち2番目のセグメントが二つの下位セグメントをもつなら，それぞれ{1.2.1}，{1.2.2}となる．このとき，セグメント{1.2.2}と{1.2}はsegment-topic-pathにより階層的につながっており，{1.2.2}の内容はsegment-topic-pathsにより階層的につながっていない{1.1}の内容よりも{1.2}の内容と関連していると解釈する．

(2) 索引語選択

ここでは，文脈をより特徴づける索引語を選択するために，Term Context Contribution based on Segments-Topic-Paths（TCC-STP）と呼ぶ索引語の重み付けを提案している．TCC-STPは，Webページ$w_i$と$w_j$は特徴的な単語が共通し，かつそれらが共通の文脈で用いられているほど類似度が高いという考えに基づき，そのような類似度における索引語$t$の貢献度$TCC\text{-}STP(t)$を次式により算出するものである．

$$\sum_{i,j \cap i \neq j} f(t,w_i) \times f(t,w_j) \times f(CT(i,j), STP(t,w_i)) \times f(CT(i,j), STP(t,w_j))$$

ここで，$f(t,w)$は索引語$t$のWebページ$w$におけるTF-IDF値，$STP(t,w)$は$t$が出現するWebページ$w$中のsegment-topic-paths，$CT(i,j)$は$STP(t,w_i)$と$STP(t,w_j)$に共通して現れる索引語を表す．また，$f(CT(i,j), STP(t,w_i))$は$CT(i,j)$の$STP(t,w_i)$における出現頻度を表す．ただし，$|CT(i,j)|=0$の場合には，$f(CT(i,j), STP(t,w_i))$と$f(CT(i,j), STP(t,w_j))$はともに1とする．直感的には，文脈に依存する単語を共有

する索引語ほど高い重みが与えられる．この定義の利点は，ある文脈を特徴づけるためには重要であるが少頻度のためにTF-IDF値が低い索引語に対しても，同一の文脈で多用される索引語の頻度に基づき高い重みを与え，結果的にその文脈をより特徴づける索引語集合を抽出できる点であると考えられる．

(3) Webページのクラスタリング

抽出した索引語を用いたWebページのクラスタリング手法として，3.3節で$k$-means法を改良したApplied $K$-Means Clustering（AKC）を提案している．一般的な$k$-means法との違いは，初期重心をWebページ間の距離に基づいて定める点，および重心へのほかのWebページの割当てにおいて独自の類似度を用いている点である．初期重心は，以下のように定義される総重心間距離$D$を最大にする$k$個のWebページとしている．

$$D=\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\overline{w_{c_i}w_{c_j}}$$

ここで，ページ間の距離$\overline{w_{c_i}w_{c_j}}$は，両者に含まれる索引語の頻度に基づき計算される．具体的には，$T$を両者に現れるすべての索引語の集合とした場合，すべての$t \in T$に対して，$t$が二つのWebページに共有される場合はその出現頻度の平均に応じて距離を減算し，逆に$t$がいずれか一方にしか現れない場合は，その頻度に応じて距離を加算する．

一方，重心への各Webページの割当てについては，以下のように定義されるクラスタ内類似度の総和$S$が最大になるように割り当てるものとしている．

$$S=\sum_{c_j=1}^{k}\sum_{w_i\in W_{c_j}}\{S'(w_i,c_j)-D'(w_i,c_j)\}$$

$c_j$は重心となるWebページ，$W_{c_j}$は重心$c_j$に割り当てられたWebページの集合，$S'(w_i,c_j)$と$D'(w_i,c_j)$はそれぞれ$c_j$と$w_i \in W_{c_j}$間の類似度，非類似度を表す．類似度$S'(w_i,c_j)$は$w_i$と$c_j$に共通に現れる索引語の両ページにおける出現頻度の和に基づき計算され，非類似度$D'(w_i,c_j)$は$w_i$に現れ，かつ$c_j$には現れない索引語のうち少なくとも一つの$c_j$以外の重心に現れる索引語の$w_i$における頻度に基づき計算される．$c_j$には現れずほかの重心に現れる索引語は，$w_i$とその重心の類似度に寄与するため，$w_i$と$c_j$の非類似度の計算に加味されるのは妥当といえよう．

ここで，$w_i$には現れるが$c_j$にもほかの重心にも現れない索引語や$c_j$のみに現れる索引語は，頻度が高くとも類似度$S$の計算では全く考慮されない点に注意されたい．他方，索引語の頻度に基づいたベクトルで表現したWebページ間のユークリッド距離を考えた場合，$w_i$と$c_j$の一方にしか現れない索引語は，その頻度に応じて両者の距離を増大させる．そのため，$w_i$が$c_j$と少しだけの索引語を共有し，かつ$c_j$に現れない索引語を多数もつ場合，$c_j$と索引語を共有しないが$c_j$に現れない索引語も非常に少ない別のWebページ$w$と$c_j$の距離よりも$w_i$と$c_j$の距離のほうが索引語を共有しているにもかかわらず大きくなる場合が生じ得る．本論文がクラスタ内のユークリッド距離の総和を最小にする一般的な方法を用いず，独自の類似度を用いるのは，このような問題を避けるためと論じられている．

また，TCC-STPでは少頻度の索引語でも文脈に特徴的であれば重みが高くなることを思い出してもらいたい．そのような索引語は索引語選択で選択される可能性が高く，クラスタ形成に寄与することも期待される．しかし，ユークリッド距離を用いる場合，重心と共有されない索引語が多数存在する場合やその頻度が高い場合，文脈に特徴的である索引語を選択できたとしても，その頻度が低ければクラスタ形成に反映できない可能性がある．ゆえに，重心と共有されない索引語の影響を極力排除し，少頻度索引語もクラスタ形成に寄与できるこのような類似度の導入は，TCC-STPを用いる場合には必然的なものと紹介者は考える．ただし，同様の効果はベクトル間の余弦[徳永 99]を用いても得られるはずなので，単一の重心だけを考慮する余弦と，ほかの重心における索引語の存在も考慮する本論文における類似度のどちらがより適切なクラスタを形成できるかは今後ぜひ比較してもらいたい点である．

## 3. 評価実験

評価実験では，Yahoo! Directoryから収集したWebページを対象に，カイ二乗値により索引語を重み付けするCHI[Yang 97]，およびWebページ間の類似度への貢献度により索引語を重み付けするTC[Liu 03]（ともに3.2節で紹介）を一般的な$k$-means法と組み合わせた手法CHI-KC，TC-KC，さらにそれらに本論文で提案したTCC-STPの重み付けを加味し，かつクラスタリング手法をAKCとしたTCC-STP-CHI-AKC，TCC-STP-TC-AKCの四つの手法を適用し，得られたクラスタの精度を比較することによりTCC-STPとAKCの有用性を示している．テストデータとしては，上位下位関係にない三つのカテゴリーに属する平均50個のWebページからなるデータセットを10個含んだT1，および特定の分野において互いに上位下位関係にある3～7個のカテゴリーに属する84個のWebページにより構成されるデータセットを10個含むT2を用い，各カテゴリーを正解のクラスタとみなしている．

評価基準は得られたクラスタの一定性を表すエントロピーと本来のクラスタに対する適合率と再現率を評価する$F$値（ともに4.2節で定義）を用いており，索引語の抽出割合を100～2%まで離散的に変化させた場合の各評価値の変化をグラフ化している（Figure 2）．この結果においては，TCC-STPとAKCを用いることでT1，T2のいずれに対しても2%だけの索引語を抽出した場合でも各評価値が向上することが示されている．また，索引語の抽出割合を100%とした場合でも，AKCによ

る評価値が通常の$k$-means法の評価値と同等かそれ以上となっており，AKCが単独でも有効に機能していることがわかる．ここで興味深いのは，T2に対してTCC-STPとAKCを用いた場合，索引語の抽出割合が100%より2%のほうが$F$値，エントロピーともに大幅に改善されている点である．これは，共通する索引語が多くなりがちな上位下位関係にあるカテゴリーが存在する場合でも，TCC-STPによる重み付けにより各カテゴリーを特徴づけるより適切な索引語が選択されていることを意味するものと考えられる．

## 4. 関 連 研 究

ここでは，2節と3.2節で述べられている内容を中心に，本論文における提案手法と従来手法との関連について，紹介者の補足・考えを交えつつ述べる．

Webページの分割においては，広告などの不要要素を排除することを主目的としたもの[Gupta 03]と，それに加えて複数のトピックを考慮した検索拡張を目指したもの[Yu 03]の二つの方向性が考えられる．紹介した手法は後者に属するものであるが，[Yu 03]とではその分割方法も分割後のセグメントの扱いも異なる．分割方法に関しては，目的にかかわらず基本的にHTMLタグの階層構造を利用することが多いようであるが，[Yu 03]ではさらに視覚的情報を用いてWebページの領域分割を試みている．これに対して紹介論文では，視覚情報に基づくアプローチは領域の大きさなどのしきい値を事前に決めなければならない，トピックが異なっていてもレイアウトが類似しているセグメントは区別できない，などの問題点があるため，視覚情報を用いない提案手法のほうが優位であると主張している．この点に関しては，すべてのWebページをいずれかの手法だけで適切に分割することは困難と思われるが，実際に利用する立場で考えると，検索結果に影響を及ぼすパラメータの設定が少ない本手法のほうが受け入れられやすいのではないかと思える．

分割後のセグメントに関しては論文中では触れられていないが，[Yu 03]では直接順位付けし適合性フィードバックに用いているのに対し，紹介した手法は文脈を考慮した索引語の重み付けに用いるのみでクラスタリングはWebページ単位である．複数のトピックが混在するWebページを前提とするなら，クラスタリングもセグメント単位とするほうが自然であり，この点は提案手法において今後の検討課題となるといえよう．

一方，索引語の選択に関しては，3.2節で紹介されたCHI[Yang 97]やTC[Liu 03]など従来から文書分類における属性選択としていくつかの手法が提案されている．なお，本論文で提案しているTCC-STPは，TCの直接的な拡張となっている．しかし，TCでは複数の文脈を区別して評価できないため，TCC-STPのほうがより実用的であるといえる．また，クラスタリングと適合性フィードバックを組み合わせた手法も数多く提案されているが，それらの多くは検索結果のWebページ自体を一般的なクラスタリング手法を用いて分類するものであり，複数のトピックが混在するWebページを適切に扱うことを考慮していない．

このように，個々のアプローチやその特定の組合せがこれまで論じられてきた中で，本論文は実世界における複数トピックの混在するWebページの存在を強く意識したうえで，それらを拡張しつつ融合する意欲的，かつ今後さらなる発展が期待できる興味深い枠組みを示すものである．

## ◇ 参 考 文 献 ◇

[Gupta 03] Gupta, S., Kaiser, G., Neistadt, D., and Grimm, P.: DOM-based content extraction of HTML document, *Proc. 12th Int. World Wide Web Conference, WWW2003*, pp. 207-214 (2003)

[Liu 03] Liu,T., Liu,S., Chen,Z. and Ma,W.Y.: An evaluation on feature selection for text clustering, *Proc. 21th Int. Conf. on Machine Learning, ICML-2003*, pp. 488-495 (2003)

[徳永 99] 徳永健伸：情報検索と言語処理，東京大学出版会 (1999)

[Yang 97] Yang, Y. and Pedersen, J. O.: A comparative study on feature selection in Text Categorization, *Proc. of the 14th Int. Conf. on Machine Learning, ICML-1997*, pp. 412-420 (1997)

[Yoo 06] Yoo, S. Y. and Hoffman, A.: Clustering-based relevance feedback for web pages, *Proc. 9th Pacific Rim Int. Conf. on Artificial Intelligence, PRICAI 2006*, pp. 464-473 (2006)

[Yu 03] Yu, S., Cai, D., Wen, J. R., and Ma, W. Y.: Improving pseudo-relevance feedback in web infromation retrieval using web page segmentation, *Proc. 12th Int. World Wide Web Conference, WWW2003*, pp. 11-18 (2003)

2007年2月21日 受理

## 著 者 紹 介

大原 剛三 (正会員)

1995年大阪大学大学院基礎工学研究科前期課程修了．1997年大阪大学産業科学研究所助手，現在に至る．博士（工学）．データマイニング，機械学習に関する研究に従事．1996年日本学術振興会特別研究員．IEEE，AAAI，電子情報通信学会，情報処理学会各会員．