Vol.2018-IOT-42 No.2
2018/6/28

Sanny: 大規模ECサイトのための精度と速度を両立した分散可能な近似近傍探索エンジン

三宅 悠介[1,a)]　松本 亮介[1,b)]

概要：EC サイトの商品種類増大に伴う情報過多問題を解決するため，利用者の要求を満たす商品を自動的に提案する機能が EC サイトにとっての関心事となる．大規模 EC サイトで商品を提案するために扱う特徴量は大規模かつ高次元ベクトルの集合となるため，類似度の比較は精度と計算量を抑えた近似解を用いなければならない．商品を自動的に提案する機能には、可用性を担保しつつ，提案内容の的確さと充分な応答速度が求められる．本報告では，大規模 EC サイトで商品を提案することを想定して，精度と速度を両立した分散可能な近似近傍探索エンジン Sanny を提案する．Sanny は，検索質問データ (クエリ) に対する高次元ベクトル集合の近傍探索結果の上位集合が，クエリと高次元ベクトル集合を任意の次元数で等分した部分ベクトル単位で近傍探索した結果と類似しやすいことに着目して，提案すべき商品の近傍探索を部分ベクトル単位での探索に分解することで分散処理可能にし，その探索結果の和集合である近傍候補から再度近傍探索を行うことにより，全体として高速に近似近傍探索を行える．実験では，従来の近似近傍探索に対する速度並びに精度面での性能比較について評価を行う．

Sanny: Scalable Approximate Nearest Neighbors Search System Using Partial Nearest Neighbors Sets

Yusuke Miyake[1,a)]　Ryosuke Matsumoto[1,b)]

Abstract: Building a recommendation system has become a big concern to solve the information overload problem of the electronic commerce sites. Implementing a feature selection system based on machine learning requires approximation of the nearest neighbor search for maintaining the availability and speed while allowing to reducing the accuracy and complexity since the feature sets consist of large and high dimensional vectors. In this report, we propose Sanny, a search engine by approximation of nearest neighbors and using the partial neighbor set. Sanny aggregates the neighbors of the low dimensional vectors using the property that the neighbors of high dimensional vectors obtained from the neural network overlap with the neighbors of partial vectors. Sanny searches the candidates generated by the aggregation, which results in reducing the total time of searching. We evaluate performance and accuracy comparison of Sanny with the existing methods.

1. はじめに

EC サイトの商品種類増大に伴う情報過多問題を解決するため，利用者が興味を持つであろう商品を自動的に提案する機能が EC サイトにとっての関心事となる．商品の提案は商品同士の類似性を根拠とすることから，商品を自動的に提案する機能では，商品の特性を機械的に扱えるよう数値化し距離空間における近傍集合を求める．この数値化した商品特性を特徴量と呼び，一般に多次元ベクトルとして表現される．深層ニューラルネットワークの発展によって，これまで適切な特徴量を導くことが難しかった画像やテキストに対しても，人の感性に近い，特性をよく表現する高精度な特徴量を得られるようになったことから [4], [16]，これらの深層ニューラルネットワークから得られる数百か

[1] GMO ペパボ株式会社 ペパボ研究所
Pepabo R&D Institute, GMO Pepabo, Inc., Tenjin, Chuo ku, Fukuoka 810-0001 Japan
[a)] miyakey@pepabo.com
[b)] matsumotory@pepabo.com

ら数千次元の高次元ベクトルである特徴量が EC サイトの商品提案機能に利用され始めている [20], [21]．また，数十万から数百万件の商品を扱うような大規模な EC サイトでは，特徴量の集合の要素数も，扱う商品数に比例して大規模なものとなっている．このような大規模 EC サイトにおける商品を自動的に提案する機能では，大規模かつ高次元ベクトルな集合に対して，提案内容の的確さと充分な応答速度を両立しながら可用性の担保が求められる．

商品提案機能における近傍集合の探索において，精度と速度の両立の観点から，正確ではあるがデータ数と次元数に比例して計算量が増加する線形探索の利用は現実的でない．そこで，精度を犠牲にして近似解を用いることで計算量の増加に対処する近似近傍探索の手法が提案されているが [1], [15], [19]，各手法の想定するデータ数や次元数を超えた大規模かつ高次元ベクトルな集合に対しては，精度と探索速度の両立が困難になる．このようなデータ数と次元数の増加に伴う近似近傍探索の性能劣化に対処するため，探索対象を分割する手法が用いられる．探索対象のベクトル集合を行方向に分割し，各部分ベクトル集合の近傍探索結果を近傍候補とする手法 [13] では，次元数の増加に伴う課題が残る．一方，列方向の分割 [10] では，検索質問データ (クエリ) を同様に分割し各部分ベクトル集合における距離計算の一部を並列して求め，これをまとめる．しかしながら，最終的な距離計算量はデータ数に比例することから，列方向の分割による探索速度の改善は限定的となる．

また，EC サイトにおいて応答速度の低下が販売の機会損失につながるため [6]，商品提案機能には安定した応答速度の確保が求められる．そこで，可用性の観点から同一データを保持する近似近傍探索機能を持つサーバを複数台用意することで負荷分散する手法が用いられるが，処理能力がサーバ台数に依存してしまう．サーバ障害時においても安定した応答速度を確保するためには，影響を局所化し処理能力の低下を最低限に留める必要がある．

そこで，本報告では，大規模 EC サイトで商品を提案することを想定して，精度と速度を両立した分散可能な近似近傍探索エンジン Sanny を提案する．筆者らは，商品特性をよく表現しており高精度に類似度が比較可能な高次元かつ密なベクトルの集合において，クエリに対する高次元ベクトル集合の近傍探索結果の上位集合が，クエリと高次元ベクトル集合を任意の次元数で等分した部分ベクトル単位で近傍探索した結果の上位集合と類似しやすいことに着目した．これにより，提案すべき商品の近傍探索を部分ベクトル単位での探索に分解することで分散処理可能にし，その探索結果の和集合である近傍候補から再度近傍探索を行うことにより，全体として高速に近似近傍探索を行う．また，近似近傍探索処理は，列方向で分割した部分ベクトル集合単位になることから，障害時の影響は近傍候補の減少のみに留めることが可能となる．

本報告の構成を述べる．2 章では EC サイトの商品提案機能で利用する特徴量について整理し，これらを利用した近似近傍探索の課題を整理する．3 章では，近似近傍探索エンジンの実装について述べる．4 章では提案手法の評価を行い，5 章でまとめとする．

2. 従来手法の課題

2.1 EC サイトの商品提案機能で利用する特徴量

EC サイトの商品提案機能では推薦根拠として，利用者間，商品間の相関や商品の特徴を表現する属性情報などを利用する．商品の提案は任意の観点での商品同士の類似性を根拠とすることから，これらの情報源を機械的に扱えるよう数値化し，任意の距離空間で近傍に位置する商品を求める．この数値化した商品特性を特徴量と呼び，一般に多次元ベクトルの形で表現する．

EC サイトの商品提案機能では，相関情報の元となる評価や購買履歴が不足している状況で，推薦精度が低下したり推薦対象に含まれないといった問題が発生する．この，いわゆるコールドスタート問題 [8] を回避するため，評価や履歴の蓄積に依存しない商品画像や説明文などの属性情報を用いた内容ベース型推薦を組み合わせて利用する [2]．

深層畳み込みニューラルネットワークを始めとする深層学習の発展 [4], [16] により，これまで適切な特徴量を導くことが難しかった画像に対しても，人の感性に近い，特性をよく表現する高精度な特徴量を得られるようになった．畳み込みニューラルネットワークは，画像の局所領域の特徴抽出を行う畳み込み層と，抽出した特徴を縮小し，位置感度を低下させるプーリング層を繰り返すネットワーク構造を持つ [5], [9], [12]．また，この層を多層にすることでより精度を向上させたものを深層畳み込みニューラルネットワークと呼ぶ．学習済みの深層畳み込みニューラルネットワークの上位層から得られる特徴量が，画像内容をよく表現することが示されており [11]，これらの特徴量が画像を用いた内容ベース型推薦の情報源として利用され始めている [20], [21]．

また，商品説明文を始めとするテキストによる内容ベース型推薦では，従来の単純なキーワード検索や TF-IDF 尺度による類似度比較で生じる単語の重要度や文脈が考慮されない課題を解決するため，前後の出現単語やパラグラフを考慮して学習を行うことで特徴表現の精度を向上させる Word2vec や Doc2vec などのモデルベースの手法 [14], [18] が利用されている．

これらの深層畳み込みニューラルネットワークや Word2vec によって出力される特徴量は，入力となる画素値や単語リストからなる多次元の特徴量からは次元数が削減されているものの，数百から数千次元の高次元ベクトルとなる．加えて，数十万から数百万件の商品を扱うような大規模な EC サイトでは，特徴量の集合の要素数も，扱

う商品数に比例して大規模なものとなっている．商品を自動的に提案する機能では，提案内容の的確さと充分な応答速度が求められることから，このような大規模かつ高次元ベクトルな集合に対して精度と速度を両立する近傍探索の手法が必要となる．

2.2 従来の近似近傍探索の課題

2.1 節で述べた大規模かつ高次元ベクトルな集合に対して，正確ではあるがデータ数と次元数に比例して計算量が増加する線形探索の利用は現実的でない．そこで，事前に近傍候補となる集団を求めておくことで，少数の近傍候補から計算量を抑えて近傍点を得る空間分割や局所性鋭敏型ハッシュといった手法が用いられる．しかしながら，高次元ベクトル集合においては正確な近傍を求めるための近傍候補を求める計算量が線形探索と同程度となると考えられており [22]，精度を犠牲にして近似解を用いることで計算量の増加に対処する近似近傍探索の手法が利用されている [1], [15], [19]．だが，近似近傍探索は精度と探索速度がトレードオフ関係にあることから，各手法のアルゴリズムが想定するデータ数や次元数を超えた大規模かつ高次元ベクトルな集合に対しては，精度と探索速度の両立が困難になる．

データ数と次元数の増加に伴う近似近傍探索の性能劣化に対処するため，探索対象を分割することで，各手法の想定するデータ数や次元数に収める手法がある．この場合，分割された対象ごとの探査結果または距離計算の結果が集約され，元のクエリに対する近傍集合が最終的に求められる．このような手法として，集合のベクトルを行に，ベクトルの持つ各次元を列とする行列とみなした場合，行方向での分割と列方向での分割の二種類がある．

行方向での分割では，クエリに対する近傍探索は分割された部分ベクトル集合ごとに行われる．各部分ベクトルの集合からの探索結果が近傍候補であり，これらの集合から元のクエリに対する近傍集合が求められる．この時，各近傍候補とクエリ間の距離は算出済みであるため，距離による並び替えのみが行われる [13]．行方向での分割は，各部分集合に対する探索は独立していることから分散処理が可能であるため，データ数に対する性能劣化への対処としては有効だが，高次元ベクトル集合に対する解とはならない．

直積量子化 [10] は，探索対象の高次元ベクトル集合を列方向に分割することで，独立した低次元ベクトル集合に対する距離計算の問題に分解する．近傍探索時は，クエリを同様に列方向に分割し，対応する部分ベクトル集合に対して距離計算を行い，その和をとる．この手法では，高次元ベクトル集合に対する性能劣化への対象として有効だが，近傍集合を求めるための計算量はデータ数に比例してしまう．転置インデックスの利用によりデータ数に対する計算量の増加を抑える方法も合わせて提案されているが，依然としてデータ数の増加に伴う性能劣化の課題は残る．

また，列方向へ圧縮する手法として，主成分分析 [17] を始めとする次元削減の手法が用いられる．これらの手法では情報量を損なわないよう表現に必要な特徴量の次元数を削減することで，次元数増加に対処する．しかしながら次元削減に伴う精度低下は避けられないため，充分な精度も求められる EC サイトの商品提案機能では削減に限界がある．

そのため，従来の近似近傍探索の課題を解決するためには，各手法の想定するデータ数や次元数に探索対象を抑えながらも分割結果の集約においてデータ数に依存しない方式が求められる．

2.3 EC サイトにおける近似近傍探索機能の課題

ここまで述べたように，EC サイトにおける商品提案機能には，推薦根拠となる商品特性を特徴量とした大規模かつ高次元ベクトルな集合に対する近似近傍探索が必要となる．利用者の効率的な閲覧を補助する商品提案機能には，提案内容の的確さと充分な応答速度が求められる．これらの性能低下は販売の機会損失につながるため，可用性の観点から同一データを保持する近似近傍探索機能を持つサーバを複数台用意することで負荷分散する手法が用いられる．一方で，計算資源の有効利用の観点から，EC サイトから求められる近似近傍探索の処理能力は負荷分散構成を前提としたものとなり，処理能力がサーバ台数に依存してしまうことから，サーバ障害時において，処理能力の低下が懸念される．サーバ障害時においても安定した応答速度を確保し，EC サイトにおける機会損失を回避するためには，障害時の影響を局所化し処理能力の低下を最低限に留める必要がある．

3. 提案手法

2 章で述べた，大規模 EC サイトの商品提案機能における大規模かつ高次元ベクトルな集合に対する近似近傍探索に関する課題を解決するためには，以下の要件を満たす必要がある．

(1) 近似近傍探索の精度と速度を両立するため，アルゴリズムの想定するデータ数と次元数に収まるよう探索対象を分割し，データ数に依存しない集約が行える．
(2) 近似近傍探索の可用性を担保するため，分散構成における障害時の影響が局所化されている．

筆者らは商品特性をよく表現しており高精度に類似度が比較可能な高次元かつ密なベクトルの集合において，クエリに対する高次元ベクトル集合の近傍探索結果の上位集合が，クエリと高次元ベクトル集合を任意の次元数で等分した部分ベクトル単位で近傍探索した結果の上位集合と類似しやすい特性があることを見出した．このような集合として 2.1 節で述べた，学習済み深層畳み込みニューラルネッ

トワークを特徴抽出器として利用して画像から得られる特徴量集合や，テキストを分散表現へ変換する Word2vec のネットワークから得られる特徴量集合がある．このような部分が類似しているものは全体としても類似する可能性が高いという特性のベクトル集合に対しては，列方向に分割した部分ベクトル単位での探索に分解し，その探索結果の和集合である近傍候補から再度近傍探索を行うことにより，全体として高速に近似近傍探索を行える．また，部分ベクトル集合単位に分散可能な構成であることから，一部の部分ベクトル集合からの近傍候補が得られずとも探索速度を落とさずに近似近傍探索の機能を継続することができる．以降の節では，提案手法を用いた具体的な探索処理を説明する．

3.1 探索対象となるベクトル集合の分割

ここでは，対象となるベクトル集合を $Y \subset \mathbb{R}^D$ と置く．これを $D^* = D/m$ 次元の部分ベクトル集合 $S_j(1 \leq j \leq m)$ に等分する．なお，部分ベクトル集合の持つ要素は各集合間で重複しないものとする．

3.2 近似近傍探索

ベクトル集合 Y に対する近似近傍探索は次のように行う．はじめに，クエリベクトル $q \in \mathbb{R}^D$ を $q_j(1 \leq j \leq m)$ に等分し，対応する添え字 j の部分ベクトル集合 S_j に対して上位 n 件を得る近傍探索を行う．

$$NN(q_j) = \arg\min_{s \in S_j} d(q_j, s) \tag{1}$$

ここで，距離関数 d にはユークリッド距離を想定している．なお，各部分ベクトル集合における近傍探索の手法は問わない．

次に，各部分ベクトル集合に対する (1) 式で得られた $NN(p_j)$ の結果を n 個の識別子からなる集合 N_j とし，全ての N_j の和集合を N としてこれを近傍候補とする．

最後にベクトル集合 Y のうち，近傍候補 N に対応するベクトル集合 $YN \subset \mathbb{R}^D$ に対して正確な類似度を比較するために元のクエリベクトル q との線形探索を行う．

$$NN(q) = \arg\min_{yn \in YN} d(q, yn) \tag{2}$$

ここで，$|YN|$ は最大 $n * m$ 個であり，(2) 式における線形探索の計算量はデータ数に依存しない．

3.3 提案手法の実装

上述の提案手法に対し，筆者らは探索速度を考慮した実装である Sanny[*1] を開発，OSS として公開した．Sanny の処理フローを**図 1** に示す．

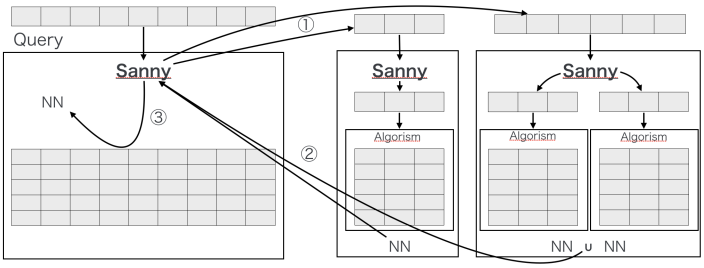

図 1 Sanny の処理フロー
Fig. 1 Process flow of Sanny.

提案手法は以下の 3 点の工程を要しており，各工程の実行時間を短縮することで，全体として高速な近似近傍探索が可能となる．

(1) 各部分ベクトル集合での近傍探索
(2) 各近傍探索結果を近傍候補として集約
(3) 固定数の近傍候補から線形探索

ここで，(1) 工程では各部分ベクトル集合は独立しており，並列に近傍探索を行うことができる．各部分ベクトル集合は元のベクトル集合に比べて低次元のベクトル集合に対する近傍探索となることから次元数に依存する計算量が削減され，並列処理によって全体として探索処理時間が削減される．また，(2) と (3) 工程では元のベクトル集合に比べて小規模かつ固定数の近傍候補に対して線形探索を行うため，探索時間は短くなるが，Sanny では線形探索処理を簡略化または並列化することで高速化する．線形探索処理の近似には，(3) 式を用いる．ここでは，距離関数はユークリッド距離とし，a を探索対象のベクトル，b をクエリベクトルとする．

$$\begin{aligned} argmin_a(a-b)^2 &= argmin_a a^2 - 2ab + b^2 \\ &= argmin_a a^2 - 2ab \end{aligned} \tag{3}$$

b はクエリベクトルであり，全てのベクトル集合に対する計算において b^2 は同じ値となることから省略することができる．また a^2 はクエリベクトルに依存しないため，ベクトル集合が与えられた際に事前に求めておくことができる．このため線形探索時には $2ab$ を求め，事前算出済みの a^2 に対する減算のみ必要となる．また，線形探索処理を実行する端末の CPU が Single Instruction Multiple Data (SIMD) 演算に対応している場合，ユークリッド距離を求める積和演算においてベクトルの複数次元に対して同時計算ができるため，Sanny ではこれに対応することで線形探索の高速化を図った．

4. 評価と考察

4.1 データ特性による有効性の評価

提案手法が有効なデータ特性を検証するため，複数の

[*1] https://github.com/monochromegane/sanny

データセットに対して提案手法を適用した．本報告では，データ数が 6 万件の各データセットを 4 つの部分ベクトル集合に分割し提案手法による上位 10 件の近傍集合に対する適合率を確認する．なお，部分ベクトル集合に対する近傍探索は線形探索を採用し，部分ベクトル集合からの上位 30 件の近傍集合を近傍候補として利用した．評価対象となる 5 つのデータセットについて述べる．一つ目は，画像を高次元密ベクトルの特徴量に変換する深層畳み込みニューラルネットワークのモデルの一つである Inception-v3[3] のうち，ImageNet の画像群で学習済みのモデルから得られたデータセットである．このモデルから得られる特徴量は 2048 次元のベクトルであり，EC サイトのプロダクション環境の商品提案機能で利用している商品画像のうち，6 万件を評価対象とした．二つ目は，単語を高次元密ベクトルの分散表現に変換する Word2vec のうち，日本語版 Wikipedia を用いて学習済みのモデル[*2]から得られたデータセットである．このモデルから得られる特徴量は 200 次元のベクトルであり，利用可能な単語のうち 6 万件を評価対象とした．三つ目は，機械学習ベンチマークで利用される Fashion-MNIST[*3]である．28x28 のグレースケール画像であり，786 次元の疎なベクトル集合である．提供されている 6 万件を評価に利用した．残り二つのデータセットは乱数を用いた．各次元の値を一様または正規分布に従う乱数とし，200 次元で 6 万件のデータセットとした．

各データセットに対する提案手法による適合率の比較を**表 1** に示す．乱数を元にしたデータセットと比較して，画像を元にした Inception-v3，日本語の単語を元にした Word2vec や服飾画像から成る Fashion-MNIST のデータセットが提案手法の精度が高い．また，画素値を元にした疎なベクトル集合である Fashion-MNIST データセットよりも，密なベクトル集合である Inception-v3 や Word2vec データセットが精度が高い．これらから，元の情報が意味や特性を表現し得る情報を持っており，高次元かつ密ベクトルで表現される特徴量から成るデータが提案手法の前提とする特性を持つと推測される．

提案手法では，部分ベクトルごとの次元数と，部分ベクトルごとに得られる近傍集合の数が少ないほど，個別の近傍探索ならびに結果の集約処理が高速になる．一方で，分割数を増やすことで部分の類似と全体の類似の関係性が薄れることが想定された．そこで，分割数と部分ベクトルごとの近傍数による適合率の変化を検証した．検証は，先の検証で提案手法の有効性が見込めるデータセットとして，画像を元にした Inception-v3，日本語の単語を元にした Word2vec から得られる特徴量を用いた．データセットの件数はそれぞれ，約 95 万件と約 100 万件である．結果をそれぞれ**図 2** と**図 3** に示す．横軸は部分集合ごとの近傍数であり，20 件から 1000 件までの評価結果を対数表示している．縦軸は適合率を表す．提案手法では，分割数を増加すると同一近傍数において精度が下がる関係にあることが分かる．また，各データセットでの分割後の次元数が 100 以下になると精度の落ち込みが大きく，データ数の 0.1%を超える近傍数が必要になることから，速度と精度を両立する近似近傍探索のためには，これらの分割数内で最適な近傍数を求めることが必要になることが分かる．

[*2] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/
[*3] https://github.com/zalandoresearch/fashion-mnist

表 1 提案手法による適合率
Table 1 Precision by proposed method.

データセット	部分 1	部分 2	部分 3	部分 4	提案手法
Inception-v3	0.57	0.37	0.50	0.65	**0.97**
Word2vec	0.67	0.68	0.68	0.69	**0.97**
Fashion-MNIST	0.10	0.23	0.31	0.16	**0.63**
一様乱数	0.05	0.05	0.05	0.05	**0.18**
正規分布	0.05	0.04	0.04	0.04	**0.17**

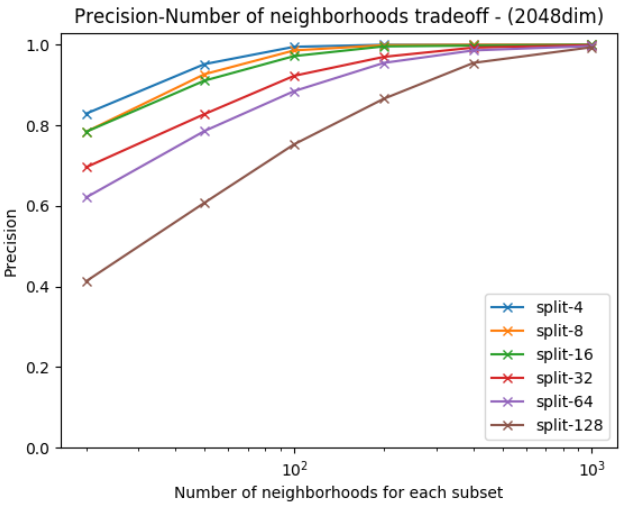

図 2 2048 次元のデータセットにおける適合率と部分集合ごとの近傍数のトレードオフ評価
Fig. 2 Precision-Number of neighborhoods tradeoff.

4.2 精度並びに速度に関する評価

提案手法による精度並びに探索速度の改善を検証するため，4.1 節で評価したデータセットのうち，提案手法が有効で大規模かつ高次元ベクトルな集合であった商品画像を元にした特徴量に対する評価を行なった．

本報告では，近似近傍探索の精度と速度の両立を評価するため，クエリベクトルに対する上位 10 件の近傍集合の適合率と探索速度を計測する．評価対象となるベクトル集合からランダムに選択した 500 件をクエリベクトルとし，残りのベクトル集合から近傍探索を行う．また，評価対象の手法には，近傍探索アルゴリズムである空間分割を利用

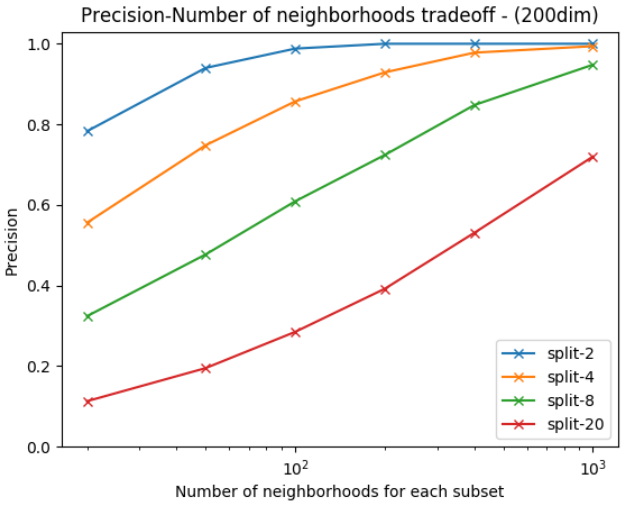

図 3 200 次元のデータセットにおける適合率と部分集合ごとの近傍数のトレードオフ評価
Fig. 3 Precision-Number of neighborhoods tradeoff.

したライブラリの Annoy[19][*4]，NGT[15][*5]，そして局所性鋭敏型ハッシュを利用した LSHForest[1][*6]を用いる．これらの手法を従来通り利用したものと，提案手法における部分ベクトル集合単位での近傍探索アルゴリズムとして利用した場合における性能改善を検証した．各手法ごとの近傍探索のパラメタを**表 2** に示す．なお，LSHForest については，アルゴリズム特性上，多くのメモリを使用することから検証機の制限上，実時間で検証が終わるパラメタによる検証に留めている．評価方法は近似近傍探索ライブラリのベンチマークである ann-benchmarks[*7]で比較可能とするため，同様の方法で結果を取得した．具体的には，評価手法ごとに近似近傍探索の挙動を調整するためのパラメタを変えながら，全クエリによる近傍探索を 3 回実行し，適合率は平均，探索速度には最小のものを採用している．

評価結果を**図 4** に示す．X 軸は適合率で右に向かうほど精度が高い．Y 軸は探索速度の逆数を表しており上に行くほど速い探索が行われたことを示している．なお，複数の探索が同一の適合率となる場合は速度が速いものを図示している．提案手法の適用により従来手法と比較して，同程度の精度での探索時間短縮，または頭打ちとなっていた精度が向上するといった性能の改善が確認できた．これらは提案手法の適用によって，元のベクトル集合よりも低次元な空間が対象となることで，より多くの近傍候補を短時間で探索できるようになったためである．また，従来手法では高次元ベクトル集合に対する精度の限界についても複数の低次元空間探索問題に分割することで，精度を向上させることが可能となっている．

[*4] https://github.com/spotify/annoy
[*5] https://github.com/yahoojapan/NGT
[*6] https://github.com/ekzhu/lsh
[*7] https://github.com/erikbern/ann-benchmarks

表 2 手法ごとの近傍探索パラメタ
Table 2 Parameters of nearest neighbors search.

手法	パラメタ	値
Annoy	Tree	100〜400
	SearchK	100〜400000
NGT	Edge	0〜80
LSH	Hash table size	20
	Hash value size	20
	Slot size	5.0
Sanny-Annoy	Split	8
	Top	20〜100
	Tree	100〜400
	SearchK	100〜400000
Sanny-NGT	Split	8
	Top	20〜100
	Edge	0〜80
Sanny-LSH	Split	8
	Top	20〜100
	Hash table size	20
	Hash value size	20
	Slot size	5.0

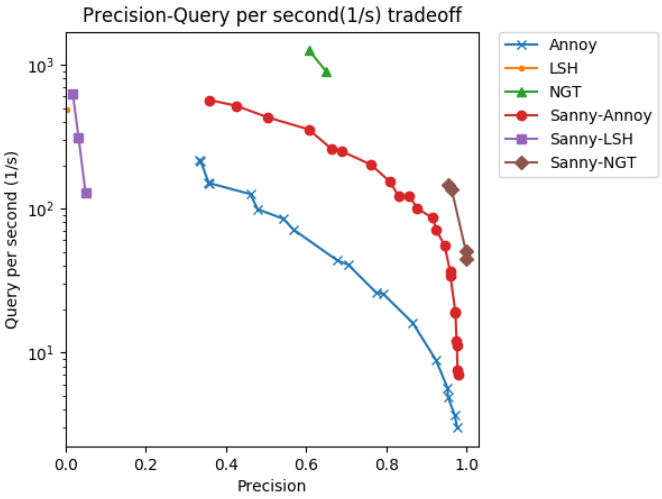

図 4 適合率と探索速度のトレードオフ評価
Fig. 4 Precision-Query per second tradeoff.

5. まとめ

本報告では，大規模 EC サイトで商品を提案するために扱う大規模かつ高次元ベクトルな集合に対して，精度と速度を両立した分散可能な近似近傍探索エンジン Sanny を提案した．商品の提案機能で利用される商品画像や説明文を情報源とするニューラルネットワークから得られる特徴量において部分の類似が全体の類似と関連することに着目することで，列方向の分割時の課題であったデータ数に依存する集約処理を短縮し，従来手法と比較して精度と探索速度が向上することを確認した．

提案手法では，部分ベクトル集合の分割は，直積量子化

に倣い単純な等分を行なったが，データの偏りを考慮したより効果的で少ない次元数での分割の可能性も検討したい [7]．また，分散処理可能な特性を活かした拡張性ある構成としてネットワークをまたぐ大規模な近似近傍探索エンジンとしての実装と評価も進めていきたい．

参考文献

[1] Bawa, Mayank, Tyson Condie, and Prasanna Ganesan. "LSH forest: self-tuning indexes for similarity search." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.
[2] Burke, Robin. "Hybrid recommender systems: Survey and experiments." User modeling and user-adapted interaction 12.4 (2002): 331-370.
[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going Deeper with Convolutions, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015
[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the Inception Architecture for Computer Vision, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016
[5] K. Fukushima. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological Cybernetics, 36(4):93-202, 1980
[6] Galletta, Dennis F., et al. "Web site delays: How tolerant are users?." Journal of the Association for Information Systems 5.1 (2004): 1.
[7] Ge, Tiezheng, et al. "Optimized product quantization for approximate nearest neighbor search." Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013.
[8] Hyung Jun Ahn, A new similarity measure for collaborative filtering to alleviate the new user coldstarting problem, Information Sciences 178, pp. 37-51, 2008
[9] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, Trevor Darrell, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In Proc. ICML, 2014
[10] Jegou, Herve, Matthijs Douze, and Cordelia Schmid. "Product quantization for nearest neighbor search." IEEE transactions on pattern analysis and machine intelligence 33.1 (2011): 117-128.
[11] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
[12] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998
[13] Muja, Marius, and David G. Lowe. "Scalable nearest neighbor algorithms for high dimensional data." IEEE transactions on pattern analysis and machine intelligence 36.11 (2014): 2227-2240.
[14] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
[15] Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki, "On Approximately Searching for Similar Word Embeddings", In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics（ACL 2016）, pages 2265-2275. Association for Computational Linguistics, 2016.
[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vision 115, 3, pp. 211-252, December 2015
[17] Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2.11 (1901): 559-572.
[18] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International Conference on Machine Learning. 2014.
[19] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Wenjie Zhang, Xuemin Lin, Approximate Nearest Neighbor Search on High Dimensional Data — Experiments, Analyses, and Improvement (v1.0), In Proc. 2016
[20] S. Zhang, L. Yao, and A. Sun, "Deep learning based recommender system: A survey and new perspectives," arXiv preprint arXiv:1707.07435, 2017.
[21] 三宅 悠介, 松本 亮介, 力武 健次, 栗林 健太郎, 特徴抽出器の学習と購買履歴を必要としない類似画像による関連商品検索システム, 研究報告インターネットと運用技術（IOT）,2017-IOT-37(4),1-8 (2017-05-18) , 2188-8787
[22] 和田俊和. "最近傍探索の理論とアルゴリズム." 研究報告コンピュータビジョンとイメージメディア (CVIM) 2009.13 (2009): 1-12.