様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 ２７ 日現在

機関番号：１４３０１
研究種目：挑戦的研究（萌芽）
研究期間：2018～2019
課題番号：１８Ｋ１８９４２

研究課題名（和文）推薦システムによる新規無機化合物予測

研究課題名（英文）Recommender system for discovering new inorganic compounds

研究代表者
世古　敦人（Seko, Atsuto）

京都大学・工学研究科・准教授

研究者番号：１０４５２３１９

交付決定額（研究期間全体）：（直接経費）　　4,200,000円

研究成果の概要（和文）：無機結晶データベースに機械学習の一つである推薦システムを応用することで，合成可能な新規無機化合物を効率的に発見する方法を提案した．特に，既知データが少ない場合においても新規無機化合物を発見することが可能な推薦システムに基づいた方法を構築した．推薦システムを材料科学に応用する研究であり，その結果，100億以上の化学組成の中から，無機化合物が存在する組成を予測することができる．新規無機化合物の発見を大幅に加速させることができる．

研究成果の学術的意義や社会的意義
料科学と機械学習の融合により材料研究の加速を目指す「材料インフォマティクス」が国内外において行われ始めており，申請者はこの分野において，最先端の研究を行っている．本研究は，材料科学に推薦システムのアプローチを応用する試みであり，材料科学と機械学習の融合による挑戦的研究である．本研究は，材料インフォマティクスという新しい学術領域に貢献する．
また，推薦システムによる方法論により，１００億もの候補全組成に対して，合成可能性の情報を定量的に与えることができる．よって，本研究は，合成可能性のある化学組成をリストアップすることができるものであり，これまでの材料探索の体系を変える可能性がある．

研究成果の概要（英文）：In this study, the relevance of chemical compositions where stable crystals can be formed, i.e., chemically relevant compositions (CRCs) was predicted. Herein we adopt recommender system approaches to estimate CRCs. This approach significantly accelerates the discovery of currently unknown CRCs that are not present in the training database.

研究分野：計算材料科学

キーワード：推薦システム　無機化合物　機械学習

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。

様　式　C－１９，F－１９－１，Z－１９（共通）

１．研究開始当初の背景

近年，材料科学と機械学習の融合により材料研究の加速を目指す「材料インフォマティクス」が国内外において行われ始めている．代表者は，無機結晶データベースに機械学習の一つである推薦システム(Amazon.com などにおける膨大な商品の中からおすすめ商品を推薦する枠組みなどの総称)を応用することで，合成可能な新規無機化合物を効率的に発見する方法を提案した．これは，推薦システムを材料科学に応用する世界初の研究である．その結果，100 億以上の化学組成の中から，無機化合物が存在する組成を効率的に予測することができることが分かった．例えば，３元系の推薦された組成の上位 20 件および 100 件中，それぞれ 16 件(80%)，56 件(56%)の組成が無機化合物の存在する組成である（図１）．これは驚異的な数字であり，推薦システムにより新規無機化合物の発見を大幅に加速できることを示している．

２．研究の目的

このような推薦システムによる予測が必要とされるのは，候補組成が膨大かつ既知データが少ない５元系のような場合である．しかし，既知データの少ない５元系では，ランダム選択と比べると発見効率は遥かに高いものの，３元系などと比べると発見効率が格段に落ちる（図１）．単純な推薦システムの方法は，既知データから化学組成間の類似度を抽出することで，無機化合物が存在する化学組成のルールを見つけ出すからであり，データが少ない場合にルールを見つけ出すことは本質的に難しい．よって，本研究は，推薦システムの方法を発展させることで，既知データが少ない場合に新規無機化合物を発見することができる方法論の構築を実施する．

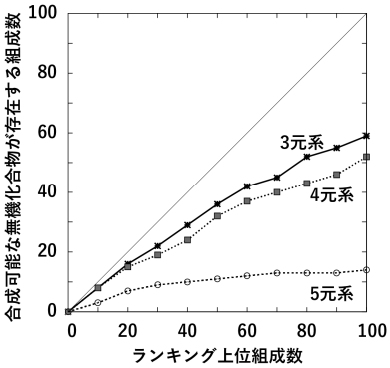

図１　推薦システム（テンソル分解）のランキング上位に含まれる合成可能な無機化合物が存在する化学組成数.

３．研究の方法

既知データが少ない場合に推薦システムの予測性能が悪くなることは，機械学習の分野では常識であり，対象の事前知識を利用することで予測性能が改善する可能性がある．よって，本研究では，化学組成の事前知識を用いる方法を採用する．本研究では，化学組成の事前知識として，化学組成の元素情報から導出される化学組成記述子を導入し，予測性能の向上を目指す．このような化学組成記述子は，申請者の研究により，化合物の凝集エネルギーなどの物性値を高精度に予測できることが分かっており[A. Seko et al., Phys. Rev. B 95, 144110 (2017)]，推薦システムにおいても有効であると期待される．また，上述の単純な推薦システムの方法で使われている化学組成表現も有効であるため，それらと化学組成記述子を組み合わせることができる Factorization

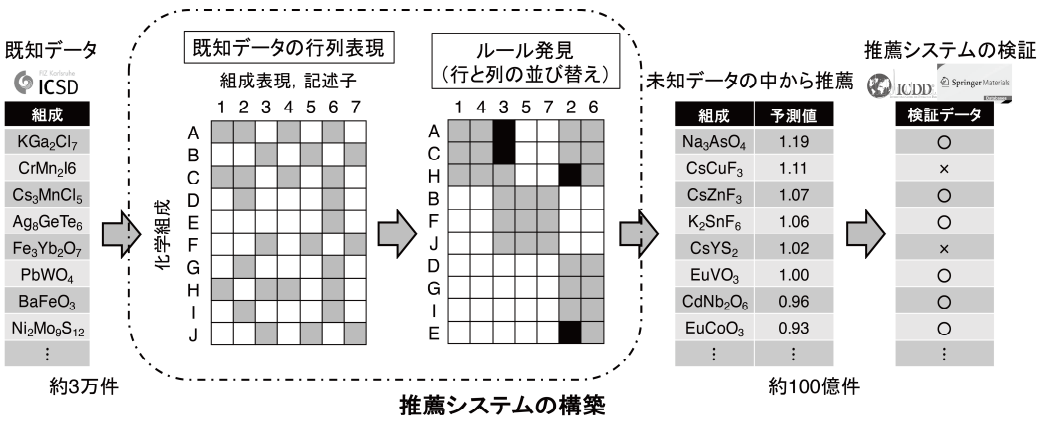

図２　研究方法の概略.

machine[S. Rendle (2012)]などの推薦システムを採用する．

図２に研究方法の概略図を示す．まず無機化合物データベース ICSD の収録組成を既知データとする．次に，組成データを適切に行列化する必要があるため，様々な行列化方法について検討した．行列化された組成データに化学組成記述子を導入し，推薦システムを構築する．得られた推薦システムをもとに約 100 億通りの化学組成に対する予測を行い，新規化合物の合成可能性が高いと予想される化学組成を列挙した．その後，推薦システムにおいて最も重要である未知データに対する予測性能の検証を行った．本研究では，既知データとして用いた ICSD 以外のデータベース（ICDD，SpringerMaterials）を検証用データとして用い，検証用データに含まれる化学組成の発見効率を評価することで，未知データに対する推薦システムの予測性能を評価した．

４．研究成果

本研究では，無機結晶データベースから新規化合物を予測する方法として，元素の情報を事前知識として導入する手法を試みた．学習器としては事前知識を導入することが容易な推薦システムのアルゴリズムである Factorization Machines を用い，構成元素と組成比の情報は Onehot encoding と呼ばれる方法でベクトル化し，事前知識の情報には化合物を構成する元素の物性値の平均・分散・共分散を使った．Factorization Machines により，事前知識を用いることなく一般的なテンソル分解と同等の性能を引き出すには，カチオンとアニオンの組を記述子として与える必要があった．また，3 元系では事前知識を導入することにより，ほぼ性能は変化しなかったが，意図的に学習データ中の既知化合物の数を減らした場合には，事前知識によってデータ不足を補えることを示した．

具体的には，3 元系において，学習データの中の既知化合物を意図的に減らした時の Factorization Machines による予測性能を示す（図３）．縦軸は予測値の大きかった上位 3000 件の中に含まれていた既知化合物の個数を示している．横軸は学習データ中の正例として ICSD に収録されている化合物全体のどれだけを用いたかを示している．右側の場合，ICSD に収録されている 3 元系化合物 (9313 種) の全てを学習データに加えている．中央の場合，ICSD に収録されている 3 元系化合物の 10%(931 種) を学習データに加え，それ以外の組み合わせの化合物は未知とみなしている．左側の場合，ICSD に収録されている 3 元系化合物の 1%(93 種) を学習データに加え，それ以外の組み合わせの化合物は 未知とみなしている．左側と中央の状況は 5 元系のような学習データ中の既知化合物の割合が 3 元系の場合以上に少ない状況を再現している．凡例は事前知識を表す成分をどれだけ記述子に含めたかを表している．青色は事前知識を一切入れていない場合，橙色は元素の情報を 1 成分に圧縮した場合，緑色は元素の情報を 2 成分に圧縮した場合である．

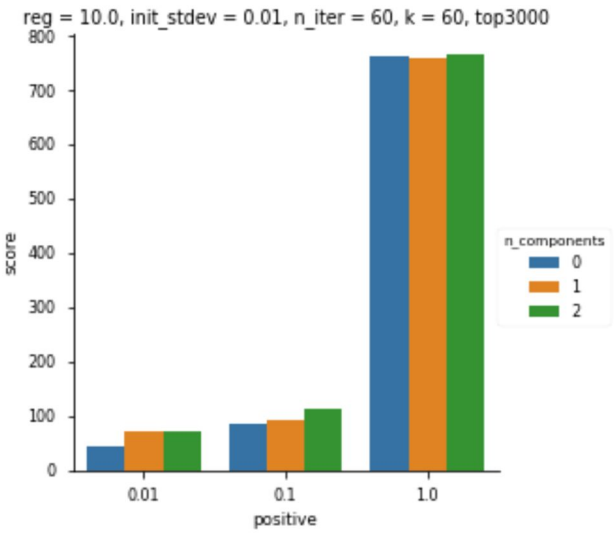

図３　３元系における事前知識を利用した推薦システムによる化合物予測正答数．学習データの中の既知化合物を意図的に減らした場合．

図 3 において右側にいくほど性能がよくなっているが，これは学習データ中の既知化合物の個 数が右側ほど多いことを考えると自然なふるまいである．図左

と図中央に注目すると，事前知識を導入することで性能が向上していることが確認できる．このことから学習データに占める既知化合物の割合が 3 元系の場合よりも更に少ない５元系などの状況では，事前知識を導入することによりデータ不足を補うことが可能である．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 A. Seko, K. Toyoura, S. Muto, T. Mizoguchi and S. Broderick	４．巻 43
２．論文標題 Progress in nanoinformatics and informational materials science	５．発行年 2018年
３．雑誌名 MRS Bulletin	６．最初と最後の頁 690--695
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Hayashi Hiroyuki、Hayashi Katsuyuki、Kouzai Keita、Seko Atsuto、Tanaka Isao	４．巻 31
２．論文標題 Recommender System of Successful Processing Conditions for New Compounds Based on a Parallel Experimental Data Set	５．発行年 2019年
３．雑誌名 Chemistry of Materials	６．最初と最後の頁 9984～9992
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1021/acs.chemmater.9b01799	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考