人工知能学会第２種研究会資料
SIG-KST-2010-02-04 (2010-12-01)

読み手の感じ方を反映させた文章可視化手法
A text visualization method that reflects viewpoints of the reader

郷原浩之　大澤幸生　西原陽子
東京大学大学院工学系研究科システム創成学専攻

要約

知識が文章化されていたとしても,単語や文脈の理解の仕方の違いによって読み手の理解が異なり,知識を継承・共有することが困難な事例がある.

本研究では,知識が記述された文章を読んだ時に読み手が感じる知識間の類似性を可視化することで,読み手に自身の文章に感じ方を認識させ,それらの知識に対する新しい見方と他人の見方を獲得することを支援することを目的としている.

そのために本研究では文章化された知識を読み手が分類し,その分類基準を教師データに用いた学習によって,文章化された知識間の類似度を計算することで,当人の分類基準を可視化するモデルを開発した.

1. 背景・目的

組織を持続的に運営していくためには,知識が継承されていくことが必須である.そのために業務の運営に必要な知識はマニュアルやレポートという形で,文章化されて保存されている.この結果生まれる業務手順や技術の勘所を網羅したマニュアルは膨大な量であるために,全てを正しく理解することは時間的資源が制約されている状況下では現実的ではなく,実際には各々が業務を通じて得た経験をベースに,マニュアルを補完的に参照する.

マニュアルを各々の経験が反映して理解しているために,組織の構成員の間で必ずしもマニュアルに対して共通の理解がなされていないケースが見受けられる.

文章化された知識に対する共通理解の欠如は組織全体の効率的かつ正確な知識継承にために克服するべき課題である.知識に対して共通の理解が無い状況では,最も適切な理解をしている者に他者からの質問が集中する.それではコストの点において On the job training と変わらず,マニュアルを整備したコストが無駄になってしまう.

特許やWebサイトなどのように膨大なテキスト情報を効率的に理解し検索する目的で,文章の類似度を計算し可視化する方法は多くの方法が提案されている[1][2].本手法では実装したシステムの適用対象を組織内に存在する技術文書としているため,特許や Web サイトに比べてテキスト量は極端に小さい.またテキスト量が小さいが故に,システムのユーザは解析対象のテキストにある程度目を通すことができる.そのため,豊富な事前情報をシステムに与えることができる.既存の可視化手法では,事前情報を与えることを想定していないために,本システムが持つ豊富な事前情報を活用できない.

嗜好や観点の相違を抽出する方法[3][4]が提案されているが,これらの既存手法では単語レベルでの嗜好や観点の相違を議論している.本研究は文章に対する感じ方を抽出したい点で既存手法を用いることができない.

本文では 2 節にて提案手法の概要を述べる.また 3 節にて本手法で使用する文章類似度計算手法について紹介する.4 章にて本提案手法を実装したシステムについて説明する.5 節にて本手法の評価実験について述べる.6 節にて考察と結論を述べる.

2. 概要

本手法では文章群をその類似度によって 2 次元平面に配置する.類似度が大きい2つの文書の 2 次元平面上での距離を小さくする.この可視化結果を得るための過程は「読み手の感じ方を反映させた文章間の類似度を計算すること」と「文章間の類似度を基に文章を 2 次元平面に配置すること」の2つに分かれる.

読み手の感じ方を反映させた文章間の類似度を計算する方法として Polynomial Semantic Indexing (PSI)[5]を用いた.PSI は文章と文章,文章と単語の類似度を計算する方法の一つで,単語間の類似度が教師データを用いた学習によって決定される点に特徴がある.単語は一つの単語からなる文章と見なすことで,単語は文章と同様に計算できる.ユーザがシステムに対して与えることができる事前情報を PSI が要求する教師データとして活用する.

文章間の類似度を基に文章を 2 次元平面に配置することを実現するために,バネモデルを用いて無向グラフの 2 次元平面上の座標配置を決定する方法の一つである Kamada-Kawai 法[6]を用いた.

3. Polynomial Semantic Indexing

全ての文章の中のユニークな単語数を N とし,各単語には一意の番号を与える.全文章の集合

を D とし,一つの文章を d で表す.

PSI では文章を以下のベクトル d で表現する.

$$\{d_t\}_{t=1}^{N} \in R^N \quad (1)$$

ここでd_jは番号 j の単語の文章 d 内での頻度を表す.ただし,d の各要素は tf-idf 値に再計算し,その後 d が単位ベクトルになるよう正規化する.

PSI では文章間の類似度を以下の式 f を用いて計算する.文章d_iと文章d_jの類似度は(2)で定義される.

$$f(d_i, d_j) = d_i^T W d_j \quad (2)$$

ここで$W \in R^{N\times N}$であり,$W_{i,j}$は番号 i の単語と番号 j の単語の類似度を表す.W は教師データからの学習によって決定される.教師データを用いた学習の更新式は(5)(6)で表される.PSI では 2 単語間の類似度のみならず,複数単語間の類似度を計算することも可能であるが,実行時間の観点から本システムでは 2 単語の類似度を計算する.

PSI の教師データは(q, d^+, d^-)を一組とし,このタプルを複数用意する.ここで$q, d^+, d^- \in D$であり,ある文章 q について,d^+は q と関係がある文章,d^-は q と関係がない文章を表す.

W の空間計算量は$O(N^2)$であるために,N が大きくなったときにメモリの確保が難しくなる.例として,N=10000 のときには W の確保に 400MB のメモリ領域が必要になる.そのために PSI では W を低ランクの行列 2 つを用いて近似する.

$$\bar{W} = U^T V + I \quad (3)$$

I は単位行列,U,V は M×N の行列である.また M≪N とし,本システムでは M=200 に設定している.

教師データによる学習によって W を作るために,まず U,V を初期化する.U,V の全ての要素を平均 0,分散 1 の正規分布に従う乱数で埋める.その後,全ての教師データに対して

$$f(q, d^+) - f(q, d^-) < 1 \quad (4)$$

(4)を満たすならば,U と V を次のように更新する.

$$U \leftarrow U + \lambda V(d^+ - d^-)q^T \quad (5)$$
$$V \leftarrow V + \lambda U q(d^+ - d^-)^T \quad (6)$$

λ は収束までの学習率を定義する定数で,本システムでは 0.01 に設定している.これを教師データの中で(4)を満たすものが一つも無くなるまで行う.(5)(6)で得られた U,V をそれぞれ(3)に代入することで,W が得られる.

4. システムの処理

本システムは「対象テキストの読み込み」,「PSI のための教師データの作成」,「PSI による文章と単語間の類似度計算」,「計算結果の可視化」の 4 段階からなる.以下では順に説明する.システムは C#で実装していて,.Net 上で動作する.無向グラフの 2 次元平面上での座標決定に使用する Kamada-Kawai 法の計算はグラフ描画ソフトである Graphviz[7]を呼び出すことで実現している.

4.1 テキストの読み込み

現在のシステムは英語と日本語の文章を解析対象としている.英語は空白によって単語を分割し,すべてを小文字にして解析している.日本語は形態素解析によって単語を分割している.形態素解析の実装にはライブラリとして SlothLib[8]を用いている.SlothLib は内部で茶筌を用いて形態素解析を行っている.

4.2 教師データの作成

PSI ではある文章に対して,その文章と関連がある文章と関連がない文章を一つの組として,その組を複数用意したものを教師データとして要求する.そのために,用意した文章集合を複数のグループに分類をする.この時の分類は各文章の意味を理解したうえで,ユーザが行う場合と,文章のメタデータを利用して機械的に分類する方法の 2 種類がある.そして,できた分類から実現可能な全ての組み合わせを教師データとする.ここで,同じ分類に含まれている文章はその文章と関連がある文章とし,他の分類に含まれている文章はその文章とは関連がない文章とする.この点について,本システムでは文章集合は固定であり,増えることはないので過学習になることを心配する必要はない.

4.3 類似度の計算

PSI で計算される文章と単語の類似度は同一のクエリに対する相対値として定義される.そのために,文章d_iとd_jについて,

$$f(d_i, d_j) \neq f(d_j, d_i) \quad (7)$$

である点に注意したい.本システムでは単語を固定して,それに対するその他全ての文章の関連度を計算し,上位 2 つを選ぶので,全て与えられた単語に対する文章の類似性の相対値のみで処理できるので,(7)について全く問題がない.

4.4 平面上への配置

同じ分類に含まれる文章は互いに関連しているとする.また先の計算によって選ばれた単語と,その単語と最も大きな類似度を示した 2 つの文章も互いに関連しているとする.このとき,単

語と文章をノードとし,関連をエッジとみなすことで,無向グラフができる.この無向グラフにKamada-Kawai 法を適用することで,この無向グラフを 2 次元平面に描画することができる.Kamada-Kawai 法でのノード間の結びつきの強さを決めるパラメタであるバネの自然長,バネ係数は全てエッジで同じとする.類似度をバネの自然長,バネ係数に反映させることで,平面上での 2 ノード間の距離を 2 ノードの類似度として表現できるが,目的は語の抽出にあるので,類似度情報は無視した.

5. 評価実験

本手法による可視化結果が読み手の感じ方を反映できていることを検証するために,個々人別の可視化結果を参考にしながら組み合わせ発想を行わせる実験をした.

組み合わせ発想は 2 つの物を組み合わせることで新しい価値を持った物を創造することであり,組み合わせ発想を行う場としてイノベーションゲーム[9][10][11] (イノベーションゲームは大澤幸生の登録商標である。)が提案されている.本実験もルールを一部変更したイノベーションゲーム上で行った.

5.1 実験環境

被験者 13 名を 2 つ（7 人,6 人）のグループに分けた.各グループ全員が組み合わせ発想を行った.組み合わせる対象は彼らが受けてきた大学の講義であり,既存の講義を組み合わせて,新しい価値を持った講義を提案することを課した.組み合わせる対象の講義は全部で 22 個あり,実験開始前に各講義の内容について説明を行う機会を設けて,その内容を周知させた.

ここで各グループでは自由にコミュニケーションを取ることを許した.また全員が組み合わせ発想を行うものの,半数はアイディアを提案する度にグループの全員の注意を向けさせプレゼンを行った（※企業家プレイヤ）.残りの半数は,自身も組み合わせ発想を行う傍ら,そのプレゼンに対してコメントを与えた（※消費者プレイヤ）.

※企業家プレイヤ,消費者プレイヤというのはイノベーションゲームでの用語である.通常の場合,企業家プレイヤはアイディア創出に専念し,消費者プレイヤはアイディア創出を行わず,企業家プレイヤによって提案されたアイディアにコメントしたり,自身が抱える問題をプレゼンしたりする.

なお,実験の前に組み合わせる対象の全ての講義について,その講義が他の講義と組み合わせやすいか,組み合わせにくいかという内容のアンケートを取った.このアンケートの結果を教師データとして用いて図 1 のような個人別の可視化図を用意した.各講義は被験者にその内容を記載していただき,その内容を解析対象の文章とした.図 1 中のノードは各文章,ここでは講義名に対応する.

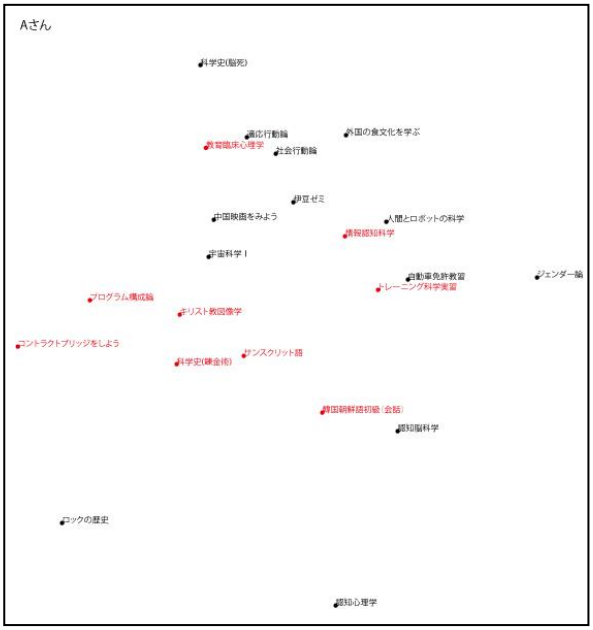

図 1：実験に用いた個人別の文章群可視化図の例

個々人に配布した可視化図に自由に線などを記述させ,組み合わせ発想の支援に使用していただいた.

ゲーム終了後に企業家プレイヤによって提案されたアイディアを各グループ内の消費者プレイヤによって評価していただいた.評価基準は実現性（現時点で技術的に実現可能かどうか）と新規性（新しいコンセプトを有しているかどうか）と実用性（恩恵を受ける人々がいそうかどうか）の 3 点を 4 段階評価によって行った.

5.2 実験結果

企業家プレイヤが可視化図に記載した線に着目した.6 人の企業家プレイヤ（うち一人は無記入）から 42 本の線が得られた.ほぼ全ての線は 2 つの組み合わせ対象を表すノードを結んでいる.例外的に他の線の中間とノードを結んでいる線とノードを囲むような線が確認できる.これらは 3 つ以上のノードを組み合わせたということを表す.線で結ばれているノードの距離は 7mm から 173mm である.なお可視化図は A4 用紙の 200mm×200mm の領域に印刷されている.

組み合わせ発想で組み合わされたアイテムの可視化図上での距離と提案されたアイディアの質の関係を表したのが表 1 である.得られた線のうち最長の線の長さを基準に区間を 4 分割した.

そして各分割された区間に属する線によって作られたアイディアとその質についての関係を表現した.質のいいアイディアとは,実現性,新規性,実用性の 3 つの項目の中で 2 つ以上の項目で平均を超えたアイディアと定義した.

距離	2 項目で平均を超えたアイディア数	提案されたアイディア数
0 ～ 43.25	6	8
43.25 ～ 86.5	4	8
86.5 ～ 129.75	2	8
129.75 ～ 173	0	1

表 1：組み合わせ発想で組み合わされたアイテムの可視化図上での距離と提案されたアイディアの質の関係を表した

表 1 から可視化図上で近い位置にあるアイテムを組み合わせることで質の高いアイディアが生み出されたことがわかる.

6 考察と結論

質が高いアイディアを創出するためには,アイテムの特徴や構成要素を適切に理解している必要がある.またプレゼンを行って,皆を納得させる必要があるために,読み手によってアイテムは十分に理解されている必要がある.この点が組み合わせ発想を用いて本提案による可視化図を評価している理由である.

本提案による可視化図上で近い距離に位置するアイテムを組み合わせて質の高いアイディアを創出できたということは,ユーザが各アイテムを理解する上でその支援をできているということに当たる.その点で本提案による可視化図の有用性が主張できる.

参考文献

[1] Fortuna, B., Grobelnik, M. and Mladenic, D.: Visualization of text document corpus, Informatica, Vol.29, No.4, pp.497--502 (2005)

[2] 岩田具治,山田武士,上田修功,トピックモデルに基づく文書群の可視化,情報処理学会論文誌, Vol.50, No.6, pp.1649--1659, (2009)

[3] 西原陽子, 辻由紀子, 田中大智, 砂山渡, 嗜好の違いの解釈を支援するアニメーションインタフェース, 知能と情報, Vol.19, No.1, pp.3—12, (2007.1)

[4] 西原陽子, 田中大智, 砂山渡, 観点の違いによるキーワード間の関係の変化を捉えるための可視化手法, 可視化情報学会論文集, Vol.29, No.6, pp.9--16, (2009.6)

[5] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Corinna Cortes, and Mehryar Mohri. Polynomial Semantic Indexing. In Advances in Neural Information Processing Systems 22, pp. 64--72 (2009)

[6] T. Kamada and S. Kawai, An Algorithm for Drawing General Undirected Graphs, Information Processing Letters 31, pp. 7--15 (1989)

[7]Graphviz www.graphviz.org/

[8] 大島裕明, 中村聡史, 田中克己：“SlothLib: Web サーチ研究のためのプログラミングライブラリ”, 日本データベース学会 Letters, 6, 1, pp. 113--116 (2007)

[9] 大澤幸生,中村潤,高市暁広,古田一雄,青山和浩,定木淳,組み合わせ発想を刺激するイノベーションゲーム,第 4 回知識・技術・技能の伝承研究会 (2007)

[10]Yukio OHSAWA, Kensuke OKAMOTO, Yuji TAKAHASHI, and Yoko NISHIHARA, Innovators Marketplace as Table Game versus as Web Agora, In Proc. IEEE International Conference on Data Mining, Workshop on Chance Discovery (2010)

[11]Yukio Ohsawa, and Yoko Nishihara, Innovators Marketplace: Process of Games as a Service System of, by, and for Innovators, In JSAI Proc. International Workshop on Innovating Service Systems, pp.115--124 (2010)