Apache Mahoutを活用したデータ分析事例のご紹介 ～50分でわかる機械学習～

株式会社エクサ

テクノロジーイノベーション部

森本淳司

お話する内容

1. はじめに
2. 社内システムへの機械学習適用事例
   - 登録漏れスキルを「レコメンデーション」
   - エクサSEの特徴を「クラスタリング」
3. 将来構想
   - SEの特性で「分類」
4. おわりに

機械学習は50年以上前に定義

Field of study that gives computers the ability to learn without being explicitly programmed

1959 Arthur Samuel

人の持つ学習能力と同様の機能を
コンピュータで実現しようとする技術・手法

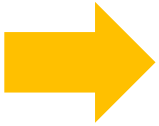
Arthur Samuel (IBM)
世界初のコンピュータチェッカーの開発
世界初のソフトウェアによるハッシュ
テーブルを考案

応用例から「機械学習」を理解する

- スパムフィルタリング
- 郵便番号の自動認識
- アンケート回答者の傾向をグループ分け

スパムフィルタリングを実現する3ステップ

1. 事前に用意したデータ（迷惑メール）をコンピュータに読み込む
2. アルゴリズムを元に有用なルール（迷惑メールを特徴付けているもの）を抽出する
3. 別のデータ（迷惑メール）をコンピュータに与え、新しい事象（迷惑メールか否か）を予測する



ビッグデータ×機械学習

ビッグデータと機械学習で何ができる？

ビッグデータの特徴

- データ量 (Volume)
- データ種類 (Variety)
- データの生成速度 (Velocity)

2012 ガートナーが定義

ビッグデータ×機械学習の代表ビジネス

異常検知

- 工場機械などの故障の予兆を知りたい
- センサーデータからモデルを構築
- モデルに適合しない外れ値の異常具合を計算し、故障の予兆に役立てる

ビッグデータと機械学習で何ができる？

ビッグデータの特徴

・データ量 (Volume)

「ビッグデータ」という単語に踊らされないで

確実に機械学習ビジネスを発展させていくために

異常検知

・ モデルに適合しない外れ値の異常具合を計算し、故障の予兆に役立てる

第一歩として、
機械学習の仕組みと応用例を学びましょう！

- **機械学習ライブラリ Mahout を活用した社内システム課題の取り組み**
  1. **レコメンデーション**
  2. **クラスタリング**
  3. **分類（将来構想として）**

社内システムへの機械学習適用事例のご紹介

技術管理システムの課題を機械学習で解決できないか

SEスキルを可視化する技術管理システム

スキルの強みを伸ばし、弱点を補強

・スキルを可視化しSE育成に活用
・個人だけでなく組織のスキルも可視化できる

柔軟な検索で的確にSEを見つける

・プロジェクトに必要なスキルを任意に設定でき、
マッチするSEを簡単に見つけることができる

活動履歴が一目瞭然！タイムライン表示

・他システムと連携し、スキルを取得した成長時期、
経験した業務、受講した研修を一括表示できる

技術管理システムのさらなるパワーアップ

① スキル登録漏れが発生

- スキル数が2000個以上あるため、登録漏れが発生する
- 登録漏れを防止するサポート機能が必要

② エクサSEの特徴を知りたい

- エクサはIBM/JFE出身者も多く、多文化企業であり、強みを把握しきれていない
- エクサSEの特徴を把握し、ビジネス戦略に活かしたい

③ 適切な研修を受講したい

- キャリアデザインにマッチする研修を受講したい
- 誰がどのキャリアに適合するか分類できる必要がある

技術管理システムのさらなるパワーアップ

① スキル登録漏れが発生

- スキル数が2000個以上あるため、登録漏れが発生する
- 登録漏れを防止するサポート機能が必要

② エクサSEの特徴を知りたい

- エクサはIBM/JFE出身者も多く、多文化企業であり、強みを把握しきれていない
- エクサSEの特徴を把握し、ビジネス戦略に活かしたい

機械学習を適用した実績について紹介します

技術管理システムのさらなるパワーアップ

① スキル登録漏れが発生

- スキル数が2000個以上あるため、登録漏れが発生する
- 登録漏れを防止するサポート機能が必要

② エクサSEの特徴を知りたい

- エクサはIBM/JFE出身者も多く、多文化企業であり、強みを把握しきれていない
- エクサSEの特徴を把握し、ビジネス戦略に活かしたい

③ 適切な研修を受講したい

- キャリアデザインにマッチする研修を受講したい
- 誰がどのキャリアに適合するか分類できる必要がある

技術管理システムのさらなるパワーアップ

将来構想の取り組みとして
紹介します

③ 適切な研修を受講したい

- キャリアデザインにマッチする研修を受講したい
- 誰がどのキャリアに適合するか分類できる必要がある

技術管理システムのさらなるパワーアップ

① スキル登録漏れが発生

- スキル数が2000個以上あるため、登録漏れが発生する
- 登録漏れを防止するサポート機能が必要

② エクサSEの特徴を知りたい

- エクサはIBM/JFE出身者も多く、多文化企業であり、強みを把握しきれていない
- エクサSEの特徴を把握し、ビジネス戦略に活かしたい

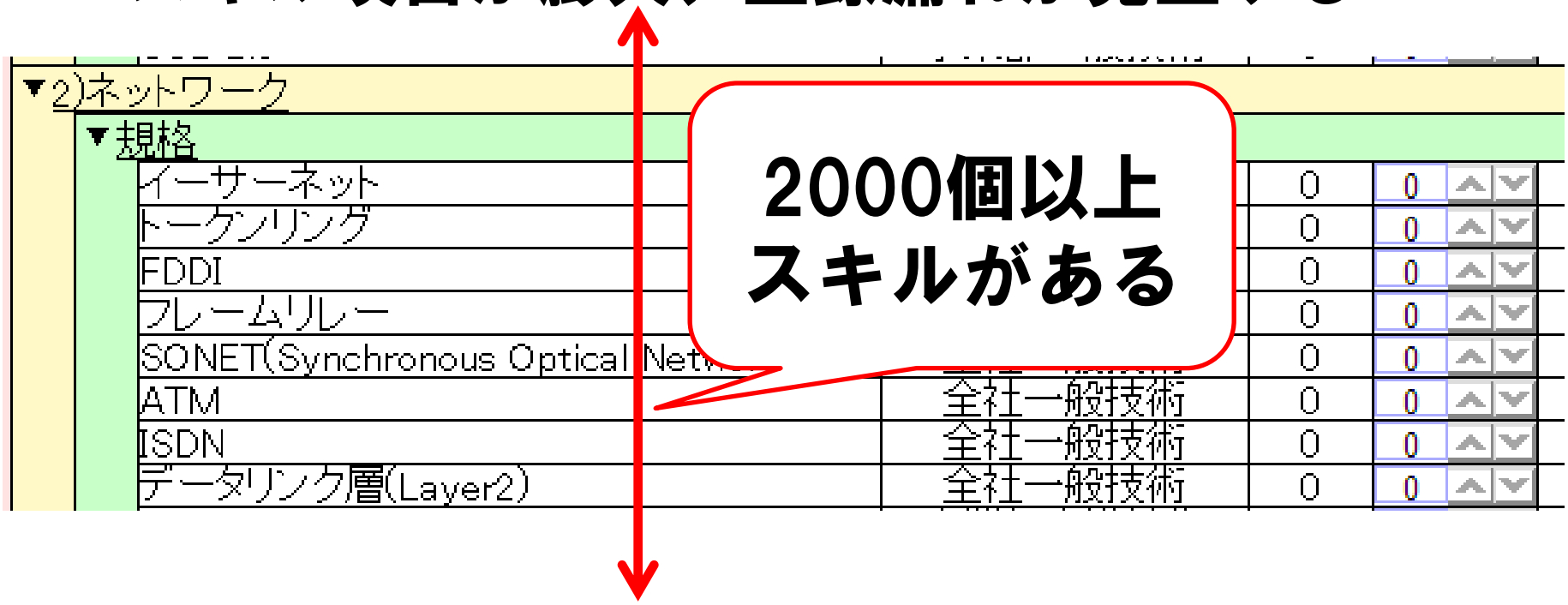
③ 適切な研修を受講したい

- キャリアデザインにマッチする研修を受講したい
- 誰がどのキャリアに適合するか分類できる必要がある

① スキル登録をサポートする

① スキルの登録が不十分

- IT業界はスキル項目が多い
  - 幅広い知識が必要
  - 業務知識
- スキル項目が膨大、登録漏れが発生する



いかに登録漏れのない仕組みを構築するか

見やすいレイアウトでユーザビリティを向上させる

・ツリー表示で全体像と詳細の切り替えを容易にする

使いやすい検索で目的のスキルを瞬時に表示

・軽快なサジェスト機能で検索をサポートする

登録の漏れていそうなスキルをシステムがSEに提案

・どのようにすればよいのか？

レコメンデーションで登録サポート

- イメージはアマゾンの商品おすすめ機能
- スキル間の類似性が高く、未登録のスキルをSEに提示する

→ どのようにレコメンデーションするのか

レコメンデーションの代表的な2つの方法

ユーザーベース　～似ているSEでレコメンド～

・登録しているスキルを元に、似ているSEを計算
・類似SEが登録済みかつ、自分が未登録のスキルを提案

アイテムベース　～似ているスキルでレコメンド～

・スキル間の類似度を計算
　・一つのスキルと残り全てのスキルの類似度を計算
・最近登録したスキルの中で、未登録かつ、類似度の高いスキルを提案

スキル一つ一つに注目するためアイテムベースを選択
具体的な計算方法を見てみよう

直感で理解するアイテムベースレコメンド

① 登録状況をマトリクスで示す

	SQL	DB2	会計業務
田中	○	○	-
鈴木	○	-	○
高橋	○	-	○

② スキル間の組合せ数をカウント

	SQL	DB2	会計業務
SQL	-	-	-
DB2	1	-	-
会計業務	2	0	-

	SQL	DB2	会計業務
SQL	-	-	-
DB2	0.33	-	-
会計業務	0.66	0.0	-

③ スキル間で割合を求める

直感で理解するアイテムベースレコメンド

① 登録状況をマトリクスで示す

	SQL	DB2	会計業務
田中	○	○	-
鈴木	○	-	○
高橋	○	-	○

SQLを登録した人は
DB2よりも会計業務を
登録した人の方が多い

SQLを登録した人には
会計業務を提案する

	SQL	DB2	会計業務
SQL	-	-	-
DB2	0.33	-	-
会計業務	0.66	0.0	-

③ スキル間で割合を求める

どのような技術でレコメンデーションするか

代表的な機械学習技術

Apache Mahout

・Hadoop、スタンドアローン両者で動作可能
・実装アルゴリズム多数

Jubatus

・日本生まれのオープンソースプロダクト
・日本語文献、ハンズオンが充実

R

・その場限りの分析に適している
・グラフなど視覚機能も充実

代表的な機械学習技術

Apache Mahout

・Hadoop、スタンドアローン両者で動作可能
・実装アルゴリズム多数

・Mahoutを選択した理由
  - スタンドアローン環境のJavaバッチで運用したい
  - データが大きくなればHadoopに移行できる
  - Jubatusは選定時にクラスタリングアルゴリズムが実装さていなかったため、対象外とした

レコメンデーション結果

- より登録しやすくなったスキル登録画面
  - 類似度を元に登録の漏れのありそうなスキルを提案
  - 好意的なコメントをいただく

- 実装所感
  - わかりやすいMahout API、類似度の計算は容易

課題と今後の展望

- 本当に登録しやすくなったのか？
  - 好意的なコメントはいただいた
  - 本当に登録してほしいものを提案できてる？
  - 定量的に変化を把握したい

- 定量データを取得する仕掛け
  - 提案したスキルが実際に登録されているのかデータを取得し、レコメンデーション機能を評価する

② エクサSEの特徴を知りたい

エクサの強み/弱みはなにか

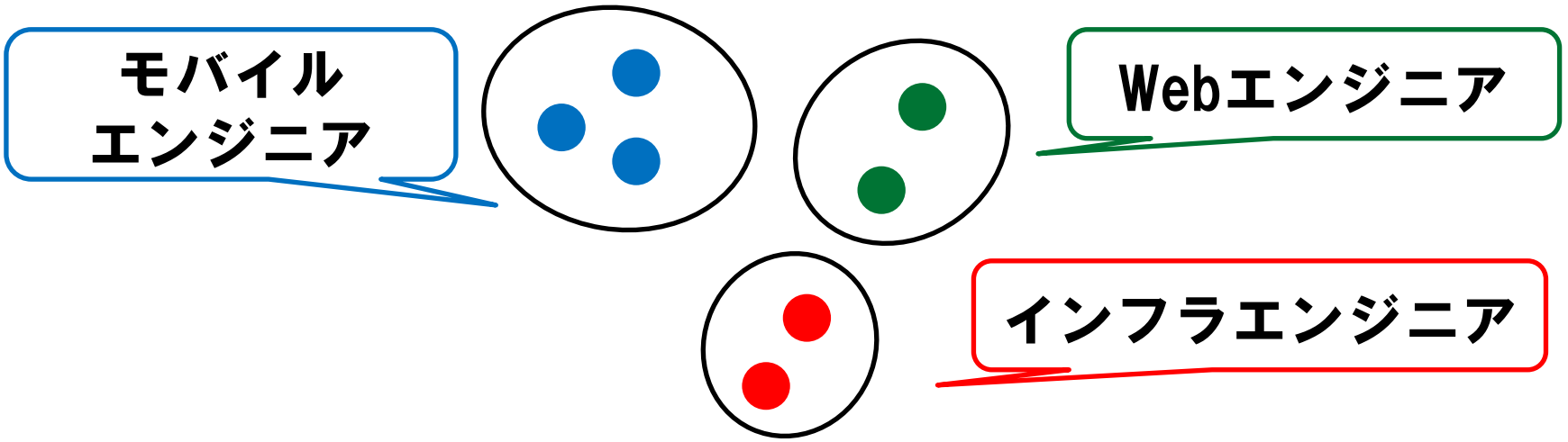
・回答は人/組織によって異なる
  - 上流/下流
  - アプリ/インフラ
  - 品質/納期/コスト
・多文化企業（エクサ/IBM/JFE）の影響も

➡ 特徴を定量的把握し、ビジネス戦略に活かす

スキルの観点からSE集団の特徴を見出す

・SEが登録しているスキルを元にグループ化

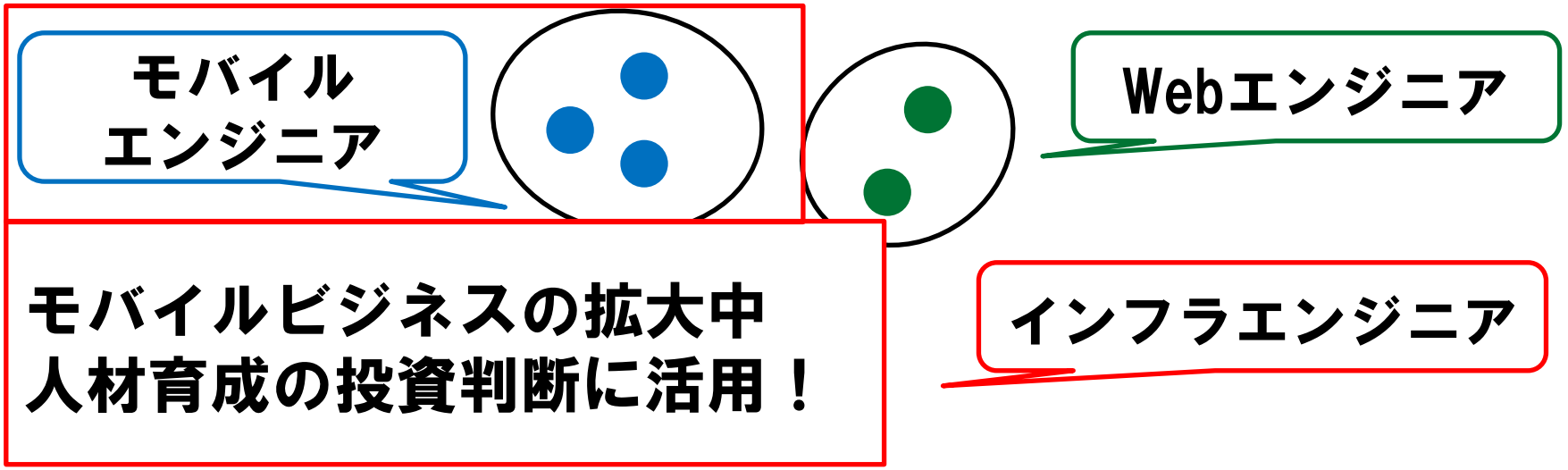
例）7人のSEを3つのグループに分け、特徴を見つける



スキルの観点からSE集団の特徴を見出す

・SEが登録しているスキルを元にグループ化

例）7人のSEを3つのグループに分け、特徴を見つける



グループ化でエクサSEの特徴がわかる
どのようにグループ分けを行うか

人によるグループ化は手間がかかり、ブレる

検索機能を用いて、人がグループ化する

- 「Webエンジニア」「インフラエンジニア」などのカテゴリを定義する
- その結果にマッチするSEを見つけ、カテゴライズする

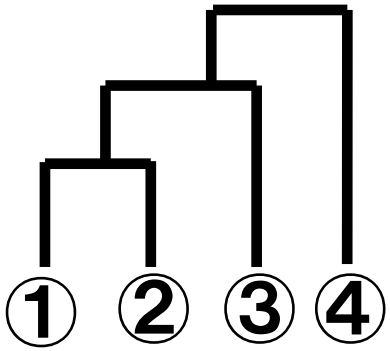
- データ規模(SE数)が大きい場合はグループ化するコストがかかるため、自動化したい
- 事前にどのようなカテゴリが存在するか定義する必要がある
- カテゴリ定義は人によって異なるため、結果がブレてしまう
- カテゴリの定義は時代とともに変化する

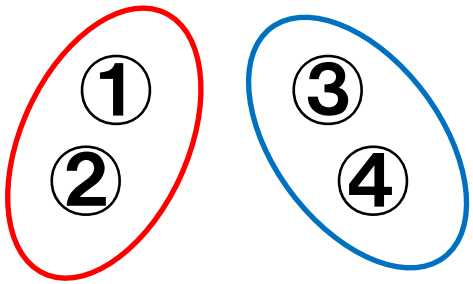
人によるグループ化は不可能
クラスタリングで機械的なグループ化を実現する

似たもの同士でまとめるクラスタリング

・クラスタリングは定められた規則の元、複数の要素(SE)を類似要素のグループにまとめること

階層的クラスタリング	非階層的クラスタリング
・ウォード法 ・重心法	・K平均法 ・スペクトラルクラスタリング





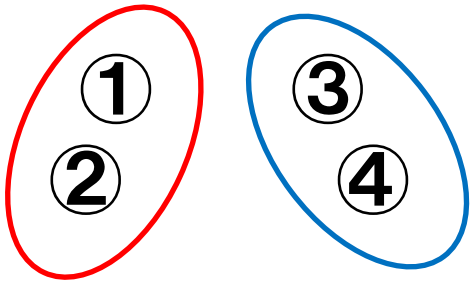
似たもの同士でまとめるクラスタリング

・クラスタリングは定められた規則の元、複数の要素(SE)を類似要素のグループにまとめること

・SEのグループを非階層的に表現したい
・代表的なアルゴリズムであるK平均法を選択
・「K」はクラスタ数
・Mahoutを利用

非階層的クラスタリング

・K平均法
・スペクトラルクラスタリング



K平均法の適用で何を求めることができるか

・グループとその重心を求めることができる

2つのグループにクラスタリングする

	SQL	DB2	会計業務
田中	2	4	-
鈴木	3	-	2
高橋	3	-	1

Aグループ（田中）
重心（2, 4, 0）

Bグループ（鈴木、高橋）
重心（3, 0, 1.5）

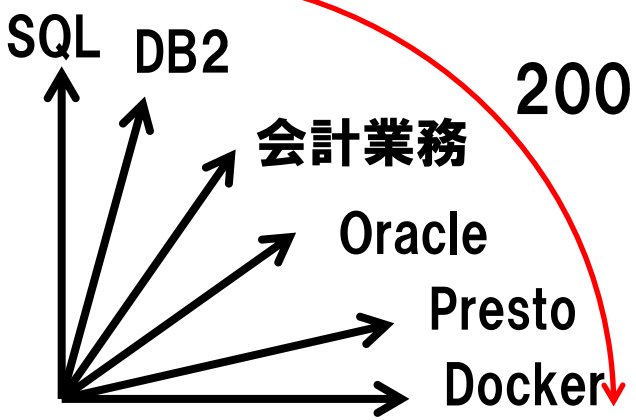
② エクサSEの特徴を知りたい

クラスタリング結果は直感的！
しかし、結果をどのように視覚化するか
- グループの違いを明確にしたい
- 3次元表示？

	SQL	DB2	会計業務
田中	2	4	-
鈴木	3	-	2
高橋	3	-	1

実際の
スキル数は
2000個以上！

2000次元以上の多次元
データを人がわかる形で
どのように表現するか



2000次元以上！

多次元データを視覚的にわかりやすく表現する

・人が分かる形＝2次元のグラフで表現するのがベスト

➡ 多次元データを低次元で表現する方法

・多次元尺度構成法
  ・重心間の距離から位置関係を計算する方法
・主成分分析
  ・分散の大きい要素を影響力のある要素と見なし、比重をおいて次元を削減する方法

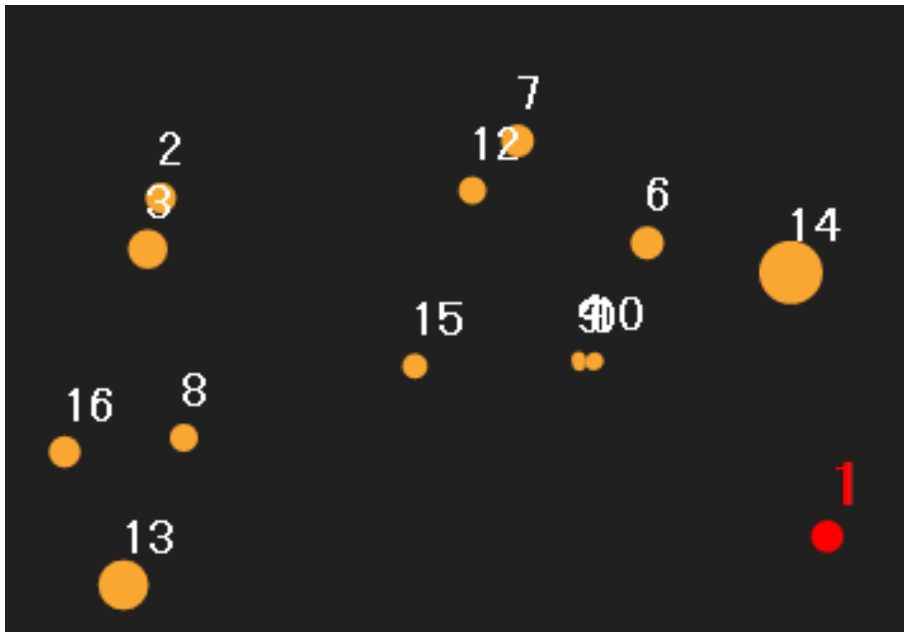
多次元データを視覚的にわかりやすく表現する

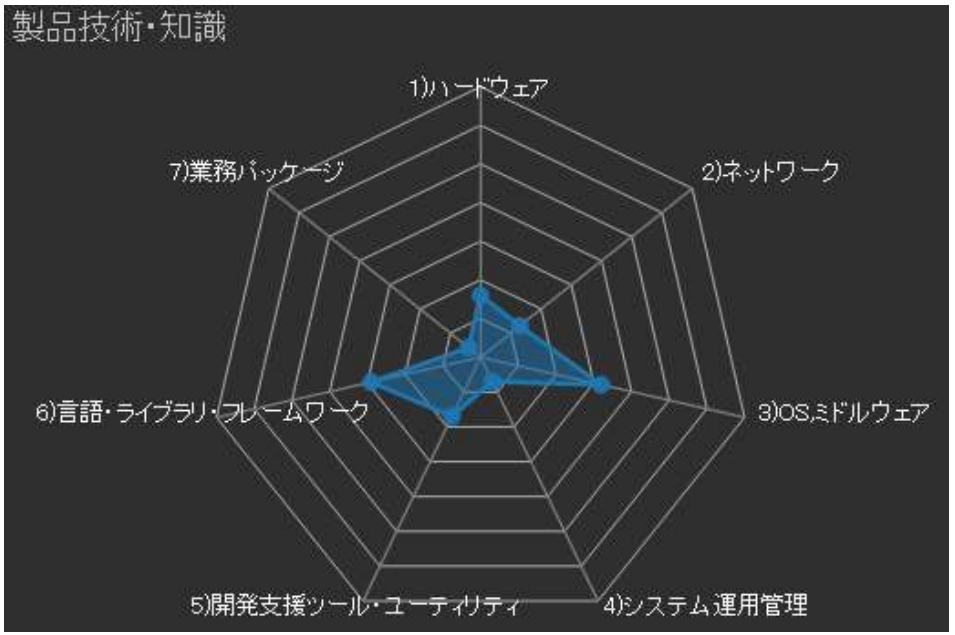
・多次元尺度構成法で2次元グラフの表現が可能

- 多次元尺度構成法
  - 重心間の距離から位置関係を計算する方法
- 主成分分析
  - 分散の大きい要素を影響力のある要素と見なし、比重をおいて次元を削減する方法

クラスタリング結果

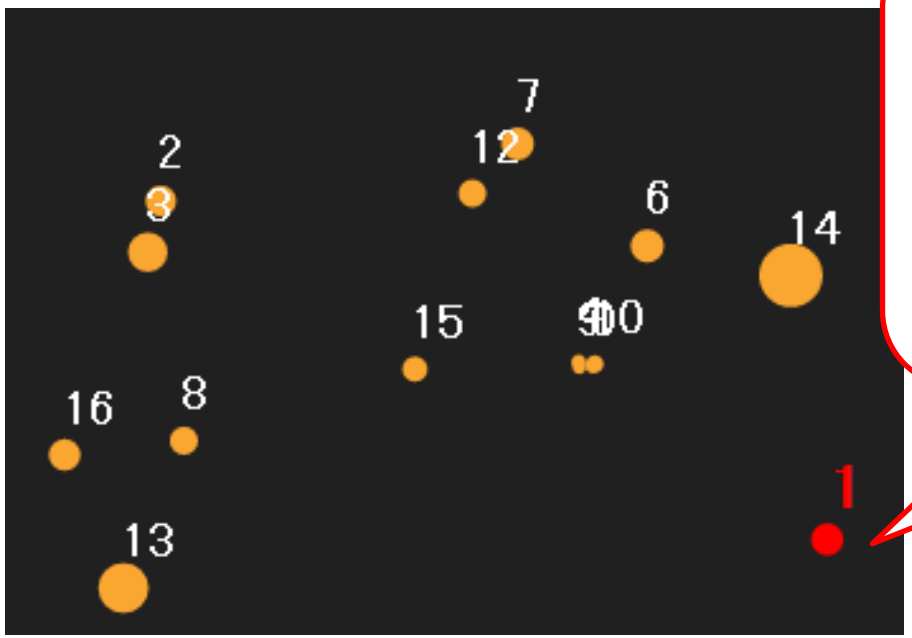
・スキルでSEをグループ化できた
・スキルの似ているグループ間の遠近がわかる
・グループのスキル特徴がわかる



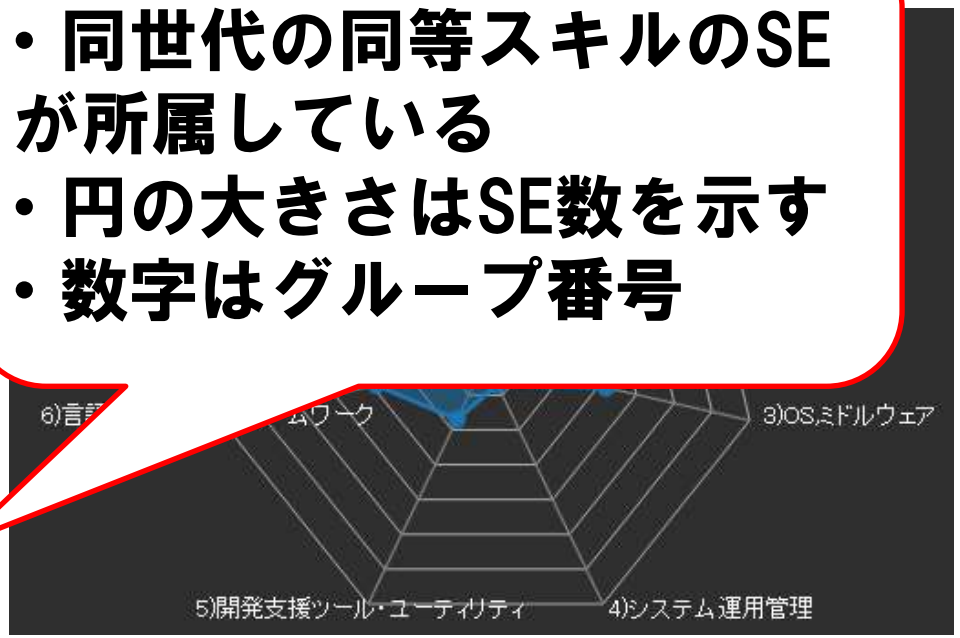


クラスタリング結果

・スキルでSEをグループ化できた
・スキルの似ているグループ間の遠近がわかる
・グループのスキル特徴がわかる



・同世代の同等スキルのSEが所属している
・円の大きさはSE数を示す
・数字はグループ番号

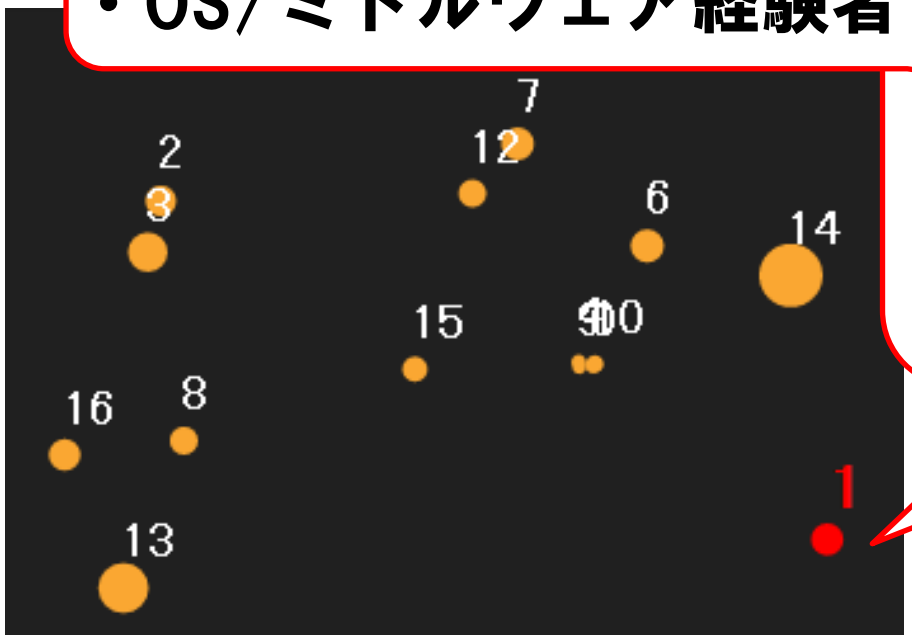


クラスタリング結果

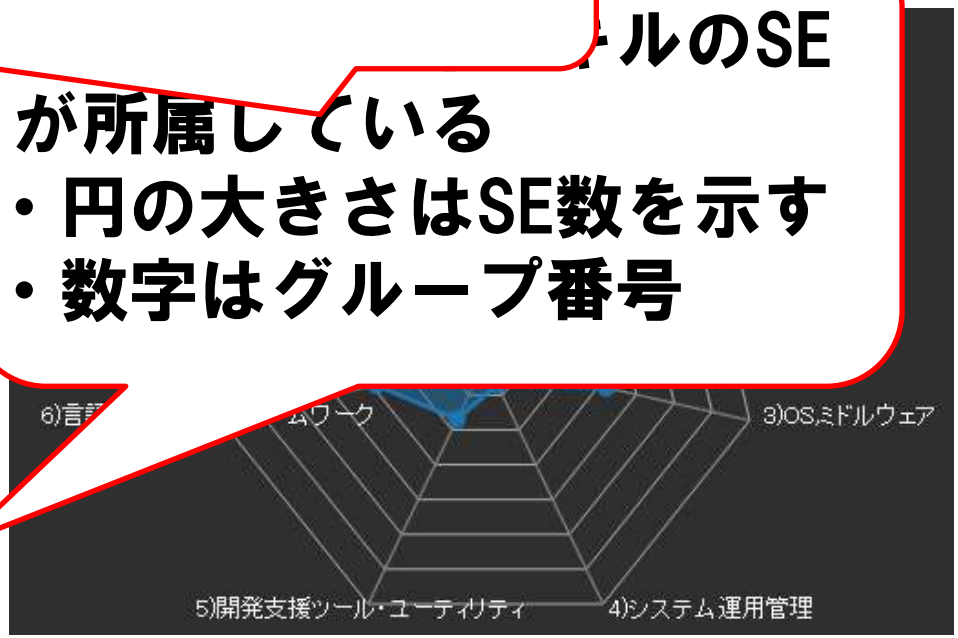
・スキルでSEをグループ化できた

・スキルの似ているグループ間の遠近がわかる

・言語/ライブラリ/フレームワーク経験者
・OS/ミドルウェア経験者



ルのSE
が所属している
・円の大きさはSE数を示す
・数字はグループ番号



課題と今後の展望

- 精度が不十分
  - SEのスキル登録が不十分
  - レコメンデーション機能でサポート
- 特徴を見い出せないグループもある
  - 様々な世代/スキルのSEグループが存在
  - 全社員のクラスタリングで結果がぼやけた
  - より鮮明な結果を得るために、別軸(スペシャリスト/世代/登録時期)でフィルタし、クラスタリング

③　SEに適切な研修を提案する（将来構想）

研修を受講するときのパターンは

1. 自ら手を上げる
2. 上司の推薦
3. SE育成室からの推薦

・プロジェクトマネージャーを目指す人のための研修
- SE育成室からPM目指している人にピンポイントで推薦したい
- 部署名から判断できるのはお客様業界やアプリ/インフラ開発など
- PMやコンサルを目指している人をどのように判断するか

研修を受講するときのパターンは

1. 自ら手を上げる

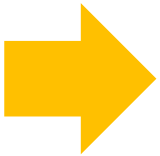
機械学習の「分類」で
SEの目指すキャリアを
分類できないか

ラ開発など

- PMやコンサルを目指している人をどのように判断するか

ロールモデルを元に「分類」する

- PMを複数人選定し、登録しているスキルを元にPMのモデルを作成する
- そのモデルにSEを照らしあわせ、PM向きの人を選出する
- PM向けの人に研修を提案する



- スパムフィルタリングと同様のメカニズム
- アーキテクトやコンサルタントの選出にも役立てることができる
- SEの効果的な育成を実現できるのではないか

おわりに

おわりに

- 技術管理システムの課題に対する取り組み
  - レコメンデーション/クラスタリング/分類

- 機械学習に取り組んでみて
  - データ分析の難しさ
  - 数学/アルゴリズム/手法の理解
  - 「仮説を立て、評価する」を何度も小さく繰り返すことが大事

- 機械学習ライブラリ Mahout について
  - 多次元データのクラスタリングでは、グループを特徴付けるものがぼやけてしまい、結果が曖昧になりがち
  - 試行錯誤の結果、納得感のあるクラスタリングを実現できた
  - その背景には、適用するアルゴリズムの調整を繰り返し実施した
  - アルゴリズムの豊富さ、利用するAPI変更の容易性を考慮するとMahoutは優秀

データ分析ビジネスのこれから

- IBMが開発した質問応答システム「ワトソン」
  - 米国の人気クイズ番組「ジョパディ」で歴代クイズ王と競演
  - この技術は医療分野などに応用
  - 機械学習は新しいビジネスを発展させていく

- データ分析ビジネスを共に盛り上げて参りましょう！

ご清聴ありがとうございました