日本ソフトウェア科学会第 27 回大会 (2010 年度) 講演論文集

Java3D を用いた企業クラスタリング

山田 裕文　大和田 勇人

この論文では, 企業クラスタリングの新しい手法を提案する. 企業クラスタリングとはウェブ上のテキストデータを用いて企業を特定の関係ごとにクラスタリングする手法である. 従来の研究では, 膨大な実験結果のデータにより実験結果が見づらくユーザーの負担が大きいものであった. そこで, 私たちは人間の認知メカニズムに基づいた Java3D を用いる視覚化を提案する. 提案手法は関係文書をウェブ上から取得した後, 形態素解析器によって名詞だけを取り出し統計解析ソフトの R によってクラスタリングを行う. クラスタリングされた後, クラスターの特徴を表すキーワードは半球状の 3 次元空間上に表示される. ここでは, テキスト情報による評価値 *TFIDF* によって高さを, 検索エンジンによる評価値 *HIT* によって大きさを決定することにより重要なキーワードはより見やすい位置に配置される.

1　はじめに

ある分野の企業を知るためには就職四季報などに存在するカテゴリサーチを用いて行う方法と，キーワードによって検索する方法がある．前者の場合，一つの企業について一つのカテゴリのみしか確認できないとう問題がある．近年，複数の事業を持つ企業も少なくないため，このような検索方法では難しい．後者の場合も多くは主要な事業のものしか記述されておらず，もし仮にすべての事業内容を記述すると膨大な量になるため現実的には不可能である．また，両者は共にある程度の予備知識を必要とすることから，これらの方法は現実的ではない．過去の研究で，テキスト情報を利用し，一つの企業を入力することで，その企業に関係のある企業のリストを入手し，リスト内の企業を入力企業の関係ごとにクラスタリングするという手法を提案した [1]．これらは，各クラスタは入力企業との関係を示したタグが付与されていて，出力結果は視覚的に見やすくなっている．しかし，データマイニングアルゴリズムよって生み出された多量のタグのために, 有益なタグを発見するのは困難である．意思決定に関連のある知識を見つけるために, ユーザーはタグをくまなく探す必要がある．そこで本研究では，Java3D を用いて企業クラスタリング結果を視覚化することを提案する．この提案手法によって ユーザーにとって実験結果のタグの比較が容易になる. さらに, ユーザーは自由にクラスターの中を探索することが出来る.

2　ユーザーの認知の限界

実験結果を見るとき, ユーザーは評価値の付いた長いタグのリストを見ることになる．この場合ユーザーは, 意思決定のための興味深いタグを見つけるために, ルールをくまなく探さなければならない. 人間の限定合理性のために, 決定プロセスは優先構造をとると考えられる．モントゴメリによると, 意思決定者は制限された部分集合を潜在的に有用な二者択一に分離 する [2]．そして, 比較をし, この意思決定プロセスは繰り返される．さらに,「属性集中」と呼ばれる KDD 方法論によると, 人間は自動的に少ない数の興味深い属性に注目することが実験データから発見された．その

Company Clustering using Java3D.

Hirofumi Yamada, 東京理科大学理工学研究科経営工学専攻, Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science.

Hayato Ohwada, 東京理科大学理工学部経営工学科, Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science.

人の限られた認識力のために, 意思決定者は, 各瞬間に少量の情報だけを処理する. これらの情報処理のタスクから, 本提案手法では, 私たちは 2 つのルールを作った.

1. ユーザーが少ない量の属性でタグの部分集合に焦点を当てることができるようにする.
2. クラスター内のタグ間で比較を行うことができるようにする.

3 Java3D

Java3D とは Java 向けの三次元グラフィックス (3D) の拡張 API である. サン・マイクロシステムズからパッケージとして提供されている. 実際の描画は OpenGL や DirectX などの 3D グラフィックス用 API を呼び出す事によって行っている. Java3D の設計思想は VRML に大きく影響を受けている. 本研究では,Java3D によって 3D 空間に半球を作成し, その半球状に評価値に応じてタグを配置していく. Java3D は VRML や OpenGL とは異なり, Java 言語の利点を継承しており, WWW 上でインタラクティブなシステムを構築するのに適している. Java3D では表示の対象となる情景を「シーン・グラフ (Scene Graph)」と呼ばれるツリー構造によって記述する. シーン・グラフの構成要素となるのは, その空間内に存在する物体の他, 光源, 音源, 運動やイベント処理を行うオブジェクト, それらをグループ化するためのオブジェクトなどが含まれる.

4 提案手法

本研究の提案手法の全体の流れを図 1 に示す.

4.1 企業リストの生成

まず，企業リストの獲得について説明する．企業リストとは，入力された企業に対して同業他社やグループ会社等なんらかの関係があると考えられる企業を集めたリストである．その取得方法は，まず初めにユーザが検索したい企業の URL である seedURL を入力する．そして，その seedURL にリンクを張っている Web ページをサーチエンジンの持つ機能である backlink 検索により獲得する．つまり，サーチエンジン (今回は AltaVista を使用) を用いて backlink 検索を行い，その検索の検結果から上位 N 件 (本実験では N = 50) を取得する．獲得した URL を fans とする．その fans の URL に順次アクセスをし HTML ファイルを取得し，各々のファイルに含まれているハイパーリンクの URL を全て抽出する．その中で出現回数の多い順にソートし，上位 M 件目 (本実験では M = 20) と同等以上の回数の URL を企業リストとする．つまり，企業リスト内の企業数は入力する企業により異なり，最低で M 社含まれているリストが作成される．

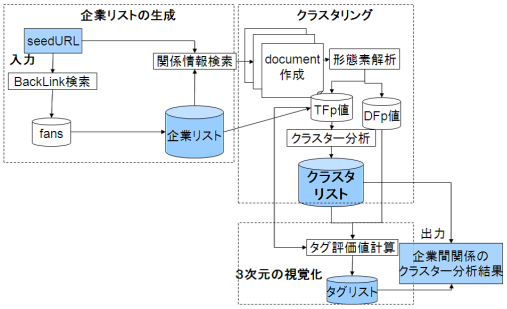

図 1 システム全体の流れ

4.2 検索エンジンからの関係情報の取得

企業間関係を抽出する方法は，検索エンジンの結果を用いる．企業リスト内のすべての企業名と入力企業を検索エンジンの機能である「AND」で結合し，クエリとして検索エンジンに投げかける.「AND」は論理積であり，リスト内の企業と入力企業の両方を含むページを検索エンジンに要求する．さらに，すべての名称は「""」で囲まれている.「""」で単語を囲むことにより，完全一致で検索することが可能である．得られた検索結果の上位 L 件 (本実験では $L = 100$) のタイトルと概要文から検索クエリとして使用した単語を除いた文書集合を各企業の企業間関係情報 (*document*) として保存する.

4.3 クラスタリング

得られた *document* は，従来手法と同様に形態素解析機 MeCab を用いて形態素解析を行い，*TFp* 値と *DFp* 値をぞれぞれ求める [3] そのうち，タグの評価値

を計算する対象は，クラスタ内の全企業の *document* での $TFp(t) > 0$ であるもの．つまり，クラスタ内の全企業の *document* で出現している名詞をタグ対象とする．

4.4 評価値算出方法

1 つ目は，前述した $TFp(t)$ 値と $DFp(t)$ 値を使用する (1)．関係情報が書かれている *document* 内での出現割合が高く，なおかつ，クラスタに属さない企業の *document* では，あまり多く出現しないという条件の下求める．

$$TFIDF(t \mid C) = \prod_{c \in C} TFp(t) \times \log\left[(1 - DFp(t)) + 1\right] \tag{1}$$

- $TFIDF(t \mid C)$: 単語 t のクラスタ C に対する TFIDF 評価値

2 つ目は，クラスタ内の企業が出現する Web ページ上に，より多く共起している単語の方が重要であると考えられるため，検索エンジン (実験では google を使用) のヒット数を用いて求める (2)．*document* に出現する単語全てを検索対象としてしまうと，膨大な時間がかかってしまうため，検索対象とする単語は，前述の TFIDF 評価値の結果上位 K 件 (本実験では 50 件) とする．

$$HIT(t \mid C) = hit\left(\bigwedge_{c \in C} c \wedge t\right) \tag{2}$$

- $HIT(t \mid C)$: 単語 t のクラスタ C に対する検索件数評価値
- $hit(t)$: t を検索クエリとした時の検索エンジンのヒット件数

そして，ウォード法によるクラスター分析を行い，デンドログラムの作成を行う [4].

4.5 視覚化

視覚化には 3 次元表現を用いる.3 次元表示の利点は, 新たに加えられた次元によてタグの特性をより反映できることである．3 次元表示の例として,Julien らは ARVis と呼ばれるアソシエーションルールを 3 次元表現するシステムを開発した [5]．本提案手法でも ARVis と同様に 3 次元の半球上にオブジェクトを配置していく．

先ほど得られたクラスターは 3D スペースの中で視覚化される．それぞれのタグは球とその上の文字列によって構成されている．タグの特徴は, 球の直径, 球の高さである．球の高さは評価値 *TFIDF* を表現する．球の大きさは評価値 *HIT* を表現する．この視覚メタファーはタグ間の比較を容易にし, よりルールを強調する．明確に, 大きく高い球は *TFIDF* と *HIT* の両方が高いので重要なタグであるということができる．一方小さく, 低く位置しているタグは両方の値が低く, 注目すべきでないということができる．

5 実験

本手法の有用性を示すためにシステムを構築し，実験を行った．*seedURL* は従来手法との比較を容易にするために，従来と同じ「日立製作所」の URL を使用した．

図 2 はクラスター分析結果のデンドログラムである． また，視覚化に関して,「日立工機」,「日立ツール」の 2 社の企業からなるクラスター 4 を 3 次元描写した結果を図 3 に示す.

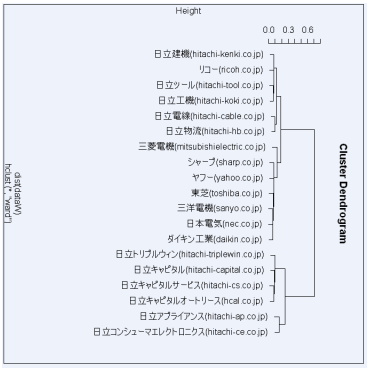

図 2 提案手法：日立製作所のデンドログラム

比較のために「日立工機」,「日立ツール」の 2 社の企業からなるクラスター 4 を従来手法でタグクラウドで表したものを図 4 に示す.

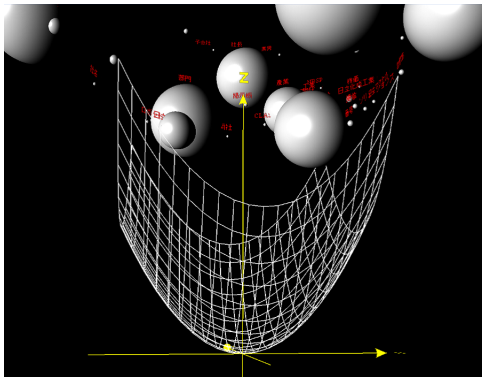

図 3　提案手法：3 次元表現 (クラスター 4)

企業名	ラベル
4	9日立ツール 18日立工機	番号 工具 大手 株価 当社 同社 産業 CLASS 社長 関社 証券 部門 エンジニアリング 日立メディコ ロボット ハイテクノロジーズ キャピタル 社名 国際 製作所 昭和 機械 市場 東証 年収 電動 プラント 平成 HITACHI 器具 ツール ドリル 卓上 日立金属 大証 ページ 工場 自動車 日立マクセル ソリューションズ 電気 順位 日立化成工業 工業 有限 NBSP グループ 業界 掲示板 子会社

図 4　従来手法：タグクラウド (クラスター 4)

6　考察

図 2 を見ると, 日立製作所のデンドログラムは他の電気機器関連の大手 3 社が同じクラスターになり, さらに図の下方で日立グループの企業が同じクラスターになるなどわかりやすい結果になった. 図 3 では, 直交座標上の半球状の空間にタグが配置されそのタグの評価値を表す球がすぐ下に配置されている. 図 4 を見ると, タグクラウドで表現されている文字では大きさや色などの差が大きくない場合どちらが評価値が高いのか判断しづらいが, 3 次元描写では大きさや高さなどが一目で解るようになっている. 大きさが判断しやすくなったのは, 従来手法は文字自体の大きさを変えていたので, 文字の種類や色によって大きさの見え具合に差があったからである. 全体として,y 軸の低い位置にスペースが空いてしまっているので, 評価値を修正して均等に配置されるように改善する必要がある.

7　まとめと今後の展望

本論文では, 調べたい企業名を入力することで, 自動的に入力企業に対する関連企業を抽出し, 3 次元空間に描写する手法を提案した. 具体的には, 調べたい企業の URL を入力し, 検索エンジンの backlink 検索を用いることで企業リストを取得し, その企業リストの企業を入力企業との関係ごとにクラスタリングを行い, 3 次元空間に描写した. 今回 3 次元の描写には,Java 言語の特性を引き継いでおりインタラクティブな言語である Java3D を使用した.3 次元に表現することによって, ユーザーは視覚的にどのタグが重要なのかを理解出来るであろう. 現時点では, 興味深さ指標として *TFIDF* と *HIT* を使用しているが, 高さや大きさだけでなく, オブジェクトの色などの表現も重要度を表すために使用できるので, 新たな指標を加えることで, よりユーザーにとって理解しやすいものにしていく必要がある.

謝辞 本論文を作成するにあたり, 指導教官である大和田勇人先生には終始適切かつ的確な助言, ご指導をいただき感謝の意絶えません. また, ともに研究してきた大和田研究室の皆さん, 母, 父, また多くの方々のご支援, ご協力のものに本論文を書き終えるに至りましたことに感謝の意を持ち, 簡潔ではありますが謝辞とさせて頂きます.

参 考 文 献

[1] 山田裕文, 松井藤五郎, 大和田勇人: “Wikipedia を類義語辞書として用いた企業クラスタリング”, 第 23 回人工知能学会全国大会, 高松,2009.

[2] Montgomery H. (1983) Decision rules and the search for a dominance structure: towards a process model of decision making. In Humphreys P.C., Svenson O., Vari A. (editors.), Analysing and aiding decision processes, Amsterdam:North Holland, pp 343-369

[3] 徳永健伸: “情報検索と言語処理”, 東京大学出版会 (1999).

[4] 永田靖, 棟近雅彦: “多変量解析法入門”, サイエンス社 (2000).

[5] Julien Blanchard, Fabrice Guillet and Henri Briand. (2007) Interactive Visual Exploration of Association Rules with Rule Focusing Methodology. Author manuscript, publshed in ”Kowledge and Informaiton System 13,1(2007)43-75”