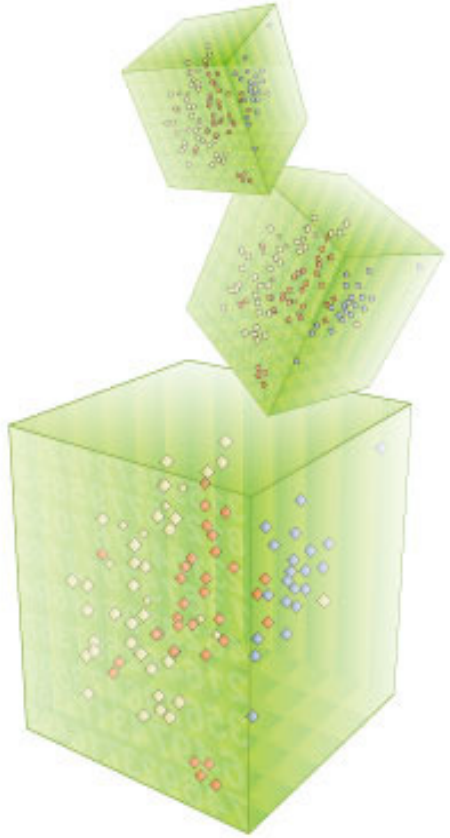
Predictive Analytics

Albert
Optimization technology

日本行動計量学会第41回大会（於：東邦大学）
特別セッション：ビッグデータ時代におけるマーケティング（1）

スパースな大量購買データで顧客クラスタリングを行なうための

顧客類似度計算方法の研究



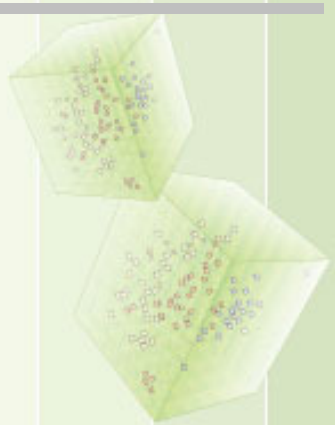
2013年9月4日（水）

山川義介・大堀あゆみ

株式会社ALBERT（アルベルト）
〒151-0053 東京都渋谷区代々木2-22-17
TEL：03-5333-3703（代表） FAX：03-5333-3723
www.albert2005.co.jp/

目次

1.はじめに



1.ビッグデータ時代の到来

大量購買データを用いた真のOne to one実現ニーズ

2.スパースなデータへの対応

従来の分析手法への疑問

3.クラスター分析の迷路問題

計算法や最適クラスターを決める基準がない

1-2.ビッグデータとは

明確な定義はない。その特徴については様々語られている。

3V＝Volume/Variety/Velocity
4V＝Volume/Variety/Velocity/Veracity
4V＝Volume/Variety/Velocity/Value

Volume（容量の大きさ）
ビッグデータの第一の特徴は、その名前の通り容量が大きいことです。企業に限らず、情報技術の進化により、黙っていてもどんどんデータが集まるようになり、データ量はテラバイトからペタバイトオーダーにもなっています。データ量が大きいことだけがビッグデータの特徴だと思われがちですが、他にも以下のようなポイントがあります。

Variety（多様性、種類）
ビッグデータは、通常表計算などで扱っているように、数値化され関連づけをされたデータ（構造化データ）であるとは限りません。テキスト、音声、画像、動画などのさまざまな構造化されていないデータ（非構造化データ）もあり、これらのデータをテキストマイニングや音声、画像解析などを行ない構造化し、ビジネスに活用する動きが広まっています。

Velocity（スピード、頻度）
サーバーのアクセスログや、東京ゲートブリッジ橋梁モニタリングシステムなど、ものすごい頻度、スピードでインターネット上やセンサーからデータが生成され、取得、蓄積されています。変化の著しい現代社会では、これらのデータをリアルタイムに処理し、対応することが求められています。

Veracity（正確さ）
従来は、サンプリングによって一部のデータで全体を推測する方法が主流でした。それに対し、ビッグデータは全てのデータを取得することも不可能ではないので、正確であり推測による曖昧さや不正確さなどを排除して、本当に信頼できるデータによる意思決定が可能になります。

Value（価値）
ビッグデータは、容量の大きさや多様性、スピードに価値があるのではありません。得られたデータを分析し有用な知識や知恵を導出し、モデル構築、検証し、課題解決をすることが本質的なビッグデータの価値です。

1-3.スパースでないデータとは

Predictive Analytics

アンケートデータのイメージ

	回答1	回答2	回答3	回答4	回答5	回答6	回答7	回答8	回答9	回答10	回答11	回答12	回答13	回答14	回答15	回答16	回答17	回答18	回答19	回答20	回答21	回答22	回答23	回答24	回答25	回答26	回答27	回答28	回答29	回答30	回答31	回答32	回答33	回答34	回答35	回答36	回答37	回答38	回答3
sample1	5	4	2	2	5	3	4	1	2	2	2	4	2	1	4	3	5	5	4	4	5	2	3	3	5	4	4	5	5	3	2	5	3	1	5	3	3	5	1
sample2	1	3	2	5	2	2	5	5	2	2	2	1	5	3	1	2	3	2	2	5	1	3	1	3	5	4	2	4	4	2	3	1	2	3	2	5	2	3	4
sample3	4	2	4	2	3	3	4	4	3	4	3	3	3	1	2	5	3	3	4	1	2	5	3	3	2	4	1	2	3	2	3	1	3	2	4	3	1	3	5
sample4	4	2	3	4	2	3	2	2	2	5	4	2	4	2	3	2	2	3	1	3	4	3	1	4	1	3	5	4	2	3	4	1	1	4	3	1	5	3	4
sample5	1	3	1	3	4	4	1	3	3	3	4	1	1	2	2	2	2	5	4	3	5	4	4	2	4	3	2	5	3	4	3	3	1	2	2	3	4	2	4
sample6	4	2	1	2	5	5	2	2	3	4	3	3	4	2	4	5	3	2	2	4	2	4	4	2	5	4	5	4	4	3	5	1	3	3	1	3	3	4	3
sample7	5	2	5	2	3	5	4	2	3	3	4	4	4	4	3	4	4	4	3	4	3	4	2	1	4	2	1	4	5	3	2	4	2	5	5	1	1	2	2
sample8	2	2	4	5	5	3	2	4	2	4	4	4	3	3	5	5	3	3	1	4	1	4	3	1	4	3	2	1	4	5	3	1	3	2	4	3	4	5	4
sample9	3	2	4	4	2	4	3	4	1	3	4	3	2	2	3	2	2	3	1	3	2	3	2	3	2	3	4	4	2	3	4	1	2	2	3	1	1	2	3
sample10	3	5	2	4	1	4	4	3	2	5	4	5	2	4	4	3	4	2	1	5	5	4	2	4	2	4	4	2	1	3	2	5	3	4	1	4	3	2	2
sample11	4	4	2	5	1	2	1	4	2	5	3	3	4	3	1	4	2	5	3	2	3	2	3	1	4	5	4	3	4	4	3	5	3	2	4	2	2	5	4
sample12	3	5	4	3	2	2	4	2	5	4	4	2	3	2	2	3	5	2	1	2	2	1	1	3	3	2	4	1	2	2	5	4	5	4	4	3	2	4	4
sample13	2	5	3	5	2	2	4	2	5	2	3	3	1	3	2	3	4	3	3	2	2	5	3	4	4	1	1	5	5	3	3	2	3	2	4	2	4	4	5
sample14	2	3	4	1	1	5	2	4	5	4	5	4	2	5	4	4	3	2	4	2	5	5	4	4	4	3	5	4	5	3	4	5	1	3	4	1	2	4	3
sample15	3	3	4	2	5	2	2	5	2	4	3	1	5	5	1	3	5	2	3	5	3	3	4	1	4	4	2	4	4	3	5	1	3	2	1	5	2	3	2
sample16	2	4	2	3	3	4	4	4	3	2	3	3	3	4	4	2	4	1	2	1	2	1	1	3	3	3	2	2	2	4	4	3	4	4	5	3	3	1	5
sample17	5	4	5	3	2	4	2	5	4	3	4	3	1	2	2	2	5	3	4	3	1	4	3	1	2	4	2	1	2	1	4	3	5	4	1	5	3	3	3
sample18	3	4	4	5	2	2	3	3	3	1	4	1	3	1	5	5	5	1	2	2	3	2	5	5	4	5	4	4	4	2	4	3	4	5	1	3	2	4	4
sample19	4	5	3	4	5	5	2	4	4	5	2	2	2	3	5	3	5	3	4	1	1	2	3	3	3	4	2	3	3	5	2	3	4	3	5	4	1	5	2
sample20	5	2	3	4	4	3	2	2	2	3	2	2	4	5	4	2	2	3	4	4	4	2	5	3	2	3	4	2	3	4	2	1	4	4	1	3	2	1	5
sample21	5	1	4	3	1	1	3	3	4	4	1	5	2	2	2	4	3	5	2	3	4	2	1	1	4	1	4	2	2	5	3	4	4	5	5	2	4	4	3
sample22	2	4	1	4	4	4	3	3	1	4	2	4	3	5	5	5	4	3	5	5	4	3	1	1	2	1	1	4	2	5	4	3	2	1	3	3	3	2	1
sample23	3	1	2	2	2	1	4	4	4	4	4	3	3	5	5	2	3	2	3	2	2	5	4	2	4	3	1	2	5	3	2	2	2	4	2	4	1	5	5

1-4.スパースなデータとは

Predictive Analytics

購買データのイメージ

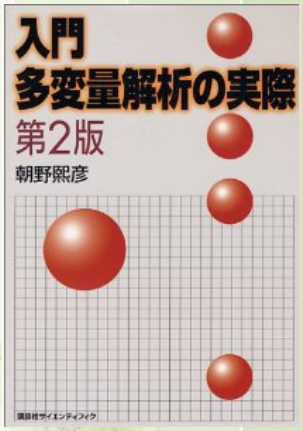
	商品1	商品2	商品3	商品4	商品5	商品6	商品7	商品8	商品9	商品10	商品11	商品12	商品13	商品14	商品15	商品16	商品17	商品18	商品19	商品20	商品21	商品22	商品23	商品24	商品25	商品26	商品27	商品28	商品29	商品30	商品31	商品32	商品33	商品34	商品35	商品36	商品37	商品38	商品39	商品40	商品41	商品42	商品43	商品44	商品45	商品46	商品47	商品48	商品49	商品**	商品**	商品**	商品**	商品**	商品**
sample1	2																																																							
sample2																												2																													
sample3																																																								
sample4																																																								
sample5																	1																																		6								
sample6																																																								
sample7																																																								
sample8																																																								
sample9							1																4																																				
sample10																																	1																									
sample11																																																								
sample12																																																								
sample13																																																								
sample14																																																								
sample15				8																																																				
sample16																1																																											
sample17																																																								
sample18																															2																									
sample19																																																								
sample20																																																								
sample21																																																								
sample22																																																								
sample23											3																																												
sample24																																																								
sample25																																																								
sample**																																																								
sample**																																																								
sample**																																			1																							
sample**																																																								
sample**																																																								
sample**																																																								
sample**																																																								

①計算法のバリエーションが多すぎる

→ともかくメジャーな解法に従う

②最適クラスターを決める基準がない

→クロス集計で決着をつける



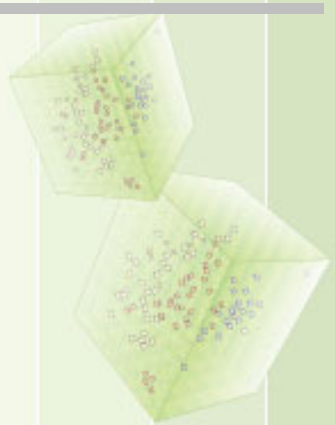
③どうやったらクラスターにアクセスできのかわからない

→クラスターとデモグラフィック変数との対応をつける

→ビッグデータは全員分析するのでアクセスできる

朝野熙彦（2000）「入門多変量解析の実際 第2版」講談社.

2.従来方法への疑問



2-1.類似度の指標の例

ピアソンの相関係数＝

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

コサイン類似度 ＝

$$\frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|}=\frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|}=\frac{\sum_{i=1}^{|V|}q_i d_i}{\sqrt{\sum_{i=1}^{|V|}q_i^2}\cdot\sqrt{\sum_{i=1}^{|V|}d_i^2}}$$

2-2.相関係数、COS距離の妥当性

caseA、caseB、caseCの類似度は同じなのだろうか？

caseA	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	
s1	1	0	0	0	0	0	0	0	0	0	1.000 cos距離
s2	1	0	0	0	0	0	0	0	0	0	1.000 相関係数

caseB	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	
s3	5	0	0	0	0	0	0	0	0	0	1.000 cos距離
s4	5	0	0	0	0	0	0	0	0	0	1.000 相関係数

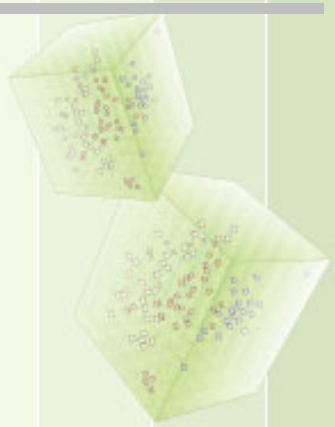
caseC	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	
s5	5	5	0	0	0	0	0	0	0	0	1.000 cos距離
s6	5	5	0	0	0	0	0	0	0	0	1.000 相関係数

caseD、caseEではcaseDの類似度のほうが高いのだろうか？

caseD	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	
s7	1	5	1	0	1	0	1	0	0	0	0.345 cos距離
s8	5	1	0	1	0	1	0	1	0	0	0.091 相関係数

caseE	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	
s9	1	10	0	0	0	0	0	0	0	0	0.198 cos距離
s10	10	1	0	0	0	0	0	0	0	0	0.089 相関係数

3.新しいクラスター分析手法



3-1.クラスター分析の目的

Predictive Analytics

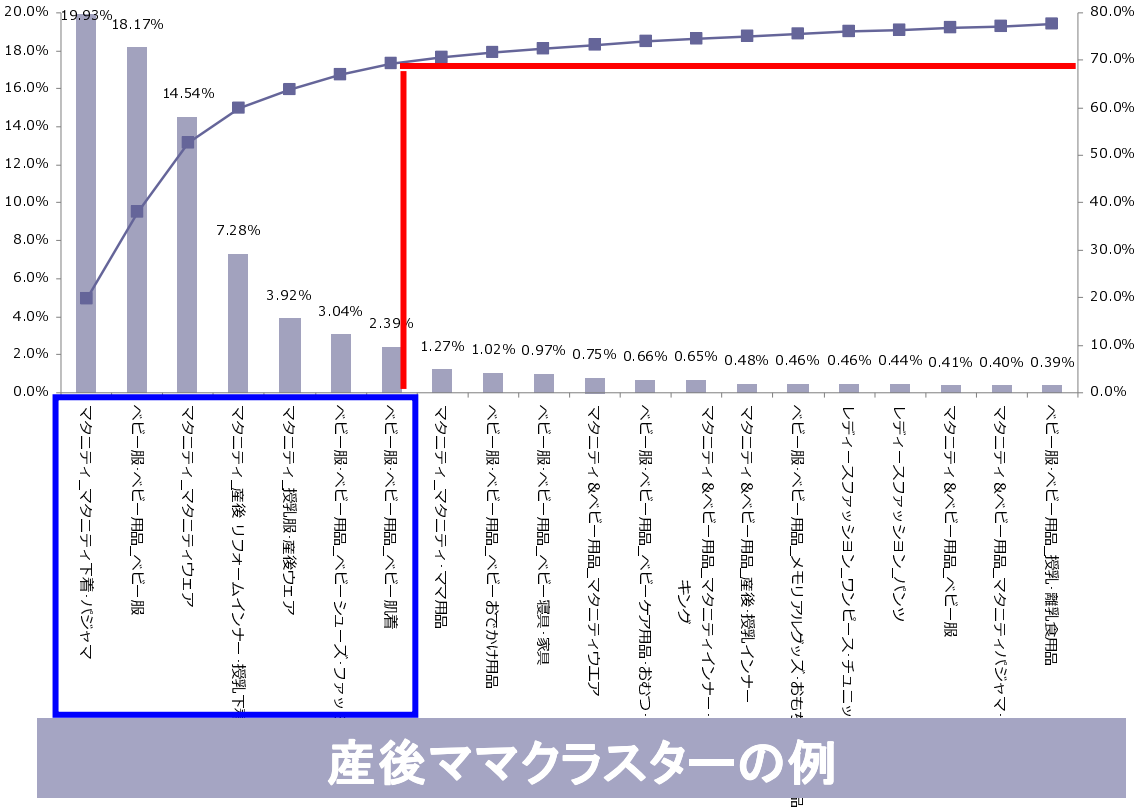
顧客に5つの商品をレコメンドすることを考えます。

極力その顧客が買いそうな商品（カテゴリ）を予測しお薦めしたい。

そのために、顧客の過去の購買履歴を元に、顧客をいくつかのクラスターに分けクラスター毎に買いそうなカテゴリの商品をお薦めする。

3-2.クラスタープロファイル例

以下は、各クラスターの購入カテゴリを降順に並べたパレート図です。例えばこの産後ママクラスターは、240カテゴリある中の、マタニティ_マタニティ下着・パジャマ、ベビー服・ベビー用品_ベビー服、マタニティ_マタニティウエアなどの7カテゴリだけで売上の70%以上を占めています。すなわち、このクラスターのユーザーは、全カテゴリをお薦めした時と比較すると、この7カテゴリの商品のお薦めで、70%以上の確率で購入が期待できることを意味しています。



産後ママクラスターの例

【購入するカテゴリ】

3-3.最適クラスターの定義

Predictive Analytics

よいクラスターとは、極力サンプル数の分散が小さく、各クラスターの累積70%到達カテゴリ数が総じて小さいものとする。一部のクラスターに人数が集中し、そのクラスターの70%到達カテゴリ数が大きければ、よいクラスターとはいわない。従って、70%到達カテゴリ数に各クラスターの人数比率を乗じた総和をクラスター数で割った値**α（重み付け平均70%到達カテゴリ数）**をクラスタのKPIとし、**α**が小さいほど、よいクラスターであると定義しました。

$$\alpha = \sum（70\%到達カテゴリ数 \times クラスター人数比率）$$

3-4.ALBERT類似度の考え方

Predictive Analytics

1.同じカテゴリを同数購入していても、
　購入数によって類似度に重み付けをする。

2.同じカテゴリを異なる数購入している場合、
　その差によって類似度に重み付けをする。

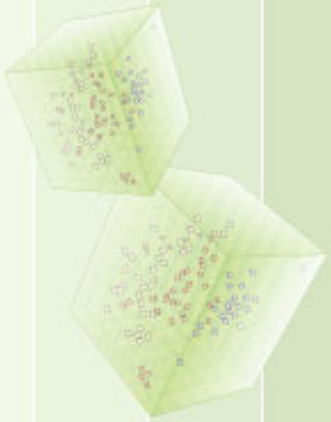
3.一方しか購入していないカテゴリに関して、
　非類似度の概念を導入する。

4.双方が購入していない場合は評価しない。

5.類似度、非類似度はその寄与度をチューニング
　できるようパラメータ化する。

3-5.ALBERT類似度計算式

Predictive Analytics

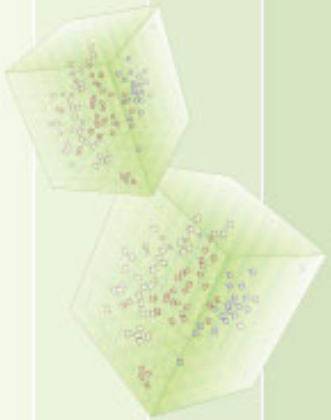


※特許出願準備中のため配布資料には提示していません。

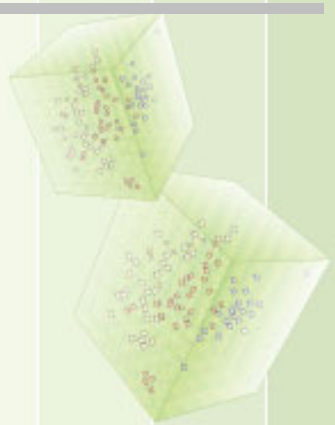
3-6.ALBERT類似度計算例

Predictive Analytics

※特許出願準備中のため配布資料には提示していません。



4.ALBERT類似度を用いた分析事例



4-1.検証に用いたデータ

Predictive Analytics

データの期間	2009/12/01～2012/10/04
用いたデータ	某通販会社購買履歴データ

【商品カテゴリ数】
97カテゴリ

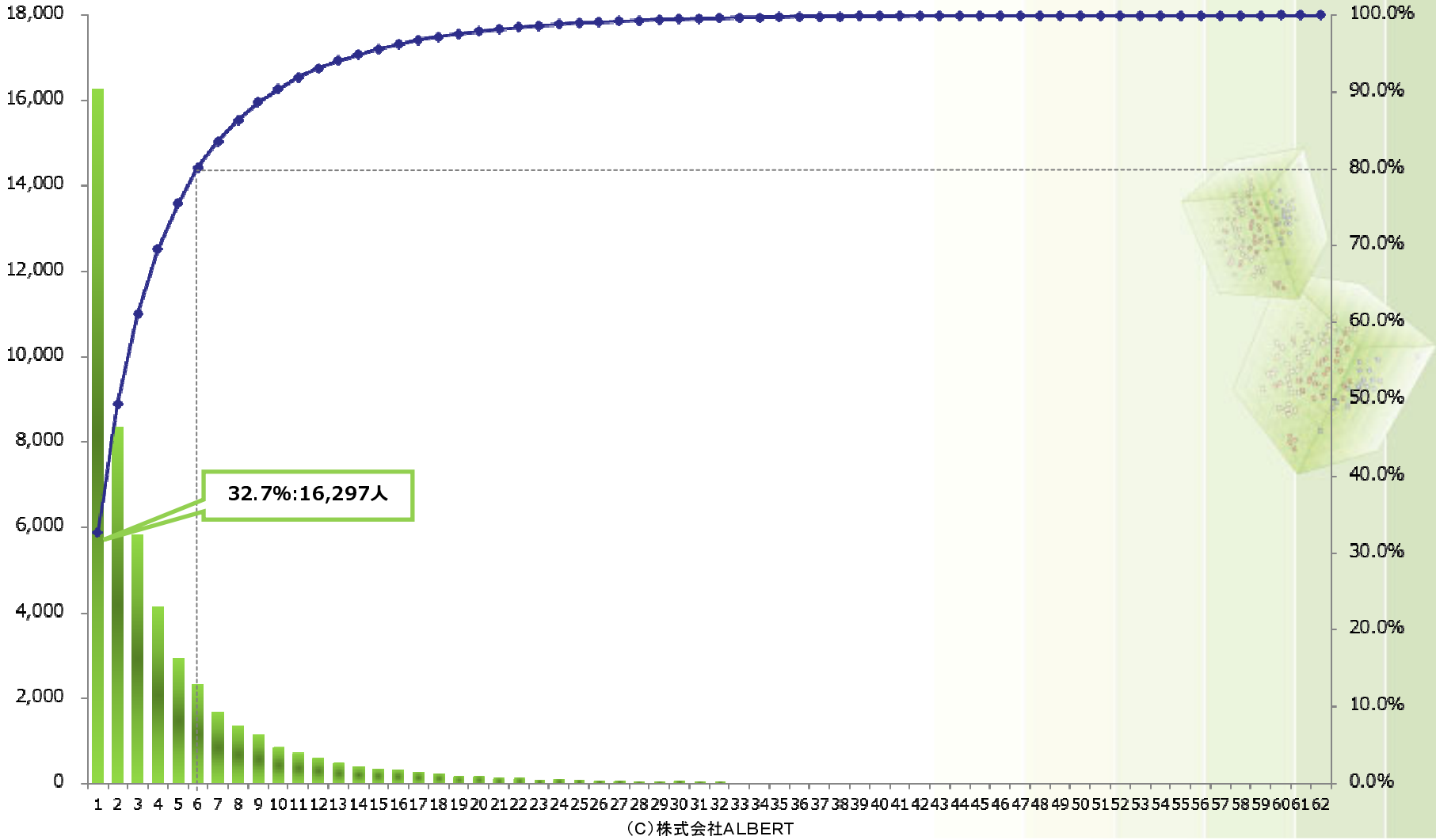
【検証するに使用するために】
用いた元データは498,889レコードありましたが、円滑に検証を行うため、
元データの10%である49,889レコードにランダムサンプリングを行い、
検証に用いました。

使用ツール:Visual Mining Studio(株式会社NTTデータ 数理システム)

4-2. 購入カテゴリ数のパレート図

Predictive Analytics

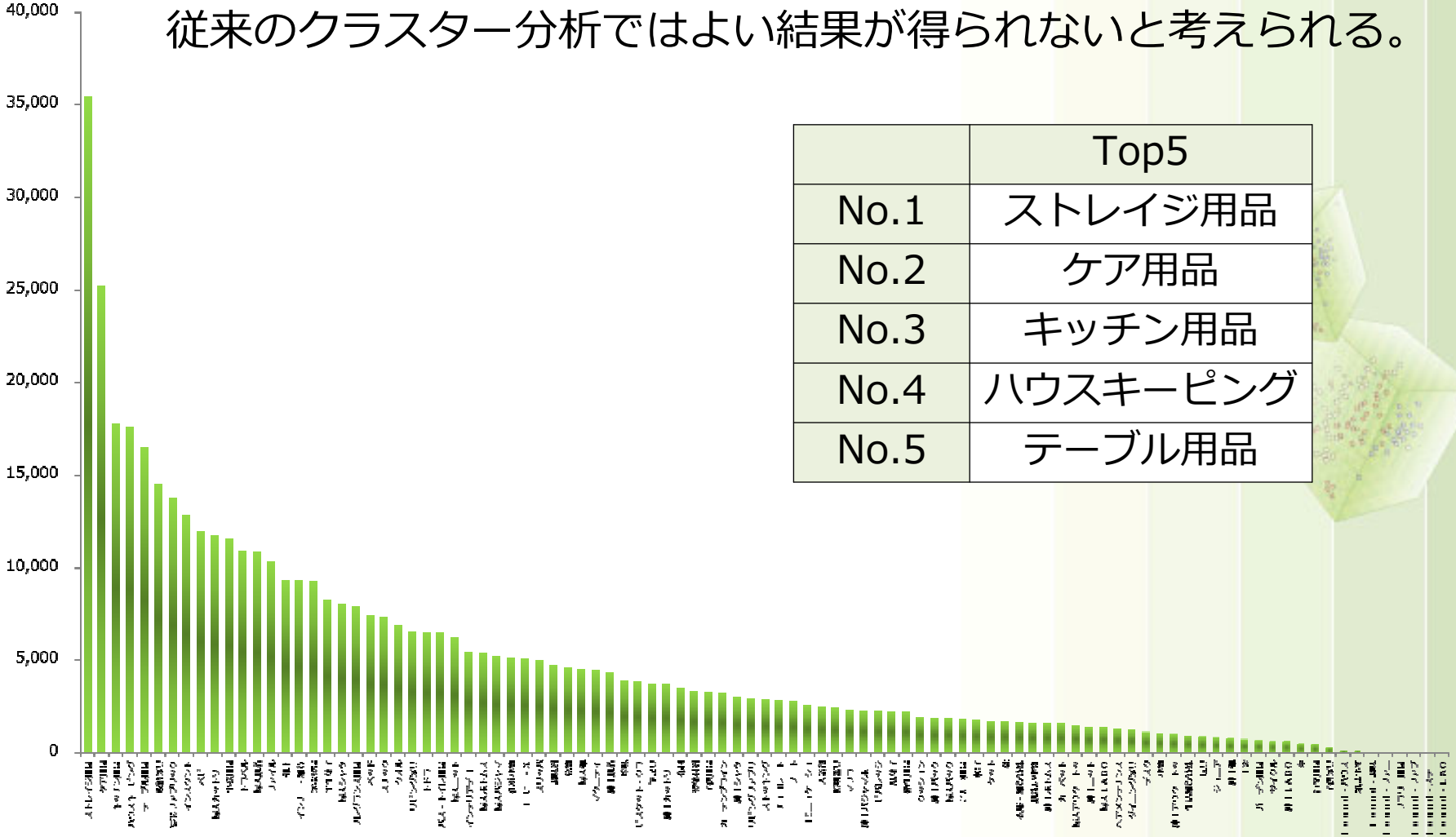
1/3の顧客が1カテゴリしか購入していない。
80%の顧客が6カテゴリ以下しか購入していない。



4-3.検証データ-カテゴリ別購入者数

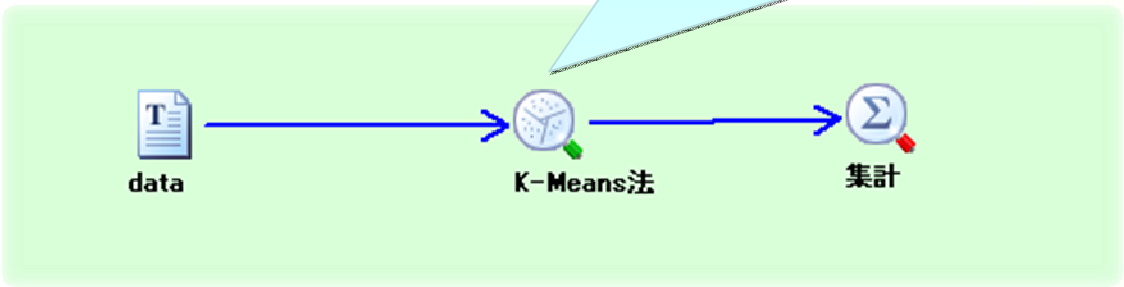
Predictive Analytics

売れるカテゴリと売れないカテゴリの差が極めて大きい。
この様に、購入カテゴリに大きな偏りがある場合、
従来のクラスター分析ではよい結果が得られないと考えられる。



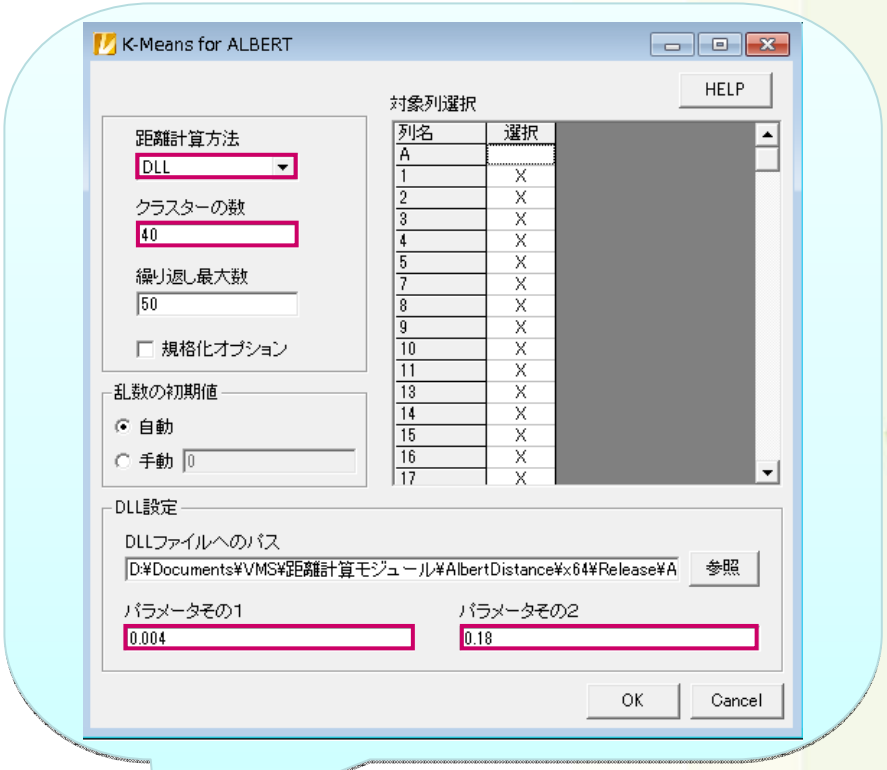
	Top5
No.1	ストレイジ用品
No.2	ケア用品
No.3	キッチン用品
No.4	ハウスキーピング
No.5	テーブル用品

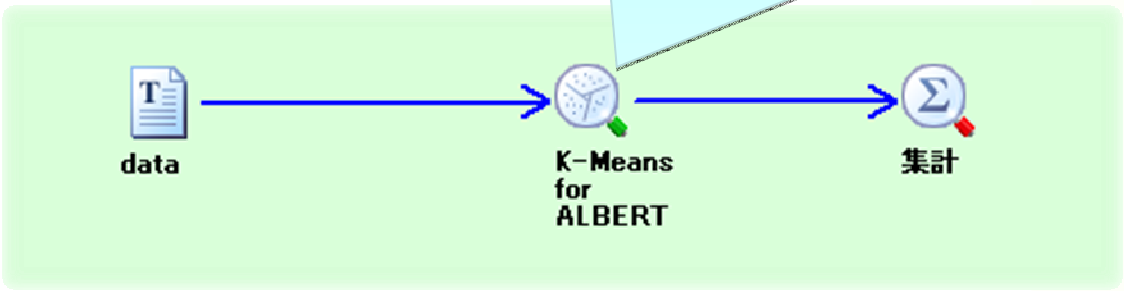
4-4.クラスター分析画面(K-means法cosine距離)



標準仕様

4-5.クラスター分析画面 (K-means法ALBERT類似度)



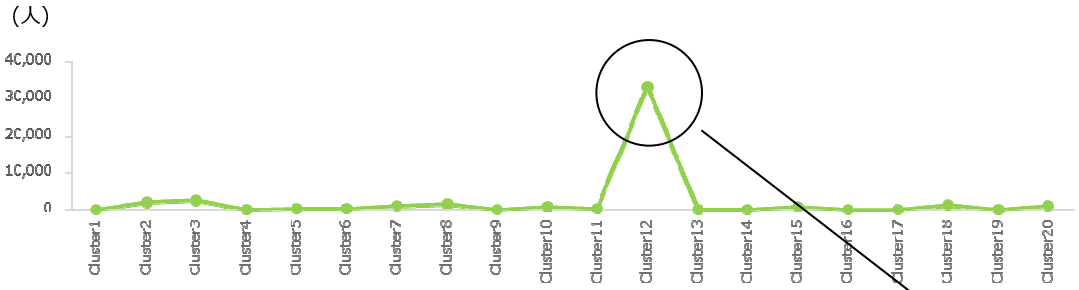


ALBERT仕様

4-6.ALBERT類似度を用いた2段階クラスター分析

Predictive Analytics

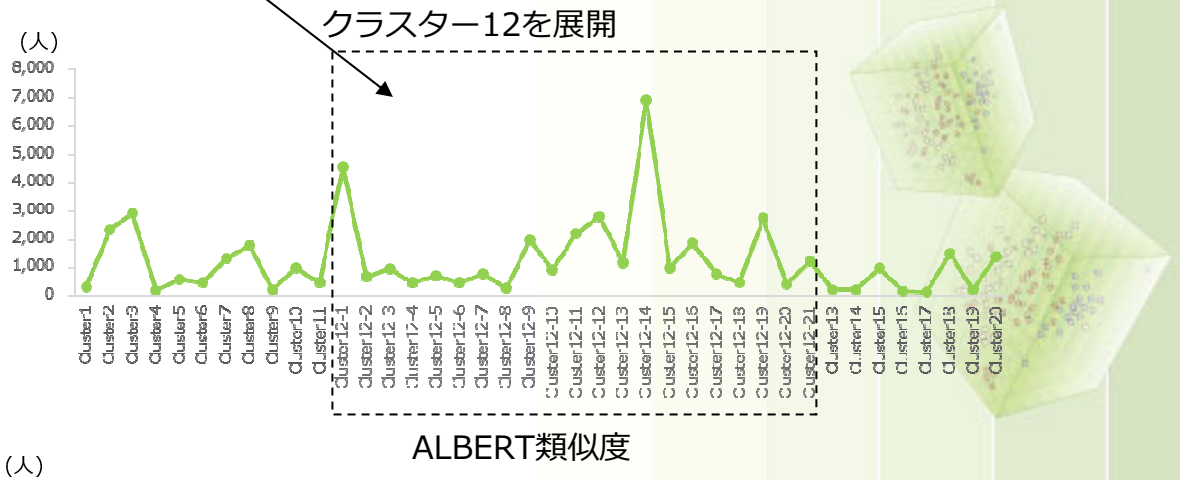
購入カテゴリ数にばらつきが大きいため、クラスター分析は2段階で行なった。



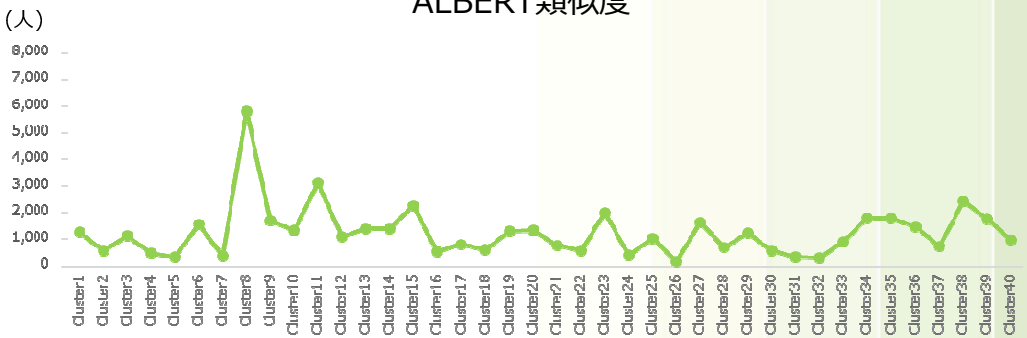
Step1：
70%到達カテゴリ数が最小になる類似度/非類似度パラメータを探索。ここでは購入カテゴリ数が大きい顧客がクラスタ12に偏在した。

Step2：
顧客数の多いクラスタの顧客のみ抽出し、さらに70%到達カテゴリ数が最小になる類似度/非類似度パラメータを探索。購入カテゴリ数が多い場合は、パラメータ設定をStep1と大幅に変更する必要がある。

クラスター12を展開



ALBERT類似度



（比較）cosine距離

4-7.ALBERT類似度を用いた70%到達度分析結果

Predictive Analytics

重み付け平均70%到達カテゴリ数の比較

	α
COSINE距離	14.21
ALBERT類似度	10.66

（25%の改善）

- 様々な購買データでの検証
- クラスター数の最適化
- 初期値問題
- 類似度/非類似度の最適解を求める方法の検討
- キャンペーンマネージメントシステムへの搭載
- 効果検証のためのPDCA
- 新規顧客対応

ご静聴ありがとうございました。

お問い合わせはこちら

y_yamakawa@mcx.co.jp

自己紹介

Predictive Analytics

◆山川義介

横浜国立大学工学部卒業。TDK株式会社記録メディア事業部門にて研究開発、商品企画に従事の後、株式会社マルマンに転じ常務取締役家電事業部長、マーケティング部長などを歴任。1995年株式会社エムアンドシーを設立し代表取締役に就任。2000年株式会社インタースコープ（マーケティングリサーチ＆コンサルティング）を設立し、取締役副社長に就任。2001年6月株式会社インタースコープ代表取締役社長に就任。2002年「EOY JAPANセミファイナリスト（スタートアップ部門）」。2005年7月インタースコープ取締役会長に就任。2005年7月株式会社ALBERTを設立し、代表取締役会長に就任。2007年2月インタースコープをヤフー株式会社に売却。（後にインフォプラントと合併、ヤフーバリューインサイトと社名変更、2010年マクロミルと経営統合）。2007年4月より関東学院大学人間環境研究所客員研究員。2008年9月より明治大学大学院グローバル・ビジネ ス研究科（MBA）非常勤講師[CRM（データマイニング）]。japan.internet.comのWebマーケティングコラム「One to oneマーケティングの本質を探る」連載。

◆大堀あゆみ

2013年3月多摩大学経営情報学部マネジメントデザイン学科卒業、豊田裕貴ゼミ。同年4月株式会社ALBERT入社、データ分析部に配属。

会社概要

Predictive Analytics

会社概要

社名	株式会社ALBERT
設立	2005年7月1日
資本金	3億3,900万円
株主	デジタル・アドバタイジング・コンソーシアム株式会社、 IVP Incubator, L.P、オリックス・キャピタル株式会社、 株式会社ジャフコ、バーチャレクス・コンサルティング株式会社、 三生キャピタル株式会社、東洋キャピタル株式会社、 ニュー・フロンティア・パートナーズ株式会社、 SMBCベンチャーキャピタル株式会社、信金キャピタル株式会社、 PE&HR株式会社、大和企業投資株式会社、 株式会社シーエー・モバイル、役員および従業員
役員	代表取締役会長　山川 義介 代表取締役社長　上村 崇 取締役　徳久 昭彦（DAC取締役CTO） 執行役員　安達章浩 木野英明 佐藤めぐみ 平原昭次 監査役　谷本 篤彦 非常勤監査役　保月 英機
事業内容	CRMソリューションの開発・提供 ・分析・コンサルティング ・データマイニング・ソフトウェア ・マルチチャネルOne to oneマーケティングソリューション 行動ターゲティング広告システムの開発・提供 ・レコメンド特化型DSP ・広告配信の最適化 ・広告クリエイティブの最適化 レコメンドエンジンの開発・提供 ・Webレコメンドエンジン ・モバイルレコメンドエンジン ・感性検索システム

事業概要

2005年7月設立。事業コンセプトは『分析力をコアとするマーケティングソリューションカンパニー』。高度なCRMソリューションをカジュアルに提供するテクノロジーとして、統計解析、データマイニング、テキスト解析、マーケティングリサーチに加え、画像解析、豊富な導入実績に裏付けられた信頼のWeb、モバイル、ITインフラ技術を保有。
これらのキーテクノロジーをベースに独自開発のビッグデータ対応データマイニング・ソフトウェア『smarticA!データマイニングエンジン』や、クラウド型キャンペーンマネージメントシステム『smarticA!キャンペーンマネージメント』、行動履歴を使ったレコメンドエンジン『おまかせ！ログレコメンダー』、対話型意思決定システム『Bull's eye』、さらに、レコメンド特化型DSP『ADreco』、広告配信の最適化を実現する『i-Effect』など、分析力を強みとしたマーケティングソリューションを提供しています。