1P2-10in

発想支援のためのテキストマイニング

Text Mining for Inspiration

イ スンジュ[*1] Soonju Lee　堀浩一[*1] Koichi Hori　赤石美奈[*1] Mina Akaishi

[*1] 東京大学大学院工学系研究科航空宇宙工学専攻
Department of Aeronautics and Astronautics, University of Tokyo

Most of text mining tools lack accuracy compared to traditional data mining tools that deal with structured data. Hence, it has been sought to find a way to apply data mining techniques to these text data. In this paper, we propose a text mining system(application) which can help revising human's thought by clustering and classification. Classification will be improved for using bridging-text techniques.

1. はじめに

創造活動を行うにあたり、人間は精神的に大きい労働をしている。新しい発想がなければ、従来の物より良い物を生み出すことはできないため、人間は精神的労働を行い、創造活動をしてきたのである。創造活動をする際、人間は言葉で表現することが多く、その言葉を用い、自分の発想を整理し、修正を行う。実際、認知心理学では、人間にとって思考と言語は深い関連があると言われている。その理由で、人工知能が発想支援を行うためには、人間の言葉を知る必要がある。しかし、自然言語を機械に理解させるのは非常に困難であり、機械独特の自然言語の分析方法が用いられている(テキストマイニング、情報検索など)。

創造活動の中で、個人のアイディアは欠かせない部分である。閃いたアイディアから、創造活動が始まると言っても過言ではない。しかし、そのアイディアは簡単に生まれる訳ではなく、長い悩みの末、何らかのきっかけや事件によって、人間の脳から生まれると考えられる。

発想を支援することは、人間からアイディアが出やすくすることと、そのアイディアの整理や修正を手伝うことだと考えられる。そのため、機械が発想支援を行うには、人間が考えるために使う手段である、テキストの情報(自然言語)の分析方法を利用する必要がある。

ここでは、発想支援のため使われるテキストマイニングシステムを提案する。

2. 関連研究

人々の創造的な活動のプロセスを変化させ、今までより創造的な活動ができるよう支援する「創造活動支援」(Creativity Support)が最初に挙げられる。

普通、創造活動支援には人間の身体、状況依存性、環境、機械学習などを合わせて考えているが、今回の研究では機械学習に集中して考える。

「機械学習」のなかで、明示されておらず今まで知られていなかったが、役立つ可能性があり、かつ、自明でない情報をデータから抽出することを意味する「データマイニング」という分野がある。その中で、テキストを重心に扱うのが「テキストマイニング」である。

ブリッジテキストを探すことは、テキストマイニングの手法の1つで、ソース、ターゲット、ブリッジの3つ文書群を用意し、加重値(tf-idf、吸引力など)を計算することで、ソースとターゲットをつなぐ文書をブリッジの文書群から探すことである。また、類似単語や文書を探すため、文書分類技術(サポートベクトルマシン(SVM)、意思決定ツリー)を利用する。

最後に、人間がこのシステムを使うことにあたり、より簡単で読みやすくするよう、ヒューマンインターフェースの研究も関連するのであろう。

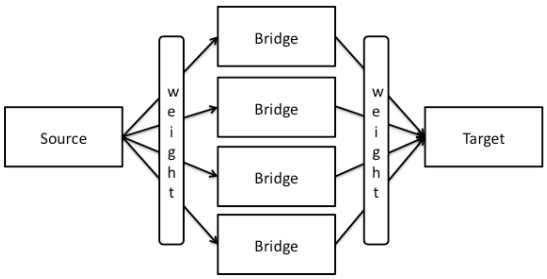

図1　ブリッジテキスト

3. システム構成

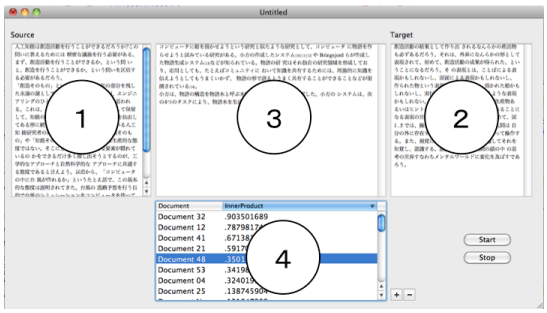

図2　アプリケーション画面

3.1 ブリッジテキストアプリケーション

すべてのデータはテキストのみである。

連絡先:東京大学大学院工学系研究科航空宇宙工学専攻,
住所:東京都文京区本郷7-3-1　工学部7号館420号室
Tel:03-5841-6637
E-mail:lee@ailab.t.u-tokyo.ac.jp

開発環境:Mac OS X Cocoa Application
使用言語:C, Objective-C
日本語形態素解析:Mecab 0.82
英語ステミング:Porter Stemmer
加重値:tf-idf、吸引力

上の図は目的ｉのためのアプリケーションである。①がソース、②がターゲット、③がブリッジテキスト、④が加重値の表である。

①、②から入力されたテキストデータから用意されているブリッジデータとの加重値計算を通し、④から加重値の表が出力され、③からその内容が見られるような構造になっている。

(1) 入力してもらうテキスト

基本的には、自分の考えたことをソースに書いてもらう。ターゲットにはアイディアの目的や予想される結果を書く。短すぎると、解析のためのデータ料が少ないため、5行くらいのデータを要求する。

(2) 用意されるデータセット

計算によりclusteringとclassificationされるデータ集である。今回は、広い分野のことをカバーできるよう、新聞記事と論文のアブストラクトを約1千件の集合を作成した。

データの準備

1。収集	新聞記事、論文（略）
2。前処理	不必要な情報処理
3。文章集合生成	データセットになる

システムの流れ

1。Clustering	文章集合の概要を獲得
2。特性抽出	目的にあったパラメタ設定
3。学習	項目選定また学習
4。分類分析	信頼度計算

図3 システム概要

3.2 システムの流れ

(1) データの準備

3.1に書いてあるようなデータが入ると、前処理として形態素分析を行う。その結果を利用し、不必要な情報（ソースコードや助詞など）を除去する。

(2) Clustering

作業の効率化を計るため、データ集のクラスタリングを行う。その結果をみて、各クラスター間の類似度をパラメタ的に調節できるようにし、できるだけテーマ別に分けられるようにしている。

ソースとターゲットと同じクラスターに分類されたものから分析は始まり、加重値も高めに設定される。

(3) Classification

学習が要るとことであって、最初に用意されたデータセットに新しいアイディアや新聞記事、論文のアブストラクトを入れる度、どのカテゴリーに入るかを決めてもらう。その学習でclassifierの特性ベクトルを作る。しかし、これには限界があって、ベクトルの信頼度のみで判断されるため、正確度が悪くなる。

(4) ブリッジテキスト

ここで、人の考えを利用する。考えの初期段階では、曖昧な表現が多い。しかし、その目的を書くことになると具体的な言葉が登場し、内容もより詳細になる。その点に着目し、信頼度の高いカテゴリーの中から、ターゲットのテキストとの類似度で最終的に融点順位を決める。

4. 今後の課題

ソースとターゲットから計算されたデータセットの各文書のランクに対するフィードバックが行われていない。そのため、Classificationとブリッジテキストの正確度の悪さを補う作業はないこととなる。

そのため、カテゴリーを候補、非候補に分け、非候補郡に対してフィードバックを行う。フィードバックは、項目のランクと信頼度の違いを利用したいと思う。

$$Dist(x,y) = RD(x,y) \times weight \quad (1)$$

$$weight = \log(\sqrt{rank + 0.1}) \quad (2)$$

Distは距離、RDはx,yの順位の差を意味する。0.1は計算の欠陥を補うためのもの。[5]

非候補は入力データとの関連する根拠が少ないため、順位の高かったものも重心として、距離を計算する。加重値は順位の差を広げるためのものである。

これを利用し、非候補の上位と、候補の最上位との距離を計算し直すことで、非候補のフィードバックができることが予測される。

参考文献

[1]「物語生成のためのトピックブリッジング手法の提案」 佐藤真 赤石美奈 堀浩一 2010年
[2]「創造活動支援の理論と応用」 堀浩一 2007年
[3]「言語処理のための機械学習入門」 奥村学 高村大也 コロナ社
[4]津田裕一, 八木秀樹, 平澤茂一, ”単語の共起を考慮に入れたナイーブベイズモデルによる文書分類”, 第 29 回情報理論とその応用シンポジウム予稿集, pp.613-616,2006
[5]Interplay of Text mining and Data mining for Classifying Web Contents, YJ Choi and S.S Park, KAIST, 2001