

MOLAP を用いたビジネスインテリジェンスのための 統計誤差を考慮した軽量キューブ生成

蔵内雄貴¹ 松田治¹ 瀬下仁志¹

概要：本稿では、データ分析に影響が少ない範囲で MOLAP キューブのサイズを小さくし、巨大なデータが対象であったとしてもローカルでの MOLAP による分析を可能とすることを目的とする。そのために、標本誤差の影響が大きくなる集計値については集計を行わないことで、データ分析への影響が少ない範囲で集計パターンを大幅に減らし、現実的に分析可能な範囲の MOLAP キューブを作成する手法を提案した。実データを用いた実験では、許容する標本誤差、および分析対象とするデータの量を変更した場合に、MOLAP キューブのサイズがどう変化するかを確かめた。特に、許容する標本誤差を 0.1 とした場合、元の 5.7% に圧縮可能であることを示し、提案法の有効性が示された。

キーワード：MOLAP, キューブ, 軽量化

Small cube building considering statistical errors for business intelligence using MOLAP

YUKI KURAUCHI^{†1} OSAMU MATSUDA^{†1} HITOSHI SESHIMO^{†1}

1. はじめに

データ分析を意思決定に役立てるビジネスインテリジェンスが広く用いられている。データ分析の際、近年では機械学習や統計を用いたデータ分析も広く行われているが、人によるデータ分析についても広く行われており、用途によって使い分けられている。人によるデータ分析では、様々なデータ分析ツールを通じてドリルダウン、スライシング、ダイニングなどの操作によって頻りに視点を変えながらのデータ分析が行われる。この概念は Online Analytical Processing (OLAP) と呼ばれている[9]。本研究では、このような人によるデータ分析に着目する。

このとき、特に大規模な企業などにおいては、データの量が多いため、また、例えば千単位などの営業所から並列に分析が行われるために、2つの問題が発生する。それは、分析用サーバに対する負荷が大きいという問題、また、視点を変える操作の際の再集計に時間がかかるため、分析において頻りに中断が発生し、分析の効率が下がるという問題である。

OLAP の実現方法は大きく分けて 2 つ、関係データベースを用いた ROLAP、多次元データベースを用いた MOLAP があり、現在も研究が続けられている[8]。ROLAP は集計のたびに元データにアクセスするため、巨大なデータを対象とする場合、分析用サーバに対する負荷が大きく、操作の際のレイテンシは大きい。一方、MOLAP は全てのクロス集計を事前に行いキューブと呼ばれる多次元配列として

保持する。このため、事前に全ての組合せについて集計を行う必要があり、集計に膨大な時間がかかる特徴を持つ。しかし、いったんキューブの構築ができれば、以降の操作の際のレイテンシが小さいという特徴があり、ROLAP の課題を解決することができる。また、MOLAP キューブをローカルで扱うことができるサイズまで小さくすることができれば、キューブの操作はローカルに任せることで、分析用サーバに対する負荷を低減することができる。

そこで本稿では、データ分析に影響が少ない範囲で MOLAP キューブのサイズを小さくし、巨大なデータが対象であったとしてもローカルでの MOLAP による分析を可能とすることを目的とする。このとき、我々が着目したのは、分析を進めてデータを細分化していくと、細分化した値における標本誤差の影響が大きくなり、分析の信憑性が損なわれることである。これをもとに、標本誤差の影響が大きくなる集計値については集計を行わないことで、集計パターンを大幅に減らし、現実的に分析可能な範囲の MOLAP キューブを作成する手法を提案する。

実データを用いた実験では、許容する標本誤差、および分析対象とするデータの量を変更した場合に、MOLAP キューブのサイズがどう変化するかを確かめた。特に、許容する標本誤差を 0.1 とした場合、元の 5.7% に圧縮可能であることを示し、提案法の有効性が示された。

以降、2 章にて関連研究、3 章にて提案法、4 章にて実験概要、5 章にて結果と考察、6 章にて結論と今後の展望を述べる。

2. 関連研究

OLAP の概念は Codd らによって提唱され[9]、キューブ

¹ NTT サービスエボリューション研究所
NTT Service Evolution Laboratories

の事前計算による操作の際のレイテンシ低減[2], 時間における年, 時, 分などの依存関係がある属性のラティスによる解決[3]などが続いて研究された。

操作の際のレイテンシ低減は依然として課題であり, 現在でも ROLAP と MOLAP それぞれの長所を組み合わせる手法[8]が検討されているほか, top-k の探索などタスクを限定した高速化[6]などが継続して検討されている。また, 本稿では取り扱わないもう 1 つの課題である MOLAP キューブの生成高速化についても, 並列での生成などの研究[1]が行われている。これらの研究はいずれも提案法との併用が可能である。

一方, 分析を補助する観点では, ドリルダウンの際に局所的な傾向を全体の傾向と誤らないようにするツール[4]の研究がある。また, 提案法と似た手法として, トレンドの変化をとらえるために, 閾値を用いてキューブを縮小して表現し, 時系列比較する研究[5]があるが, いずれも目的が異なり, 分析可能な軽量キューブの生成に用いることはできない。

近年の分析ツールとしては, 演算速度が上がったために, Tableau[12]のようにデータベースの変更履歴を全て保持する大福帳形式を扱うものも出ているが, これは小規模なデータのみを対象としている。また, 2015 年に公開されたオープンソースの分散多次元分析エンジン Apache kylin [7]では, MOLAP キューブを軽量化するための機能が数多く含まれており, 軽量の MOLAP キューブの必要性の高さを裏付けていると言える。軽量化の機能の例としては, 以下の 4 つがある。

- **partial cubing:** 組み合わせて分析することがない属性は別にキューブ化する
 - **mandatory dimension:** 指定した属性を含む組合せだけをキューブ化する
 - **hierarchy dimension:** 国, 県, 市など階層のある属性で, 下位の属性を含む場合は上位の属性を含むキューブを生成しない
 - **derived dimension:** 外部キーに紐づく情報がある場合に外部キーのみをキューブ化し, クエリ時に紐づける
- そのほか, Google は列指向 DB で分散処理をする BigQuery を提供しているが[10], 社外に出せないデータの場合に利用が難しく, 自社で同等の環境を構築するのはコストが高い問題がある。

3. 提案法

本研究の目的は, データ分析に影響が少ない範囲で MOLAP キューブのサイズを小さくし, 巨大なデータが対象であったとしてもローカルでの MOLAP による分析を可能とすることである。

このとき, 我々が着目したのは, 分析を進めてデータを細分化していくと, 細分化した値における標本誤差の影響

が大きくなり, 分析の信憑性が損なわれることである。例えば, 通信業に関するデータ分析の例では, ある地域 A においてフィーチャフォンでパケットを多く利用し, かつ長期契約者の 60 代男性のうち, スマートフォンに変更した比率の推移を時系列で見る, などといった分析が行われうる。これを整理すると, 下記 8 属性項目のクロス分析であると言える。

- 前端末種別: フィーチャフォン
- 後端末種別: スマートフォン
- パケット利用量: ○○パケット以上
- 契約期間: ○○年以上
- 居住地域: 地域 A
- 年代: 60 代
- 性別: 男性
- 期間: ○年○月～○年○月

このように細分化した分析を行う場合, たとえサンプル数が多いデータであったとしても, 細分化された値は数件～数十件の単位まで減ってしまうことがある。その場合, 標本誤差の影響を無視することができなくなり, 分析の信憑性が損なわれてしまう。また, 人がデータ分析を行うという観点でも, 細分化していくと分析すべきパターンが増えすぎるため, 解釈が難しくなってしまうという問題もある。先ほど例示した通信業に関するデータ分析で言えば, 他の性年代ではどうか, 地域ごとに傾向が無いかなどを同時に見ようとしても難しいという問題である。いずれにせよ, データ分析を行う上では, 細分化に限界があると言える。

そこで, 標本誤差の影響が大きくなる集計値について集計しないことで, 集計パターンを大幅に減らし, 現実的に分析可能な範囲の MOLAP キューブを作成する手法を提案する。前述した通りデータ分析においては細分化に限界があることから, データ分析作業への影響も小さいものと想定できる。

標本誤差の影響が大きいかどうかを判定する方法としては, 調査を設計する際などに利用される, 調査対象数の決定手法を用いる。具体的には, 標本誤差分析の目的が母比率を得ることであるとすれば, 標本誤差を δ 以内とするために必要な母集団の標本数 n は, 下式で求められる。

$$n \geq \left(\frac{z_x}{\delta}\right)^2 P(1-P)$$

このとき, P は母比率, δ は標本誤差, z_x は標準正規分布の上側 100x%点である。必要な母集団の標本数 n は母比率 P , すなわち分析する属性値の比率が 50%の場合に最も大きくなる。

例えば, 標本誤差 δ を 0.1 以下, つまり $\pm 10\%$ 以下の誤差で結果を得たい場合, 母比率 P を 50%として必要な母集団

表1 求める標本誤差に応じた必要サンプル数

標本誤差 δ	必要サンプル数
0.1	96.04
0.2	24.01
0.3	10.67
0.4	6.00
0.5	3.84
0.6	2.67
0.7	1.96
0.8	1.50
0.9	1.19

の標本数 n が最も大きい場合を求めると、 z_α は信頼係数 95% の場合は 1.96 となるので、

$$n \geq \left(\frac{1.96}{0.1}\right)^2 0.5(1 - 0.5) = 96.04$$

となる。すなわち、母集団の標本数が 96 以下となる場合は集計表を作る必要は無い。先に出したデータ分析の例で言えば、ある地域においてフィーチャフォンでパケットを多く利用する長期契約者の 60 代男性が 96 サンプル以下であれば、それより細かく集計する必要は無い。

誤差率を変更した場合に必要なサンプル数は表 1 の通りである。許容する標本誤差を大きくするほど、必要なサンプル数が減少し、より多くのデータを分析対象に含めることができるようになるが、MOLAP キューブのサイズは大きくなってしまいう傾向にある。

以上の方法により、分析したいデータと、母集団の標本数に対して誤差が±何%以下であってほしいかが入力されれば、分析可能な範囲を機械的に判定し、その範囲だけの MOLAP キューブを構築できる。範囲外の値については、サンプル数が何件以下であることを示すか、または必要であれば元データにアクセスして集計し実際の値を集計することが考えられる。

4. 実験概要

提案法について、有効性を確かめるために実データを用いて 2 つの実験を行った。実験 1 では、許容する標本誤差を変更した場合に、どの程度 MOLAP キューブのサイズを小さくできるかを確かめ、実験 2 では、分析対象とするデータの量を変更した際に、MOLAP キューブのサイズがどう変化するかを確かめた。

実験で用いるデータは、集計されていないローデータである必要があるが、このようなデータのうちオープンであるものに一般用マイクロデータ [11] がある。これは、集計表から作成するなどの調査票情報を直接的に用いない方法により作成する擬似的なローデータであり、秘匿を気にする必要なく自由に扱うことができる。本稿ではこのうち、就業構造基本調査の 1997, 2002, 2007, 2012 年の計 4 年分を実験に用いた。全数は 873,888 件であり、属性項目の数は

表2 用いたデータの属性概要

属性項目	属性値の数	属性値の例
調査年	4	1997, 2002
都道府県	47	北海道, 青森
政令指定都市	22	札幌, 仙台
市部	2	市部, 市部以外
性別	2	男, 女
年齢	13	15~75 歳で 5 歳刻み
就業状態	2	有業者, 無業者
雇用者	2	雇用者, それ以外
正規就業者	5	正規, 非正規就業者
産業	36	農業・林業, 漁業
就業希望	3	就業希望者, 非就業希望者
求職	3	求職者, 非求職者
非就業者年齢	16	15~85 歳で 5 歳刻み
配偶関係	5	未婚, 配偶者あり

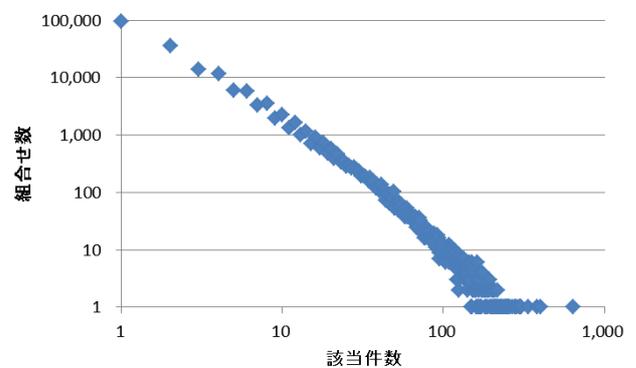


図1 属性値の組合せごとに該当する件数のヒストグラム

14, 各属性項目における属性値の種類数および属性値の例は表 2 の通りである。想定している巨大なデータよりも小さなデータセットではあるが、誤差率やデータ量を変更した場合の MOLAP キューブの変化傾向については検証可能と考える。

このようなデータを用いた場合の分析において、扱うデータ量がどの程度になるかを以降で述べる。まず、全ての属性値の組合せパターン数は、最大で下式の通りとなる。

$$\prod_{a \in A} n_a$$

ここで、 a は属性項目、 A は全ての属性項目集合、 n_a は属性項目 a における属性値の種類数を表す。すなわち属性値の種類数を全ての属性項目について掛け合わせたものであり、実験で用いたデータにおいては 111,493,324,800 通りとなる。このうち実際に 1 件以上のローデータが該当し、値を含む組合せは 197,796 通り、最大パターン数の 0.00002% であった。その属性値の組合せパターンごとに該当する件数のヒストグラムは図 1 の通りであり、ロングテールの傾向が見てとれる。このとき、例えば標本誤差を 0.1 とした場合、

表4 許容する標本誤差を変化させた場合のクロス数ごとの MOLAP キューブのサイズと、値を含む組合せに対する比率

クロス数	許容する標本誤差									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	(参考)全数
1	161	162	162	162	162	162	162	162	162	162
2	6,966	7,851	8,087	8,140	8,211	8,234	8,259	8,259	8,259	8,351
3	84,197	115,659	129,340	134,958	140,776	143,475	146,258	146,258	146,258	154,272
4	470,953	777,579	948,148	1,031,068	1,119,182	1,161,917	1,207,118	1,207,118	1,207,118	1,355,955
5	1,511,014	2,949,722	3,914,649	4,437,013	5,019,680	5,317,054	5,639,415	5,639,415	5,639,415	6,830,262
6	3,103,428	7,077,299	10,178,372	12,010,676	14,163,889	15,321,235	16,603,724	16,603,724	16,603,724	21,875,993
7	4,329,972	11,458,295	17,780,138	21,809,873	26,790,097	29,602,975	32,779,890	32,779,890	32,779,890	47,287,048
8	4,224,387	12,963,575	21,633,262	27,546,454	35,209,828	39,745,008	44,951,256	44,951,256	44,951,256	71,329,644
9	2,903,564	10,390,778	18,617,029	24,585,850	32,675,067	37,678,568	43,502,942	43,502,942	43,502,942	76,206,058
10	1,390,959	5,874,205	11,302,460	15,475,595	21,378,788	25,188,314	29,674,391	29,674,391	29,674,391	57,571,349
11	449,865	2,285,703	4,734,217	6,722,502	9,656,912	11,630,923	13,975,700	13,975,700	13,975,700	30,135,741
12	92,037	580,261	1,300,534	1,916,443	2,865,740	3,531,897	4,326,860	4,326,860	4,326,860	10,412,886
13	10,407	86,001	210,520	322,198	502,325	634,548	792,216	792,216	792,216	2,138,423
14	457	5,593	15,196	24,168	39,359	51,082	64,952	64,952	64,952	197,796
合計	18,578,367	54,572,683	90,772,114	116,025,100	149,570,016	170,015,392	193,673,143	193,673,143	193,673,143	325,503,940

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	(参考)全数
1	99.4%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
2	83.4%	94.0%	96.8%	97.5%	98.3%	98.6%	98.9%	98.9%	98.9%	100.0%
3	54.6%	75.0%	83.8%	87.5%	91.3%	93.0%	94.8%	94.8%	94.8%	100.0%
4	34.7%	57.3%	69.9%	76.0%	82.5%	85.7%	89.0%	89.0%	89.0%	100.0%
5	22.1%	43.2%	57.3%	65.0%	73.5%	77.8%	82.6%	82.6%	82.6%	100.0%
6	14.2%	32.4%	46.5%	54.9%	64.7%	70.0%	75.9%	75.9%	75.9%	100.0%
7	9.2%	24.2%	37.6%	46.1%	56.7%	62.6%	69.3%	69.3%	69.3%	100.0%
8	5.9%	18.2%	30.3%	38.6%	49.4%	55.7%	63.0%	63.0%	63.0%	100.0%
9	3.8%	13.6%	24.4%	32.3%	42.9%	49.4%	57.1%	57.1%	57.1%	100.0%
10	2.4%	10.2%	19.6%	26.9%	37.1%	43.8%	51.5%	51.5%	51.5%	100.0%
11	1.5%	7.6%	15.7%	22.3%	32.0%	38.6%	46.4%	46.4%	46.4%	100.0%
12	0.9%	5.6%	12.5%	18.4%	27.5%	33.9%	41.6%	41.6%	41.6%	100.0%
13	0.5%	4.0%	9.8%	15.1%	23.5%	29.7%	37.0%	37.0%	37.0%	100.0%
14	0.2%	2.8%	7.7%	12.2%	19.9%	25.8%	32.8%	32.8%	32.8%	100.0%
合計	5.7%	16.8%	27.9%	35.6%	46.0%	52.2%	59.5%	59.5%	59.5%	100.0%

表3 クロス数ごとの値を含む組合せ数

クロス数	属性値の組合せ数	値を含む組合せ数
1	14	162
2	91	8,351
3	364	154,272
4	1,001	1,355,955
5	2,002	6,830,262
6	3,003	21,875,993
7	3,432	47,287,048
8	3,003	71,329,644
9	2,002	76,206,058
10	1,001	57,571,349
11	364	30,135,741
12	91	10,412,886
13	14	2,138,423
14	1	197,796
合計	16,383	325,503,940

横軸である該当件数が 97 の位置に縦線を引き、それよりも左にプロットされている組合せについて集計しないと解釈することができる。

次に、MOLAP キューブのサイズ、すなわち集計される数値の数は、最大で下式の通りとなる。

$$\prod_{a \in A} (n_a + 1)$$

すなわち、各属性値の種類に加えてその属性値について絞り込みをしない場合の組合せの数となるため、属性値の種類数に 1 を足した数の掛け合わせとなり、実験で用いたデ

ータにおいては 1,776,527,230,464 通りとなる。このうち実際に値を含む組合せは 325,503,940 通りであり、最大パターン数の 0.02% であった。全ての属性値の組合せパターン数と比べて、理論値では 10 倍以上、値を含む組合せでは 1,000 倍以上大きいことが見てとれる。本稿における課題は、この値を含む組合せをいかに小さくできるかである。

これをクロス数ごとに確認すると、表 3 の通りである。14 クロス、すなわちすべての属性値の組合せにおいて値を含むパターン数は前述の 197,796 通りである。しかし、例えばクロス数が 2 の場合、14 個の属性項目のうち、どの属性項目が選ばれるかによって ${}_{14}C_2=91$ の組合せが生じる。そのため、値を含む組合せの数が大きくなる傾向があり、最も大きいのはクロス数が 9 のときで、76,206,058 通りとなっている。

5. 結果と考察

本章では、前章において述べた実験の結果および考察について述べる。

5.1 実験 1: 許容する標本誤差を変えた場合

許容する標本誤差を 0.1 から 0.9 まで変えた場合の、クロス数ごとの MOLAP キューブのサイズと値を含む組合せに対する比率は表 4 の通りである。参考として、最右列に、値を含む組合せの全数を記載している。

まず、合計の行を確認すると、許容する標本誤差を小さくするほどに圧縮率が高くなる傾向が確認できた。許容する標本誤差を 0.1、すなわち ±10% 以下の誤差で結果を得た

表5 データ量を変えた場合のクロス数ごとの MOLAP キューブのサイズと、各年数分の値を含む組合せに対する比率

クロス数	許容する標準誤差							
	0.1				0.3			
	利用する年数				利用する年数			
	①1年分	②2年分	③3年分	④4年分	①1年分	②2年分	③3年分	④4年分
1	140	150	156	161	141	153	157	162
2	4,533	5,767	6,390	6,966	6,114	7,173	7,601	8,087
3	40,920	61,121	73,104	84,197	79,331	104,284	116,240	129,340
4	183,230	305,843	391,543	470,953	479,455	697,865	821,952	948,148
5	497,757	900,084	1,215,109	1,511,014	1,681,566	2,664,172	3,300,582	3,914,649
6	902,056	1,728,274	2,432,131	3,103,428	3,827,413	6,488,347	8,395,246	10,178,372
7	1,143,396	2,283,662	3,322,725	4,329,972	6,001,435	10,727,452	14,399,538	17,780,138
8	1,034,402	2,126,800	3,181,848	4,224,387	6,680,666	12,442,246	17,232,633	21,633,262
9	668,820	1,400,675	2,147,688	2,903,564	5,331,898	10,247,617	14,585,564	18,617,029
10	304,273	642,926	1,009,320	1,390,959	3,028,869	5,960,299	8,693,201	11,302,460
11	93,905	198,433	319,524	449,865	1,192,786	2,387,711	3,562,080	4,734,217
12	18,330	38,384	63,812	92,037	308,231	624,461	952,066	1,300,534
13	1,970	4,036	7,037	10,407	46,706	95,446	148,810	210,520
14	83	161	305	457	3,114	6,416	10,267	15,196
合計	4,893,815	9,696,316	14,170,692	18,578,367	28,667,725	52,453,642	72,225,937	90,772,114

	①1年分	②2年分	③3年分	④4年分	(参考)④-①	①1年分	②2年分	③3年分	④4年分	(参考)④-①
1	99.3%	98.0%	99.4%	99.4%	+0.1%	100.0%	100.0%	100.0%	100.0%	+0.0%
2	71.5%	77.3%	81.2%	83.4%	+11.9%	96.5%	96.2%	96.6%	96.8%	+0.4%
3	39.5%	46.6%	51.5%	54.6%	+15.1%	76.5%	79.6%	81.9%	83.8%	+7.3%
4	22.5%	27.7%	31.9%	34.7%	+12.3%	58.8%	63.2%	66.9%	69.9%	+11.1%
5	13.4%	16.8%	19.8%	22.1%	+8.7%	45.2%	49.6%	53.9%	57.3%	+12.1%
6	8.2%	10.4%	12.5%	14.2%	+5.9%	35.0%	38.9%	43.1%	46.5%	+11.5%
7	5.2%	6.5%	7.9%	9.2%	+3.9%	27.3%	30.5%	34.3%	37.6%	+10.3%
8	3.3%	4.1%	5.0%	5.9%	+2.6%	21.5%	23.9%	27.3%	30.3%	+8.8%
9	2.1%	2.6%	3.2%	3.8%	+1.7%	16.9%	18.8%	21.6%	24.4%	+7.5%
10	1.3%	1.6%	2.0%	2.4%	+1.1%	13.3%	14.7%	17.1%	19.6%	+6.3%
11	0.8%	0.9%	1.2%	1.5%	+0.7%	10.4%	11.4%	13.4%	15.7%	+5.3%
12	0.5%	0.5%	0.7%	0.9%	+0.4%	8.0%	8.7%	10.3%	12.5%	+4.5%
13	0.3%	0.3%	0.4%	0.5%	+0.2%	6.1%	6.6%	7.9%	9.8%	+3.8%
14	0.1%	0.1%	0.2%	0.2%	+0.1%	4.5%	4.8%	5.9%	7.7%	+3.2%
合計	3.5%	4.1%	4.9%	5.7%	+2.2%	20.6%	22.3%	25.0%	27.9%	+7.3%

い場合には、MOLAP キューブのサイズは 18,578,367 となる。値を含む組合せ数は 325,503,940 であるため、そこから 5.7%に圧縮できた。一方、許容する標準誤差を 0.9 とした場合には、MOLAP キューブのサイズは 193,673,143 となり、59.5%に圧縮できた。許容する標準誤差が 0.7 以上の場合は、いずれも必要となるサンプル数が 2 以上であるため(表 1)、結果が変わらなかった。

これをクロス数ごとに確認すると、クロス数を増やした場合に圧縮率が上がっていることがわかる。許容する標準誤差が 0.1 の例を見ると、1 クロスの場合は 99.4%のキューブが残っているのに対して、14 クロスの場合は 0.2%残っており、大きな差があることが見て取れる。

これらの結果から、現実的な分析において、±10%程度までの誤差を許容できる状況は多いと思われるため、大幅に MOLAP キューブのサイズを小さくすることができると言える。標準誤差を 0.1 とすると分析対象が少なくなりすぎる場合は、例えば標準誤差を 0.3 として 10 サンプル以下を集計しないようにすることで、標準誤差は大きくなるものの、MOLAP キューブのサイズを 27.9%として分析対象を増やすことができる。逆に、誤差が許されない状況であれば、標準誤差を 0.1 よりも小さくすることで、分析対象が少なくなるものの、さらに圧縮率を高めることが可能である。特に、ほぼ全ての属性を同時に用いるような分析は多くないと考えられるため、クロス数が多い場合に圧縮率

が高い本手法は、相性が良いと想定できる。

5.2 実験 2: 分析対象とするデータ量を変えた場合

分析対象とするデータの量を変更した際に、MOLAP キューブのサイズがどう変化するかを確認するため、扱うデータを 2012 年 1 年分、2007 年以降 2 年分…と変化させた場合の MOLAP キューブのサイズと各年数分の値を含む組合せに対する比率は表 5 の通りである。標準誤差については、例示のために 0.1 と 0.3 の場合を掲載している。

まず、合計の行を確認すると、データ量が増えるほどに圧縮率が低くなる傾向を見て取ることができ、データ量が多い場合には圧縮性能が下がることが想定される。例えば、許容する標準誤差が 0.1 の場合、2012 年 1 年分を対象とした場合は 3.5%に、2007 年以降 2 年分では 4.1%、4 年分では 5.7%に圧縮されており、それぞれ+0.6%pt、+2.2%pt 悪化している。これは、許容する標準誤差が 0.3 の場合でも同様の傾向である。

これをクロス数ごとに確認すると、許容する標準誤差によって傾向が異なっていることがわかる。許容する標準誤差が 0.1 の場合、3 クロスをピークに圧縮率が下がっている。1 年分と 4 年分で比較すると、3 クロスの際に 39.5%から 54.6%で+15.1%pt と最も悪化しており、4 クロスでは 22.5%から 34.7%で+12.3%pt、以降クロス数が増えるほど圧縮率の差が小さくなっている。一方、許容する標準誤差が 0.3 の場合、5 クロスの際に 45.2%から 57.3%で+12.1%と最

も悪化しており、以降は同傾向である。

これらの結果から、分析対象とするデータ量が線形に増えた場合、ほぼ比例して線形に MOLAP キューブのサイズも大きくなると考えられる。しかし、クロス数ごとに見るとその増加率は一定ではなく、許容する標本誤差が小さい場合にはクロス数が小さいところで増加し、許容する標本誤差が大きい場合にはクロス数が大きいところで増加する傾向があった。そのため、実用上は、分析する際のクロス数はどの程度の場合が多いかなどのどのような分析を行うのかおよび分析対象とするデータ量に応じて、許容する標本誤差を調整する必要があると考えられる。

6. 結論と今後の展望

本稿では、データ分析に影響が少ない範囲で MOLAP キューブのサイズを小さくし、巨大なデータが対象であったとしてもローカルでの MOLAP による分析を可能とすることを目的とし、標本誤差の影響が大きくなる集計値については集計を行わないことで、データ分析への影響が少ない範囲で集計パターンを大幅に減らし、現実的に分析可能な範囲の MOLAP キューブを作成する手法を提案した。

実験では、許容する標本誤差、および分析対象とするデータの量を変更した場合に、MOLAP キューブのサイズがどう変化するかを確かめた。特に、許容する標本誤差を 0.1 とした場合、元の 5.7% に圧縮可能であることを示し、提案法の有効性が示された。

今後は、データベースを用いて分析作業の際のレイテンシの変化を確認するほか、実際に分析作業をしてもらい、レイテンシが低減されることによる分析の効率改善効果や、MOLAP キューブを軽量化したことによるデータ分析作業への影響を確認する必要がある。

参考文献

- [1] Tatsuo Tsuji, and Dong Jin: "A New Parallel Data Cube Construction Scheme". International Journal of Grid and High Performance Computing, Vol. 4, pp. 32-45, 2012.
- [2] Jim Gray, Adam Bosworth, Andrew Layman, and Hamid Pirahesh: "Data cube: A Relational Aggregation Operator Generalizing Group-by, Cross-tab, and Sub-totals". IEEE ICDE, pp. 152-159, 1996.
- [3] Venky Harinarayan, Anand Rajaraman, and Jeffrey D. Ullman: "Implementing Data Cubes Efficiently". ACM SIGMOD, pp. 205-216, 1996.
- [4] Doris Jung-Lin Lee, Himel Dev, Huizi Hu, Hazem Elmeleegy, and Aditya Parameswaran: "Avoiding drill-down fallacies with VisPilot: assisted exploration of data subsets". ACM IUI, pp.186-196, 2019.
- [5] Nedjar, S., Casali, A., Cicchetti, R. and Lakhali, L: "Emerging cubes for trends analysis in OLAP databases", DaWaK, Vol. 4654, pp.135-144, 2007.
- [6] Dalsu Choi, Chang-Sup Park, and Yon Dohn Chung: "Progressive top-k subarray query processing in array databases". VLDB, Vol. 12, No. 9, pp.989-1001, 2019.
- [7] Apache kylin, <http://kylin.apache.org/> (accessed 2020-1-31)

- [8] Yansong Zhang, Yu Zhang, Shan Wang, and Jiaheng Lu. "Fusion OLAP: Fusing the Pros of MOLAP and ROLAP Together for In-memory OLAP". IEEE ICDM, 2019.
- [9] Codd E.F., Codd S.B., and Salley C.T.: "Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate", ComputerWorld, 1993.
- [10] Google: "An Inside Look at Google BigQuery". 2012.
- [11] 統計センター, <https://www.nstac.go.jp/services/ippan-microdata.html> (accessed 2020-1-31)
- [12] Tableau, <https://www.tableau.com/> (accessed 2020-1-31)