E-018

Latent Sematic Indexing を利用した日本語文書間類似度による文書群の話題構造抽出

Subject Structure Extraction of Document Group using Latent Semantic Indexing

井上浩* 荒井秀一* 宮内 新*
Hiroshi INOUE Shuichi ARAI Arata MIYAUCHI

1 はじめに

近年, 多くの文書検索手法が提案され, 自然文入力による検索や類似文書検索といった柔軟な検索手法も提案されている [1][2][3].

こういった検索手法では同一の話題やテーマの文書を抽出することができている. しかしながら, 新聞記事やディベートといった文書群の場合, 同一の話題であっても時間と共に内容が詳細になったり話題が広がりをみせたりする. そこで, 本稿では同一の話題に含まれる微細な類似関係を元に, 文書全体の話題の推移や分岐点になるような重要な文書の抽出を行うことで, 文書群の話題構造の表現を試みる.

2 Latent Semantic Indexing

文書検索手法での一般的な手法であるベクトル空間モデルにおいて, 文書-索引語空間を低次元の潜在的な意味空間に射影する Latent Semantic Indexing[1] は, 従来の手法よりも精度の向上が報告されている.

Latent Semantic Indexing では, SVD(Singlar Value Decomposition) を用いて高次元ベクトル空間を低次元の空間に射影することで, 共起しやすい複数の単語を圧縮することができる. 数学的に基づいた手法により次元の削減を行うことができ, 次元を削減することで類似度の計算量の低減が期待できる. また, 射影した低次元空間では潜在的な意味を考慮することができ, 類義語や多義語に対しても効果があると言われている.

一方, SVD による処理の重さやメモリ・記憶容量の増大, 文書の追加・削減毎に SVD を必要としメンテナンスコストが高い, などといった問題を含んでいる.

2.1 semidiscrete matrix decomposition

少ない記憶容量で次元を削減する方法として, semi-discrete matrix decomposition(SDD)[2][3] が提案されている. SDD では文書-索引語行列 A_k の特異値行列 D_k を求めることで, 式 (1) のように分解する.

$$A_k = X_k D_k Y_k^T \tag{1}$$

ここで X_k と Y_k^T の要素を ± 1,0 の値として近似することで記憶容量を抑えている.

検索精度を低下させることなく, 記憶容量を削減することができるため, 大量文書への適用には SDD が有効であると考えられる.

3 本手法の流れ

日本語文書を対象とした, 本手法の検索の流れを本節で説明する.

1. 単語の区切りを形態素として抽出する.
2. 索引語を選定し, 文書-索引語行列の増大を抑える.
3. SDD を行い行列を分解する.
4. 検索要求-文書間の類似度を計算する.

3.1 形態素解析による単語抽出

文書ベクトルと検索要求ベクトルを作成するために, 索引語の候補となる単語を抽出する必要がある. 本研究では日本語文書を対象とするため, 対象文書と検索要求に対して, 形態素解析を行い形態素を単語とする.

3.2 索引語の選定

SDD による行列の分解の処理量を軽減するために, 索引語の選定を行う.

- 単独で意味を持たない非自立語は削除する.
- 活用する語は表記を用いず, 終止形を用いる.
- 検索精度に影響が少ないとされる単一文書のみに出現する語は削除する.
- 共起関係の強い 2 単語をあらかじめ合成する.

*武蔵工業大学,Musashi Institute of Technology

3.3 SDD と低次元空間への射影

文書-索引語行列を式 (1) の SDD を用いて分解する. 文書ベクトル列の低次元空間への射影は式 (2) となる.

$$\tilde{a}_k = D_k Y_k^T \tag{2}$$

また, 検索要求ベクトルの低次元空間への射影は式 (3) となる.

$$\tilde{q}_k = X_k^T q_k \tag{3}$$

3.4 類似度計算

検索要求-文書間の類似度には余弦を用いる. 検索要求ベクトル$\tilde{q_k}$と文書ベクトル列$\tilde{a_k}$から式 (4) で算出する.

$$sim(a_k, q_k) = \frac{\tilde{q}_k^T \cdot \tilde{a}_k}{\sqrt{|\tilde{q}_k||\tilde{a}_k|}} \tag{4}$$

4 実験

本手法により微細な類似関係を元に文書群の話題構造を表現できるかを確認するために実験を行う. 実験には, 毎日新聞 ('95) の同一ニュースを伝える記事 45 文書を用いる. 検索要求にはその中の任意の記事を用いることとする. ここで記事の内容は全日空機ハイジャック事件に関する記事で, 時間の経過に伴って記事番号は大きくなっている.

4.1 実験方法

新聞記事の話題の時間的推移を探るために最も類似している文書を抽出していく. 古い記事が新しい記事の内容を受けることがないため, 各記事の検索対象文書は検索要求である記事自身よりも過去の記事とする. 全ての記事において最も類似している文書を抽出し, その文書間をリンクで繋ぎ図示する.

4.2 実験結果

全ての記事に対して最も類似している記事をリンクで繋いだ結果を図 1に示す. 図 1の番号は記事を表し, リンクは各文書の最も類似している文書に繋がっている.

4.3 考察

図 1中の太線で表している, 最も多く文書の繋がっている経路が記事の時間経過に伴う話題の主な記事といえる. 実際に文書の内容を見てみると, 経路の序盤では事件の経過を伝える記事や警察・政府の対応を伝えている. 記事 11 では時間の流れと事件の対応関係を示しており, 記事 11 に繋がる多くの記事は事件を分析する内容が多く含まれている. 記事 16 ではここまでの事件経過をまとめており, 記事 20 では事件解決を伝える内容でどちらも

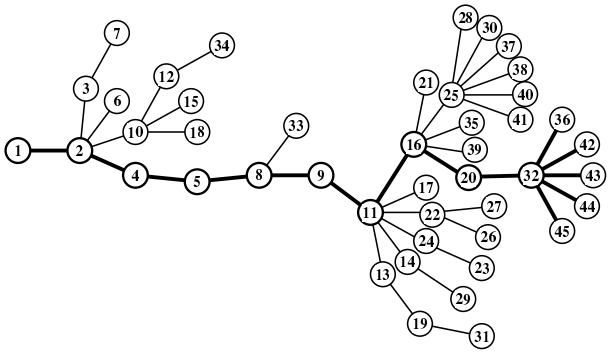

図 1: 話題の構造抽出結果

事件内容を知るためには重要な記事であると考えられる. 記事 32 では犯人の供述を伝えるもので, その葉となっている記事は犯人像を探る内容となっている. また, 記事 25 と記事 25 の葉となっている記事は乗客や乗務員の話伝える内容となっている.

しかしながら, 警察の対応や事件の経過を伝えている内容である記事 10 や, 事件の総括的内容の記事 22 が主経路に含まれていないといったこともみられた.

全体として, 最も長い経路が事件の経過を伝える記事を多く含み, 話題の時間的推移を表すことのできる可能性があると考えられる. また, 話題の分岐点や話題の広がりをも表現できる可能性がある.

5 まとめ

本稿では, 微細な文書間の類似性を元に文書群の話題構造を抽出した.

同一ニュースの新聞記事群を図示した結果, 事件の経過を伝える主な記事を示すことができ, また話題の分岐点や広がりを表現できる可能性を示した.

本稿では微細な類似度を用いて話題構造の抽出の可能性を示したものの, 定量的な評価を行う必要がある.

参考文献

[1] S.Deerwester,S.T.Dumais,G.W.Furnas, T.K.Landauer,R.A.Harshman, "Indexing by Latent Semantic Analysis",Journal of the Society for Information Science, 1990

[2] T.G.Kolda, D.P.O'Leary, "A semidiscrete matrix decomposition for latent semantic indexing information retrieval", ACM Transactions on Information Systems, 1998

[3] Jason Dowling, "Information Retrieval using Latent Semantic Indexing and a Semi-Discrete Matrix Decomposition", 2002