3K2-NFC-3-2

テキストマイニングにおける語句計量化指標群の利用に関する一考察

Implementing Term Evaluation Indices in the Total Environment for Text Data Mining

阿部 秀尚[*1]
Hidenao ABE

[*1]文教大学情報学部情報システム学科
Faculty of Information and Communications, Bunkyo University

In this paper, I describe indices for evaluating words and phrases for an integrated text data mining environment called TETDM. The indices calculate the values of each term on a given document based on their usages. The values indicate particular aspects of the terms such as frequency, importance, meaningfulness, and so forth. Then, the values are used for evaluating their usefulness in the term selection task by human experts. For implementing the term evaluation indices on TETDM, I also discuss several visualization methods for convenient evaluation support of the users, in addition to the definitions of the representative indices.

1. はじめに

テキストマイニングにおいて，分析者が実行しようとするタスクには，文書毎の類似性や分類を分析するほかに，文書群中の重要語や特徴語を抽出するタスクがある．このとき，抽出される語や語句は，分析者の目的によって重要度や特徴的である基準を基に選定される．重要度や特徴的である基準は，語句の用いられ方などに依存して決定されるが，どのような数値をもって決定されるかは自明では無い．しかし，重要な語句の選定は，文書を特徴付け，文書毎の分析タスクの実行結果に大きな影響を及ぼす．このため，重要語・特徴語の選定は，テキストマイニングにおいて，分析対象領域の専門家の専門知識と評価に時間のかかるタスクとなっている．

本稿では，テキストマイニングにおける重要語・特徴語の選定タスクを支援するため，語句の用いられ方を数値として計量化する指標を統合テキストデータマイニング環境（TETDM: Total Environment for Text Data Mining）[TET] 上に実装するために必要な事項について考察を行う．語句の計量化指標は，語句を分割せずに出現頻度を用いる指標と語句が１つ以上の形態素から成ることを想定した指標である．これらの指標は，これまで，それぞれの分野や適用対象で独自に提案され，比較可能な形態で利用者に提示されることは少なかった．これに対し，本研究では，TETDM を利用して，１つの語句に対する複数の計量化指標の値の算出とより適切な可視化手法による提示を組み合わせ，語句の評価を行うタスクの支援ツールの構築をめざす．

以降，2. 章では，テキストマイニングにおいて，本研究で支援するタスクについて述べる．次に，語句の評価タスクで利用可能な代表的な計量化指標を 3. 章に述べる．4. 章では，これらの計量化指標を実装したマイニングモジュールと組み合わせ，語句の評価作業を行うための可視化モジュールについて考察する．

連絡先: 阿部秀尚，文教大学情報学部情報システム学科，〒253-8550 神奈川県茅ヶ崎市行谷 1100，電話番号 (0467)53-2111，hidenao@shonan.bunkyo.ac.jp

2. 想定するテキストマイニングタスク

テキストマイニングは,「テキスト中の統計情報を用いて，有用な情報を取り出すこと」と大まかに定義されるが，その対象は大きく分けて，文書と文書を構成する語句である．テキストマイニングのタスクは，文書群中の語句の出現頻度に基づく統計情報を基に文書間の比較を行うアプローチと語句同士の比較を行うアプローチに大別される．さらに，文書群に内在する潜在的な構造であるトピックを抽出するタスクは，２つの方向性を同時に併せ持った性質のタスクと考えられる．これらのタスクを概観すると図 1 のようになる．

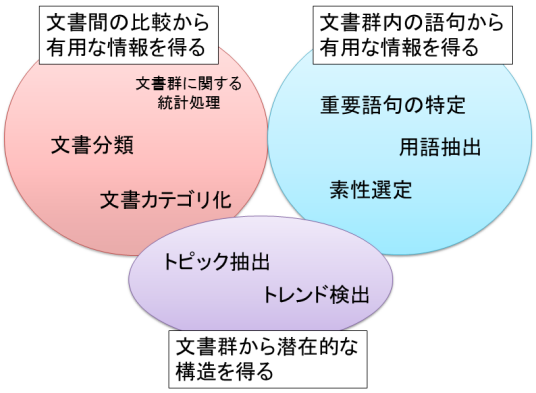

図 1: テキストマイニングでの代表的タスクとアプローチの概観.

以上のタスクは，それぞれが独立して完結するものではなく，文書毎の比較を行う文書分類やカテゴリ化では，語句を比較して特徴語を取り出す必要がある．特徴語の選定によって，文書毎の比較結果が変化するため，語句の比較による評価が重要なタスクとなる．語句の比較による特徴的な語句の選定は，素性マイニング [工藤 03] と呼ばれるが，語句の価値を評価する計量化指標は出現頻度に基づく特定の指標の利用に限られていた．

そこで，本研究では，文書群中での語句の出現頻度から算出可能な複数の計量化指標を用い，利用者にとっての重要語や特徴語を選定するタスクを支援する環境の構築を行う．これにより，語句の重要度や特徴的な度合いを評価し，重要語や特徴語を選定するだけでなく，選定した語句を他のマイニングタスクに適用するによる結果の改善が可能であると考えられる．

3. 語句の計量化指標

語句の計量化指標とは，文書群中でそれぞれ語句が使用されたことによって数え上げられる値やいくつかの値を演算して得られる値を算出する方法を指す．これらの計量化指標の値は，語句の使われ方によって変化し，分析の目的によって意味づけされ，比較の対象となる．最もよく用いられ，基本的な語句の計量化指標に，各語句 $term_i$ の文書群 D での出現頻度 $tf(term_i, D)$ がある．語句の使用頻度を数え上げる tf は，文書群 D 中に語句 $term_i$ が出現するすべての回数を数え上げた値である．一方，文書群中の各文 d で語句 $term_i$ を一回でも含む文 $d_{\in term_i}$ の数を数え上げた値が $df(term_i, D)$ である．これらの出現回数を基本として，これまで重要度指標あるいは評価指標と呼ばれる計量化指標が提案されてきた．以下では，このうちの代表的な指標について算出方法を述べる．

3.1 単語を前提とした計量化指標

文書群 D 中での語句 $term_i$ の出現頻度 $tf(term_i, D)$ や語句 $term_i$ を含む文書の頻度 $df(term_i, D)$ は，語句の構成要素を考慮せず，語句 $term_i$ を見ている．このような計量化指標を本研究では，単語を前提とした計量化指標と呼ぶ．

語句の出現頻度 tf や語句を含む文書数 df は，語句 $term_i$ がそれ自体で意味の無い付属語であっても頻度が多ければ，大きな値となる．また，意味のある自立語であっても，文書中で全般に用いられている語句も一般的であると考えられる．そこで，文書全体での出現頻度を考慮して提案された指標が以下に示す tf-idf である．

$$TFIDF\quad(\quad term_i, D) = tf\quad(\quad term_i, D) \times log_e \frac{|D|}{df(term_i, D)}$$

このほかに，tf と df の双方について，文書群中での出現しにくさに注目したオッズを計算可能である．語句の出現頻度 $tf(term_i, D)$ について，オッズを算出する計算式は以下の通りである．

$$Odds = \frac{tf(term_i, D)}{1 - tf(term_i, D)}$$

3.2 複数の語を許容する計量化指標

語句を１つ以上の要素に分け，語句の出現頻度や特徴的な度合い，有用性を評価する指標は，情報検索の分野でも提案されてきた．中でも，Jaccard 係数は，情報検索の要求フレーズの評価に用いられる指標であり，複数の単語 $w_l(1 \leq l)$ から成るフレーズ $term_i$ について，以下のような定義式で値を算出する．このとき，$df(w_1 \cap \cdots \cap w_l, D)$ は $df(term_i, D)$ に等しい．

$$Jaccard(term_i, D) = \frac{df(w_1 \cap \cdots \cap w_l, D)}{df(w_1 \cup \cdots \cup w_l, D)}$$

Jaccard 係数と同様に，構成要素の順序関係を考慮せず，語句を単語の集合と見なせば，頻出アイテム集合に関する評価指標 [Tan 02] のうち，相関ルールの条件部か結論部のどちらか一方に着目した指標が算出可能である．

また，語句に含まれる構成要素を考慮して，自然言語処理で数値によって語句の用語性を評価し，自動抽出する手法が提案されている [Nakagawa 00]．単語数 L の語句 $term_i =< w_1, \cdots, w_L >$ についての用語性指標は，以下のように定義される．

$$FLR(term_i, D) = tf(term_i, D) \times \left(\prod_{j=1}^{L} (FL(w_j) + 1)(FR(w_j) + 1) \right)^{\frac{1}{2L}}$$

このとき，$FL(w_j)$ は w_j に対する bi-gram で左側に異なる単語があるときの頻度（左異なり数）を表す．同様に，$FR(w_j)$ は右異なり数を表す．FLR のような用語性判定指標は，形態素解析による名詞の同定と組み合わせて用いることで用語を自動抽出する手法として提案されている．

さらに，テキスト中の語句を順序関係のあるアイテムの部分列である系列パターンと見なすと興味度や時間変化に対応できる重要度として，語句の様々な側面を計量化できることが明らかとなっている [櫻井 12, 岩沼 12]．これら系列パターンに関する計量化指標についても，利用者の目的に合致する評価結果を提示するため，語句の構成要素をアイテムとして定義式を利用し，実装を行っていく．

4. 語句評価のための可視化インタフェース

TETDM では，テキストマイニングの各タスクを行うにあたって，必要となる処理を実装した「マイニングモジュール」と入力テキストやマイニング結果を可視化するための「可視化モジュール」を組み合わせることが前提となっている [砂山 11]．そのため，本研究で扱う語句の評価タスクにおいても，計量化指標による評価値の算出に加え，語句の評価結果を利用者に視覚的に提示する可視化モジュールを利用する必要がある．本章では，実装した複数の語句計量化指標の値を効果的に比較して，評価タスクを実行するための可視化インタフェースについて考察する．

4.1 既存の可視化モジュール

現在，公開されている TETDM において，3. 章に述べた計量化指標を利用者に提示できる可視化モジュールは，表 1 に示す各モジュールである．

表 1: 語句の計量化指標を用いた評価における公開済の可視化モジュール

名称	可視化内容
テキスト	テキストを表示
Html テキスト	html 形式でテキストを表示
タグクラウド	タグクラウド状に語句を強調表示
表	表形式でテキストの統計情報を表示

表は，TETDM のテキストファイル読み込み時に自動的に計測される用語の使用頻度について，基本マイニングモジュールなどを通して表リスト形式で可視化する．利用者は，指標による並べ替えをマイニングモジュール側で行い，表示された並び替え結果から目的に沿った語句の選定を行う．

また，タグクラウド形式の表示では，計量化指標によって強調される語句が変化するため，指標による順位付けをより多面

的に把握できる．HTML を付すことができるテキスト形式の表示では，文字の強調を大きさだけではなく，色情報を加えることができるため，複数の計量化指標による語句の比較評価が可能になる．

4.2 複数文章での比較のための可視化モジュール

4.1 節であげた既存の可視化モジュールでは，入力された文章に対する語句の出現度合いをそれぞれの視覚化手法によって表示することが可能であった．しかし，語句の重要度や特徴的な度合いを評価する際には，語句の時間経過に伴う変化や文書群間での比較が大きな役割を果たす [Kontostathis 03]．例えば，語句の時間経過に伴う計量化指標を可視化する手法としては，折れ線グラフが実用的である [阿部 10, Goo]．このため，利用者が時間経過や文書群間での語句の計量化結果を把握できるグラフ表示による可視化モジュールが必要であると考えられる．

5. おわりに

本稿では，テキストマイニングにおける語句の評価作業に注目し，計量化指標による語句の比較を提示する TETDM の実装について考察した．3. 章で述べた語句の出現頻度に基づく計量化指標と語句を 4. 章で述べた可視化インタフェースを通して提示することにより，利用者がより効果的に目的に即した語句を選定今後は，本稿で述べた計量化指標群を TETDM のマイニングモジュールとして実装し，既存の可視化モジュールに加えて，複数の文章での語句の評価値について比較が可能な新たな可視化モジュールの実装を行っていく．さらに，これらのモジュールの組み合わせによる語句の選定作業について，専門家の支援を行い，作業時間の短縮や選定後の文書分類結果の改善を評価していく予定である．

参考文献

[Goo] Google Insights for Search, `http://www.google.com/insights/search/#`

[Kontostathis 03] Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S., and Phelps, D. J.: A Survey of Emerging Trend Detection in Textual Data Mining, *A Comprehensive Survey of Text Mining* (2003)

[Nakagawa 00] Nakagawa, H.: "Automatic Term Recognition based on Statistics of Compound Nouns", *Terminology*, Vol. 6, No. 2, pp. 195–210 (2000)

[Tan 02] Tan, P. N., Kumar, V., and Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns, in *Proceedings of International Conference on Knowledge Discovery and Data Mining KDD-2002*, pp. 32–41 (2002)

[TET] TETDM, `http://tetdm.jp`

[阿部 10] 阿部 秀尚, 津本 周作：専門用語の用法に関する計量化指標の時系列パタン分析, 2010 年度人工知能学会全国大会 (第 24 回) (2010)

[岩沼 12] 岩沼 宏治：テキスト系列マイニングにおける有用性尺度について, 人工知能学会誌, Vol. 27, No. 2, pp. 136–145 (2012)

[工藤 03] 工藤 拓, 松本 裕治：系列パターンマイニングを用いた有効な素性の組み合わせの発見, *IPSJ SIG Notes*, Vol. 2003, No. 4, pp. 147–154 (2003)

[砂山 11] 砂山 渡, 高間 康史, Danushka, B., 西原 陽子, 徳永 秀和, 串間 宗夫, 松下 光範：Total Environment for Text Data Mining, 人工知能学会誌, Vol. 26, No. 4, pp. 483–493 (2011)

[櫻井 12] 櫻井 茂明：多様なデータに対する系列パターンマイニングの適用, 人工知能学会誌, Vol. 27, No. 2, pp. 128–135 (2012)