# 特徴把握のための医学系研究課題の可視化

渋木 英潔† 木村 泰知‡ 一瀬 信敏* 武田 善行◇

†北海学園大学ハイテク・リサーチ・センター ‡小樽商科大学社会情報学科
*札幌医科大学附属産学・地域連携センター ◇東京大学大学院工学系研究科

## 1 まえがき

近年，大学の社会貢献活動の重要性が増しており，大学の研究を産業界へ積極的に技術移転することが社会から求められている．こうした背景のもと，異分野同士の大学間連携や産業界との産学連携の取り組みが進められているが，大学の研究活動は専門性が高いため，他の分野，一般社会からはその特徴が見えにくいのが現状である．そこで，各大学の研究活動の特徴を専門外の人々にも理解しやすいように提示することは，大学の技術移転活動を促進する基礎材料のひとつになると考えられる．

これまで医学系研究課題に関する情報提示は，案件ごとに各研究を一覧表として提示する形式が一般的であったが，研究間の共通性や差異性を一目で判断するのに適した形式とはいえない．また，課題とする病気の種類や対象となる身体部位など共通性や差異性を判断するための観点はいくつか存在するが，そのような観点ごとに各研究の特徴を把握することは医学関係者以外にとって困難であった．それゆえ，我々は医学系研究課題を，対象となる身体部位，課題とする病気や生命現象，それらの解決・研究手法という３つの観点から自動的に分類し，分類結果をバブルチャートにより表現することで各研究の特徴把握を容易とすることを目的とした研究を行っている [1]．本稿では，SVM[2, 3] を用いた分類結果の可視化を行い，現状の把握と問題点の整理を行う．

**2** 節では，本稿で扱う分類カテゴリーについて説明する．**3** 節と **4** 節では，それぞれ分類処理と可視化処理について述べる．**5** 節はまとめと今後の予定である．

## 2 分類カテゴリー

本研究では，消化器や呼吸器といった「対象」となる部位，感染症や癌・腫瘍といった「課題」とする病気，遺伝子治療や予防医療といった「解決手法」の３通りの観点から分類を行う．各観点における分類カテゴリーを表 1 に示す．

これらのカテゴリーは，札幌医科大学の基礎医学・臨床医学研究を対象としたデータを参考に人手により設定した．また，医学関係者以外であっても直観的な理解を可能とすることを目的としているため,「その他」といった具体的イメージを喚起しないカテゴリーは設定しなかった．表 1 のカテゴリーには将来的に検討する余地があるが，本稿で対象とするデータにおいては重大な問題はなかった．

## 3 分類処理

本稿では，SVM を用いて研究課題の分類を行う．分類対象となる研究課題には，科学研究費補助金データベース [4] から，2005 年度の札幌医科大学，北海道大学，京都府立医科大学のデータを用いた．科学研究費補助金データベースには，研究課題のタイトル，概要，研究分野，キーワードなどの情報が含まれている．SVM の素性空間は，科学研究費補助金データベースに含まれる情報から，茶筅 [5] を用いて抽出した名詞により構成した．

素性としてどの情報を用いるのがよいかを調査するために，タイトルを基本として，概要，研究分野，キーワードを組み合わせた場合を考慮した．具体的には，タイトル，タイトルと概要，タイトルと研究分野，タイトルとキーワード，タイトルと概要と研究分野，タイトルと概要とキーワード，タイトルと研究分野とキーワードの 7 通りである．本稿では，素性名詞の抽出元となる情報の違いは考慮せず，頻度を素性値とした．

**2** 節で述べた観点ごとに one-versus-rest 法による分類を行い，その正解率を調査した．10 分割交差検定による結果を表 2 に示す．また，全体の正解率を表 3 に示す．表の値は，“正解数/評価件数 (正解率)” の形式で表わされており，ボールド体は各観点における最高値を示している．

表 2 と表 3 から，タイトルと研究分野を組み合わせ

表 1: 各観点における分類カテゴリー

| 対象 | 課題 | 解決手法 |
|---|---|---|
| 家族・社会 | 感染症 | 機器診断・開発 |
| 血液・骨髄 | 運動・行動 | 疫学・調査 |
| 生殖 | 疾患・外傷 | 遺伝子医療 |
| 胎児・乳幼児・小児 | 移植 | 機能解析 |
| 免疫系 | 薬剤・放射線障害 | 標的化・免疫医療 |
| 消化器 | 癌・腫瘍 | 再生医療 |
| 呼吸器 | QOL | 機構解析 |
| 目・耳鼻・皮膚 | 認知・運動・行動 | 遺伝子・ゲノム解析 |
| 青少年・成人 | 発生・分化 | 看護・介護・支援 |
| 泌尿器 | 活性・調節 | 薬物・化学医療 |
| 老人 | 生体高分子 | 新規治療・診断法開発 |
| 脳・神経・精神 | ストレス | 予防医療 |
| 細胞・組織 | 炎症・疼痛 | |
| 骨・筋・関節 | | |
| 生体高分子 | | |
| 心臓・血管 | | |

表 2: 観点ごとの正解率

| 素性 | 対象 | 課題 | 解決手段 |
|---|---|---|---|
| タイトル | 110/271(40.6%) | **146/271(53.9%)** | 131/271(48.3%) |
| タイトルと概要 | 99/271(36.5%) | 115/271(42.4%) | 86/271(31.7%) |
| タイトルと研究分野 | **129/271(47.6%)** | 145/271(53.5%) | **149/271(55.0%)** |
| タイトルとキーワード | 120/271(44.3%) | 133/271(49.1%) | 116/271(42.8%) |
| タイトルと概要と研究分野 | 102/271(37.6%) | 117/271(43.2%) | 87/271(32.1%) |
| タイトルと概要とキーワード | 103/271(38.0%) | 115/271(42.4%) | 84/271(31.0%) |
| タイトルと研究分野とキーワード | **129/271(47.6%)** | 136/271(50.2%) | 115/271(42.4%) |

た場合の正解率が「課題」を除いて最も高く,「課題」においても 2 番目の正解率であった．反対に概要を組み合わせた場合の正解率は，タイトルのみを素性とした場合の正解率よりも全ての観点において低く，その有効性を確認できなかった．しかしながら，人手による正解データの作成において，タイトルのみでは判断が難しく概要を参照しなくては分類できないデータも存在していた．それゆえ，概要の情報をタイトルなどの情報と同じレベルで処理したことに問題があったと考えられ，タイトルなどの情報で分類できない場合に概要情報を用いるといった段階的な分類機構が必要になると考えられる．

また，最も正解率が高かったタイトルと研究分野の組み合わせにおいても，その値は全体として 52.0%であり，十分に高い値とはいえない．上で述べたように，人間にとっても分類が困難なデータが存在しており，このようなデータに対しては最終的な判断をユーザに任してしまうという支援ツールとしての用法も考えられ，正解率の向上とともに今後の課題である．

## 4 可視化処理

可視化システムの実装は，以下の三点に留意して行われた．第一に，各大学の研究の特徴が一望でき，重点的に研究されているカテゴリーが明瞭に表示されること，第二に，研究されているカテゴリーの年度ごとの推移が表現され，今後どのように変化していくかの判断材料となること，第三に，大学間の特徴の比較が容易にでき，それぞれの大学で必要とされている知識や技術が何かを直観的に把握できることである．

各大学の特徴を一望するために，図 1 や図 2 に示すようなバブルチャートを用いた．縦軸と横軸は **2** 節で述べた各観点のカテゴリーであり，円の大きさは研究課題の数を表している．年度ごとの推移を概観できるようにアニメーションで時系列の変化を表示し，大学間の比較を可能とするため図 2 のように三大学までの

表 3: 全体の正解率

| 素性 | 全体 |
|---|---|
| タイトル | 387/813(47.6%) |
| タイトルと概要 | 300/813(36.9%) |
| タイトルと研究分野 | **423/813(52.0%)** |
| タイトルとキーワード | 369/813(45.4%) |
| タイトルと概要と研究分野 | 306/813(37.6%) |
| タイトルと概要とキーワード | 302/813(37.2%) |
| タイトルと研究分野とキーワード | 380/813(46.7%) |

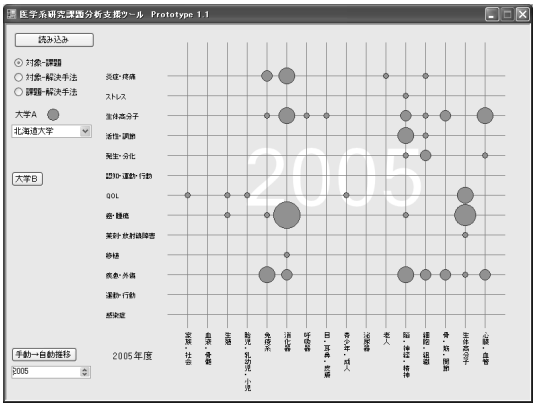


図 1: 出力の例 (a)

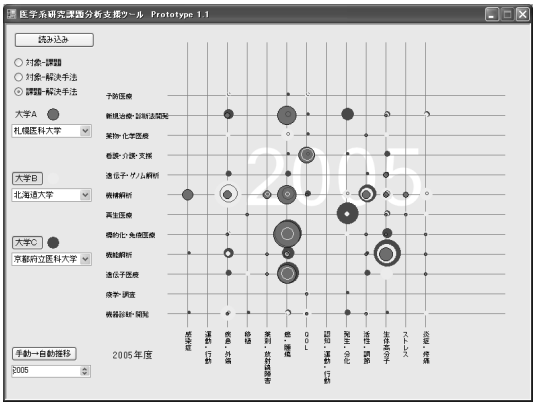


図 2: 出力の例 (b)

バブルチャートを重ねて表示できるようになっている．加えて，研究課題の数が増加傾向にあるカテゴリーの円は徐々に明度を増し，減少傾向の円は暗く変化させることで，各カテゴリーの動向情報を一望できるようにした．また，円をクリックすることでそのカテゴリーの属する研究課題の一覧表が表示される．

これらを実装することにより，例えば，同じ「癌・腫瘍」という課題を研究している 2 大学において，一方は「遺伝子治療」という解決手法を中心に取り組んでいるのに対して，もう一方は「薬物・科学医療」を中心に取り組んでいるといった比較が容易にできるようになる．その結果，それぞれの大学が必要としている知識や技術を補うような連携が可能かといった判断材料として役立つと考えられる．

## 5 まとめ

本稿では，医学系研究課題を，対象となる身体部位，課題とする病気や生命現象，それらの解決・研究手法という 3 つの観点から SVM を用いて自動的に分類し，分類結果をバブルチャートにより表現することを行った．これにより，大学ごとの研究の特徴を一望することができ，観点ごとに共通性や差異性の把握を容易とすることができた．今後は，分類精度の向上を図るとともに，研究課題の特徴把握および比較分析をさらに容易とするよう改良する予定である．

## 参考文献

[1] 木村泰知, 渋木英潔, 一瀬信敏: 医学系研究課題の網羅的分析に向けての分類支援, NLP 若手の会第 2 回シンポジウム (2007).

[2] T. Joachims, Learning to Classify Text Using Support VectorMachines. Dissertation, Kluwer, 2002.

[3] T. Joachims, Optimizing Search Engines UsingClickthrough Data, Proceedings of the ACM Conference on KnowledgeDiscovery and Data Mining (KDD), ACM, 2002.

[4] 科学研究費補助金データベース: http://seika.nii.ac.jp/

[5] ChaSen/茶筅: 奈良先端科学技術大学大松本研究室, http://chasen-legacy.sourceforge.jp/