Saarbrücken, den 14.05.99

Inkrementelle Verwaltung von Histogrammen und Samples

Referent: Matthias Rinck
Richard-Wagner-Str. 2d
66125 Dudweiler

Betreuer: Arnd Christian König

Seminarleiter: Prof. Dr.-Ing Gerhard Weikum

Wir leben in einem Zeitalter, in dem Begriffe wie Datenautobahn oder Datenschutz schon zum Alltag gehören. Dies hat einen ganz einfachen Grund. In unserer Welt werden immer mehr Daten elektronisch gespeichert. Dies geschieht nicht nur im Internet, sondern auch in Bereichen wie Marketing und Statistik. In solchen Bereichen ist es nicht unbedingt wichtig, die einzelnen Daten genau zu kennen, was bei großen Datenmengen zeitaufwendig ist, sondern Aussagen zu machen, die nicht genau sein müssen, sondern nur eine Tendenz anzeigen. Dafür sollen diese aber schnell verfügbar sein. Aus diesem Grund sich verschiedene Techniken entwickelt worden, um die Daten so zu verwalten, daß man schnelle Antworten auf bestimmte Fragestellungen erhält. Zwei davon sind Sampling[1] und Histogramme[2]. In diesem Referat geht es darum, wie man Samples und Histogramme verwaltet, wenn es sich nicht um eine festgelegte Datenmenge handelt, sondern um solche, bei denen mit fortschreitender Zeit neue Daten hinzukommen.

In meinem Referat beziehe ich mich besonders auf die Papiere *Fast Incremental Maintenance of Approximate Histograms* und *New Sampling-Based Summary Statistics for Improving Approximate Query Answers.* Eine vollständige Literaturliste befindet sich am Ende des Referates.

[1] In diesem Proseminar ist über das Thema *Sampling-Techniken* ein Referat von Vanessa Walter geschrieben worden.

[2] In diesem Proseminar ist über das Thema *Ein Vergleich verschiedener Histogramm-Techniken* ein Referat von Martin Wanke geschrieben worden.

Inhaltsverzeichnis:

Deckblatt 1
Einstimmung auf das Thema 2
Inhaltsverzeichnis 3
Begriffserläuterungen 4
1. Stichproben
1.1 Ersatz-Stichprobe 5
1.1.1 Verwaltung von Ersatz-Stichproben bei Datenänderungen *5*
1.2 Kompakte Stichprobe 5
1.2.1 Vor- und Nachteile von kompakten Stichproben *6*
1.2.2 Eigenschaften von kompakten Stichproben *6*
1.2.3 Erzeugung einer kompakten Stichprobe aus vorhandenen Daten *6*
1.2.4 Behandlung von neuen Daten bei kompakten Stichproben *7*
1.2.5 Festlegung der Erhöhung des Schwellenwertes *8*
1.2.6 Mehrere Werte „gleichzeitig“ behandeln *8*
1.3 Zählende Stichproben 8
1.3.1 Zählende Stichproben in kompakte Stichproben umwandeln *9*
1.3.2 Behandlung von neuen Daten bei zählenden Stichproben *9*
1.3.3 Verwaltung von zählenden Stichproben beim löschen von Daten *9*
2. Histogramme 10
2.1 Gleich hohe Histogramme (equi-depth histograms) 11
2.1.1 Fehlermaße bei gleich hohen Histogrammen *11*
2.1.2 Verwaltung von neuen Daten bei gleich hohen Histogrammen *12*
2.1.3 Auswirkungen des Parameters γ auf die Fehlermaße *12*
2.1.4 Der teilen&zusammenfassen Algorithmus (split&merge) *12*
2.1.5 Verwaltung eines Histogrammes beim löschen von Daten *13*
2.2 Kompakte Histogramme (Compressed histograms) 13
2.2.1 Besonderheiten bei kompakten Histogrammen *14*
Zusammenfassung 14
Literaturverzeichnis 15

Im folgenden Text verwende ich einige Fachbegriffe, die ich aus dem Englischen übernommen habe. Auf dieser Seite möchte ich diese zusammenfassen und einige kurz erläutern.

Stichprobe (sample): Eine Stichprobe ist ein Teil eines Ganzen, aus der man auf die Zusammensetzung des Ganze schließen kann.

Gleichmäßige zufällige Stichprobe (uniform random sample): Bei einer gleichförmige zufällige Stichprobe werden, wie der Name schon sagt, die Werte für die Stichprobe durch ein Zufallsverfahren gleichmäßig aus der Gesamtheit aller Werte ausgewählt.

Ersatz Stichprobe (backing sample): Stichproben, die sich bei Änderungen in der ursprünglichen Datenbasis anpassen lassen. Dieser Begriff kommt aus [GMP97]. Dort werden diese Stichproben im Zusammenhang mit der Verwaltung von Histogrammen benutzt.

Kompakte Stichprobe (concise sample)

Zählende Stichprobe (counting sample)

Fußabdruck (footprint): Mit diesem Ausdruck ist der Platz gemeint, der verwendet wird, um eine Stichprobe zu speichern. In diesem Begriff steckt die Anspielung darauf, daß man auch aus einem Fußabdruck auf das Wesen schließen kann, daß ihn hinterlassen hat. Man muß sich also nur den Fußabdruck merken.

Stichprobenwerte (samling point) : Mit diesem Ausdruck werden die einzelnen Werte bezeichnet, die aus der Gesamtmenge in die Stichprobe eingeflossen sind.

Stichprobengröße (sample-size): Anzahl der Stichprobenwerte in einer Stichprobe

Gleich hohe Histogramme (equi-depth histograms)

Komprimierte Histogramme (compressed histograms)

Körbe (buckets): Bereiche eines Histogrammes

1.1 Ersatz Stichprobe (Backing samples)

Eine *Ersatz-Stichprobe* ist eine *gleichmäßige zufällige Stichprobe*, die bei Änderungen der zugrundeliegenden Daten angepaßt wird.
Um eine *Ersatz-Stichprobe* zu erzeugen wird zuerst die Größe der Stichprobe auf n festgelegt. Dann werden die ersten n Werte aus der zugrundeliegenden Datenmenge in die Stichprobe geschrieben. Danach wird eine zufällige Anzahl von Werten übersprungen. Der nächste Wert wird an einer zufällig ausgewählten Stelle in der Stichprobe eingefügt und überschreibt den dort stehenden. Dann werden wieder einige Werte übersprungen, der nächste wird wieder an einer zufälligen Stelle eingefügt und so weiter, bis alle Werte abgearbeitet sind. Die Sprungweite wächst mit der Anzahl der schon betrachteten Werte, so daß jeder Wert mit der Wahrscheinlichkeit 1/N in die Stichprobe aufgenommen wird, wobei N jeweils die Anzahl der schon betrachteten Werte ist. Vergleiche hierzu auch Kapitel 1.2.6

1.1.1 Verwaltung von Ersatz-Stichproben bei Datenänderungen

Werden neue Werte hinzugefügt, werden diese behandelt, als hätten sie hinter dem letzten Wert gestanden, der betrachtet wurde. Dies ist möglich, da keine Annahmen über die Größe der Datenmenge gemacht wurden.
Werden Werte aus der Datenmenge gelöscht, wird in der Stichprobe nachgesehen, ob dieser Wert auch dort vorhanden ist. Wenn dies der Fall ist, wird er aus der Stichprobe gelöscht, andernfalls passiert nichts mit der Stichprobe. Das Löschen von Werten aus der Stichprobe führt aber dazu, daß die Stichprobe kleiner wird. Aus diesem Grund wird eine untere Schranke eingeführt. Wird diese unterschritten, wird die Stichprobe aus der aktuellen Datenmenge aufgefüllt. Solche Fälle kommen aber nur vor, wenn eine große Anzahl von Löschungen in kurzem Abstand auftreten, da ja die „Lücken“, die beim Löschen entstehen, von eingefügten Daten wieder geschlossen werden können.

1.2 Kompakte Stichproben (Concise samples)

Wenn man Fragestellungen betrachtet, bei denen es um Werte geht, die mit einer hohen Frequenz vorkommen, kommen diese natürlich auch häufiger in einer *gleichmäßigen zufälligen Stichprobe* vor. Bei diesem Verfahren verschenkt man aber wertvollen Speicherplatz, wenn die Werte öfter als zweimal vorkommen, denn man kann einen Wert w der n mal vorkommt auch als geordnetes Paar <w,n> speichern. Dabei belegt man nur 2 Speicherstellen und es werden logischerweise n-2 Speicherstellen frei.

Es ist ein sehr verbreiteter Ansatz, das man <Wert,Anzahl> Paare benutzt, der auch in zahlreichen Zusammenhängen verwendet wird. Um so verwunderlicher ist es, daß dieses mächtige Verfahren erstmals 1998[3] im Zusammenhang mit *zufälligen Stichproben* untersucht wurde.

1.2.1 Vor- und Nachteile von kompakten Stichproben

Der **Vorteil** von *kompakten Stichproben* liegt darin, daß bei gleichem Speicherplatz mehr *Stichprobenwerte* gespeichert werden können und dadurch die Qualität der Antworten steigt. Wenn man Werte, die mehr als einmal vorkommen, als <Wert,Anzahl> Paar darstellt, ist die Anzahl der *Stichprobenwerte* mindestens genau so groß wie der *Fußabdruck*. Sei $S = \{<v_1,c_1>,...,<v_j,c_j>,v_{j+1},...,v_k\}$ eine *kompakte Stichprobe*. Es gelten folgende Formeln:

$$\textit{Stichprobengröße}(S) = k - j + \sum_{i=1}^{j} c_i$$

$$\textit{Fußabdruck}(S) = k + j$$

Der **Nachteil** liegt darin, daß es schwieriger wird, neu hinzukommende Daten zu verwalten, als bei gewöhnlichen Stichproben. Wie dies funktioniert, wird weiter unten beschrieben.

1.2.2 Eigenschaften von kompakten Stichproben

Jede *kompakte Stichprobe* repräsentiert eine *gleichmäßige zufällige Stichprobe* mit der Größe *Stichprobengröße*(*S*) und kann daher problemlos in allen Techniken verwendet werden, die sich auf *gleichmäßige zufällige Stichproben* stützen.

Wenn man n Werte hat, aber nur m/2 unterschiedliche darunter sind, ist der *Fußabdruck* maximal m groß. In diesem Fall entspricht die *kompakte Stichprobe* genau einem „high-biased“ Histogramm in dem allen Wert mit ihrer Anzahl gespeichert sind. Die *Stichprobengröße* einer *kompakten Stichprobe* kann bei einem festen *Fußabdruck* eine beliebige Größe erreichen.

1.2.3 Erzeugung einer kompakten Stichprobe aus vorhandenen Daten

Um eine *kompakte Stichprobe* mit einem *Fußabdruck* von m zu erzeugen, wählt man erst zufällig m Werte aus der kompletten Datenmenge. Danach sortiert man diese Werte und faßt Sequenzen von gleichen Werten zu <Wert,Anzahl> Paaren zusammen. Danach fügt man solange neue Werte hinzu, bis man entweder alle Werte gespeichert hat, oder der *Fußabdruck* großer wird, als er soll. Im letzten Fall läßt man den letzten Wert wegfallen, damit der *Fußab-*

[3] This simple idea is quite powerful, and to the best of our knowledge, has never before been studied [GM98]

druck nicht zu groß wird. Um einen neuen Wert hinzuzufügen, schaut man erst nach, ob er schon vorkommt. Wenn er noch nicht vorkommt, fügt man ihn einfach hinzu. Kommt er als einzelner Wert vor, wird ein <Wert,Anzahl> Paar gebildet. Kommt er schon als <Wert,Anzahl> Paar vor, wird nur die Anzahl um eins erhöht.
Dieses Verfahren wird nur einmal durchgeführt. Die Anzahl der Plattenzugriffe ist gleich mit der *Stichprobengröße*. Um Zeit einzusparen kann man die <Wert,Anzahl> Paare als Hashtabelle speichern, was das nachschauen, ob ein Wert schon vorhanden ist, auf eine konstante Zeit reduziert. Wie man neu hinzukommende Daten verwaltet, wird im nächsten Abschnitt beschrieben.

1.2.4 Behandlung von neuen Daten bei kompakten Stichproben

Bei *kompakten Stichproben* ist es schwieriger, als bei gewöhnlichen Stichproben, neu hinzukommende Daten zu behandeln. Um die Korrektheit des folgenden Algorithmus einfacher zu zeigen, werden zuerst alle hinzukommenden Daten einzeln betrachtet. In Kapitel 1.2.6 wird dann gezeigt, wie man mehrere Werte „gleichzeitig“ behandeln kann.
Wir führen einen Schwellenwert t ein, der am Anfang bei 1 liegt. Wenn nun ein neuer Wert hinzukommt, wird dieser mit einer Wahrscheinlichkeit von 1/t in die *kompakte Stichprobe* eingefügt. Dabei gibt es drei Möglichkeiten. Der Wert ist schon als <Wert,Anzahl> Paar gespeichert. In diesem Fall können wir einfach die Anzahl um eins erhöhen. Wenn der Wert nur als einzelner Wert vorkommt, oder noch gar nicht vorhanden war, wird er wie oben beschrieben, eingefügt. Die letzten beiden Fälle können aber dazu führen, daß der vorgegebene *Fußabdruck* überschritten wird. Ist dies der Fall, wird der Schwellenwert auf t‘ erhöht. Alle *Stichprobenwerte* bleiben mit einer Wahrscheinlichkeit von t/t‘ in der Stichprobe. Dadurch sinkt die *Stichprobengröße* erwartungsgemäß auf das t/t‘ fache. Man muß dabei aber berücksichtigen, daß dadurch der *Fußabdruck* nicht zwangsläufig sinkt, sondern nur, wenn <Wert,Anzahl> Paare zu Einzelwerten werden, oder wenn Einzelwerte wegfallen. Da jeder Wert vor einer Schwellenwerterhöhung mit einer Wahrscheinlichkeit von 1/t in das Sample hineingekommen ist, und mit einer Wahrscheinlichkeit von t/t‘ auch darin bleibt, entspricht dies genau einer Wahrscheinlichkeit von (1/t) * (t/t‘) = 1/t‘. Aus diesem Grund gilt auch der neue Schwellenwert 1/t‘ für Werte, die vor der Erhöhung in das Sample hereingekommen sind.

1.2.5 Festlegung der Erhöhung des Schwellenwertes

Bei der Wahl der Erhöhung des Schwellenwertes ist man also ziemlich frei. Wenn man den Schwellenwert stark erhöht, kann es sein, daß die *Stichprobengröße* mehr gesenkt wird, als notwendig ist, was eine schlechtere Repräsentation der gesamten Datenmenge zur Folge hat. Andererseits ist dann aber mehr Platz, um neue Daten hinzuzufügen, ohne jedesmal den Schwellenwert erhöhen zu müssen, was ja zeitaufwendig ist. Wenn man den Schwellenwert zu wenig erhöht, kann es sein, daß der *Fußabdruck* nicht genügend gesenkt wird und der Schwellenwert nochmals erhöht werden muß. Eine Erhöhung des Schwellenwertes um 10% hat sich als sinnvoll herausgestellt.

1.2.6 Mehrere Werte „gleichzeitig“ behandeln

Anstatt jeden Wert mit einer Wahrscheinlichkeit von 1/t in die Stichprobe aufzunehmen, kann man auch mit einer Wahrscheinlichkeit von $(1-1/t)^i * (1/t)$ die nächsten i Werte ignorieren und den nächsten Wert hinzufügen. Dabei spart man Zeit, vor allem bei großen Werten für t. Bei t = 5 werden mit einer Wahrscheinlichkeit von 51,2 % die nächsten 3 Werte ignoriert. Bei t = 10 werden mit einer Wahrscheinlichkeit von ca 53% die nächsten 6 Werte ignoriert.
Auch bei der Erhöhung des Schwellenwertes kann man auf ähnliche Weise Zeit sparen. Man muß nicht für jedes einzelne Vorkommen des gleichen Wertes eine Entscheidung treffen. Wird der Schwellenwert um 10% erhöht, bleiben die nächsten 7 Vorkommen des Wertes mit einer Wahrscheinlichkeit von ca. 51 % in der Stichprobe erhalten.

1.3 Zählende Stichproben (counting sample)

Zählende Stichproben sind eine Variation von *kompakten Stichproben*, bei denen sich die Anzahl eines Wertes auf die Gesamtzahl des Vorkommens eines Wertes in der gesamten Datenmenge bezieht.
Um eine *zählende Stichprobe* mit einem Schwellenwert von t zu erhalten, wird für jeden Wert, der c > 0 mal vorkommt, eine Münze geworfen, die mit einer Wahrscheinlichkeit von 1/t Kopf zeigt. Die Münze wird maximal c mal geworfen. Zeigt sie nie Kopf an, wird der Wert nicht übernommen. Zeigt die Münze beim i-ten Wurf Kopf, wird der Wert mit der Anzahl c-i+1 als <Wert,Anzahl> Paar gespeichert, vorausgesetzt, das Ergebnis ist noch größer als 1. Ist das Ergebnis gleich 1, wird der Wert als einzelner Wert gespeichert.

1.3.1 Zählende Stichproben in kompakte Stichproben umwandeln

Ein *zählende Stichprobe* entspricht nicht direkt einer *gleichmäßigen zufälligen Stichprobe*. Man kann aber aus einem *zählenden Stichprobe* ein *kompakte Stichprobe* erzeugen. Dies geschieht auf folgende Weise. Für jeden Wert, der c>1 mal vorkommt, wird eine Münze mit der Wahrscheinlichkeit 1/t daß Kopf fällt, c-1 mal geworfen. Jedesmal wenn nicht Kopf fällt, wird die Anzahl der Vorkommen des Wertes um eins reduziert.

1.3.2 Behandlung von neuen Daten bei zählenden Stichproben

Am Anfang haben wir einen Schwellenwert von t=1. Kommt ein neuer Wert hinzu, der schon als <Wert,Anzahl> Paar vorhanden ist, wird die Anzahl einfach um eins erhöht. Ist der Wert als Einzelwert vorhanden, wird ein <Wert,Anzahl> Paar gebildet. Kommt der Wert noch nicht vor, wird er mit einer Wahrscheinlichkeit von 1/t hinzugefügt. Durch die letzten beiden Aktionen kann der *Fußabdruck* den angenommenen Wert überschreiten. In diesem Fall wird der Schwellwert t angehoben, um Platz frei zu bekommen.
Für jeden Wert, der in der Stichprobe vorkommt, wird eine Münze geworfen. Diese hat beim ersten Wurf eine Wahrscheinlichkeit von t/t‘ für Kopf, danach 1/t‘. Jedesmal wenn Zahl fällt, wird die Anzahl des Wertes um eins erniedrigt. Dies geschieht, bis die Anzahl auf 0 gesunken ist, dann wird der Wert aus der Stichprobe entfernt, oder bis Kopf oben ist. Auch hier kann die Anzahl der Münzwürfe mit einem ähnlichen Verfahren, wie oben geschildert, reduziert werden.

1.3.3 Verwaltung von zählenden Stichproben beim löschen von Daten

Ein Vorteil von *zählenden Stichproben* gegenüber *kompakten Stichproben* ist es, daß es einfacher ist, Löschungen von Daten aus der Stichprobe zu behandeln. Bei *kompakten Stichproben* ist dies schwierig, weil die Verteilung der Werte dadurch verzerrt werden kann. Da man bei *zählenden Stichproben* gegenüber *kompakten Stichproben* einen direkten Bezug zu allen Daten aus der Datenmenge hat, müssen wir nur nachschauen, ob der zu löschende Wert vorhanden ist. Ist dies der Fall, wird die Anzahl der Vorkommen um eins erniedrigt, andernfalls passiert nichts.
Es muß jetzt noch gezeigt werden, daß die Eigenschaften von *zählenden Stichproben*, die oben beschrieben sind, beim einfügen und löschen von Daten nicht verletzt werden.

Wird ein Wert hinzugefügt, steigt seine Anzahl in der Datenmenge um eins. War dieser Wert in der *zählenden Stichprobe* enthalten, zeigte einer seiner Münzwürfe Kopf und die Anzahl der Vorkommen wird um eins erhöht. Andernfalls zeigte keiner der Würfe für diesen Wert Kopf und der Wert wird mit der Wahrscheinlichkeit 1/t hinzugefügt. Alle anderen Werte bleiben unberührt.

Wird ein Wert gelöscht, sinkt seine Anzahl in der Datenmenge um eins und auch die Anzahl in der Strichprobe wird um eins erniedrigt, was dazu führen kann, daß der Wert ganz aus der Stichprobe fällt. War ein zu löschender Wert nicht in der Stichprobe enthalten, zeigten alle c Münzwürfe Zahl, also zeigten auch die ersten c-1 Münzwürfe Zahl und der Wert gehört nicht in die Stichprobe.

Wird der Schwellenwert von t auf t‘ erhöht, und kommt ein Wert c > 0 mal in der Datenmenge vor, gilt folgendes. Kam dieser Wert nicht in der Stichprobe vor, ist eine Münze mit der Wahrscheinlichkeit eines Kopfes von 1/t c mal auf Zahl gefallen. Wären diese Würfe mit der neuen Münze, bei der die Wahrscheinlichkeit eines Kopfes nur noch 1/t‘ ist, wäre auch jedesmal Zahl gefallen.

Wenn der Wert in der Stichprobe mit einer Anzahl von c‘ enthalten ist, dann gab es c-c‘ Münzwürfe mit der Wahrscheinlichkeit eines Kopfes von 1/t, die Zahl zeigten. Diese hätten auch mit der neuen Münze (1/t‘) Zahl gezeigt. Danach ist die Münze mit der Wahrscheinlichkeit von 1/t auf Kopf gefallen. Nach dem Algorithmus muß zuerst eine Münze mit der Wahrscheinlichkeit t/t‘ geworfen werden. Das ergibt zusammen (1/t) * (t/t‘) = (1/t‘), was der Wahrscheinlichkeit der neuen Münze entspricht. Wenn diese Münze Zahl zeigt, ist die Wahrscheinlichkeit der weiteren Münzwürfe auch 1/t‘. Die Bedingungen einer *zählender Stichprobe* werden also nicht verletzt.

2. Histogramme

Seien D alle möglichen Werte, die in einer Datenmenge vorkommen können und V alle Werte, die tatsächlich darin vorkommen, mit $V = \{v_i: 1 \le i \le |V|\}$, wobei $v_i < v_j$, wenn i < j. Die Frequenz, also die Anzahl der Vorkommen eines Wertes wird mit f_i bezeichnet. Die Datenverteilung ist die Menge $\Gamma = \{(v_1, f_1), (v_2, f_2), \ldots, (v_{|V|}, f_{|V|})\}$.

Ein Histogramm besteht aus β ($\ge$1) elementfremden Teilmengen der Menge Γ, so daß alle Werte in einer Menge aufeinanderfolgend sind. Diese Teilmengen werden auch *Körbe* genannt. Zwei Möglichkeiten, die Anzahl der Werte in einem *Korb* festzulegen, sind: Es werden entweder alle m Werte aufgezählt, die sich zwischen dem größten und dem kleinsten Wert in einem *Korb* befinden, oder nur $k \le m$ Werte, die dazwischen liegen und den gleichen

Abstand voneinander haben. Die Anzahl der Werte in einem *Korb B* wird auch mit f_B bezeichnet. Der Näherungswert für das Vorkommen eines Wertes in einem *Korb* ergibt sich aus den Formeln f_B / m bzw. f_B / k.

2.1 Gleich hohe Histogramme (equi-depth histograms)

Die verschiedenen Histogramme unterscheiden sich darin, wie die Gesamtmenge der Werte auf die *Körbe* verteilt wird. Bei *gleich hohen Histogrammen* werden alle vorkommenden Werte so aufgeteilt, daß die Anzahl von Werten f_B in jedem Korb *B* gleich groß ist. Bei *komprimierten Histogrammen* werden die *n* häufigsten Werte einzeln gespeichert, und die restlichen werden wie bei *gleich hohen Histogrammen* aufgeteilt. Der Vorteil von Histogrammen liegt darin, daß sie weniger Speicherplatz benötigen, als die entsprechende *Ersatz Stichprobe*, dafür sind diese aber bei bestimmten Datenverteilungen nicht so genau, wie eine Stichprobe. Um ein Histogramm zu speichern, wird von jedem *Korb B* der größte Wert *B*.maxwert und die Anzahl der Werte in diesem *Korb B*.count gespeichert.

2.1.1 Fehlermaße bei gleich hohen Histogrammen

Bei gleich hohen Histogrammen gibt es zwei wichtige Fehlermaße, die angeben, wie gut das Histogramm nach Änderungen der zugrundeliegenden Datenmenge noch ist. Das μ_{count} Fehlermaß, daß auch für andere Histogrammarten gilt und zu den Verteilungsfehlern gehört, gibt an, wie groß der Unterschied zwischen dem gespeicherten Histogramm und der tatsächlichen Datenverteilung ist. In den folgenden Formeln ist *N* die Anzahl der Werte, die in der gesamten Datenmenge vorkommen.

$$\mu_{\text{count}} = \frac{\beta}{N}\sqrt{\frac{1}{\beta}\sum_{i=1}^{\beta}(f_{B_i} - B_i.\text{count})^2}$$

Das andere Fehlermaß μ_{ed} gilt nur für *gleich hohen Histogrammen* und gibt an, wie groß der Unterschied zwischen der idealen *Korbgröße* N/β und der tatsächlichen *Korbgröße* f_B ist. Die Formel lautet hier ganz ähnlich.

$$\mu_{\text{ed}} = \frac{\beta}{N}\sqrt{\frac{1}{\beta}\sum_{i=1}^{\beta}\left(f_{B_i} - \frac{N}{\beta}\right)^2}$$

Die beiden Formeln sind mit der *durchschnittlichen Korbgröße* N/β normalisiert.

2.1.2 Verwaltung von neuen Daten bei gleich hohen Histogrammen

Für die Verwaltung von hinzukommenden Daten wird eine *Ersatz Stichprobe* mit der Größe *m* angelegt. Dabei muß die Anzahl der *Körbe* mindestens 3 sein, und es muß ein $c \geq 4$ geben, so daß $m=(c \ln^2 \beta)\beta$. Außerdem muß $N \geq m^3$ sein. Wird das Histogramm aus dieser Stichprobe erzeugt, sich die Fehlermaße mit einer Wahrscheinlichkeit von mindestens $1-\beta^{-(c^{1/2}-1)} - N^{-1/3}$ kleiner als $\alpha=(c \ln^2 \beta)^{-1/6}$.

Man kann jetzt eine obere Schranke $T= (2+\gamma)N'/\beta$ einführen, wobei N‘ die Anzahl der Werte ist, die sich beim erzeugen der Histogrammes aus der Stichprobe in der Datenmenge befunden haben. γ ist ein wählbarer Parameter, der größer als –1 sein muß. Werden jetzt neue Werte eingefügt, wird einfach die Größe des entsprechenden *Körbe* erhöht. Dies geschieht solange, bis ein Korb die obere Schranke erreicht. In diesem Fall wird das Histogramm aus der Ersatz Stichprobe neu berechnet, wobei sich natürlich auch das N‘ und somit die neue Schranke ändert.

2.1.3 Auswirkungen des Parameters γ auf die Fehlermaße

Der Parameter γ hat nur geringe Auswirkungen. Die Formel für die neue Wahrscheinlichkeit lautet $1-\beta^{-(c^{1/2}-1)} -(N/(2+\gamma))^{-1/3}$. Hier wirkt sich der Parameter, wenn er klein ist, und das N groß ist, also kaum aus. Das Fehlermaß μ_{count} bleibt von diesem Parameter völlig unberührt. Das Fehlermaß μ_{ed} ist noch kleiner als $\alpha+(1+\gamma)$, also nur um $(1+\gamma)$ größer als vorher.

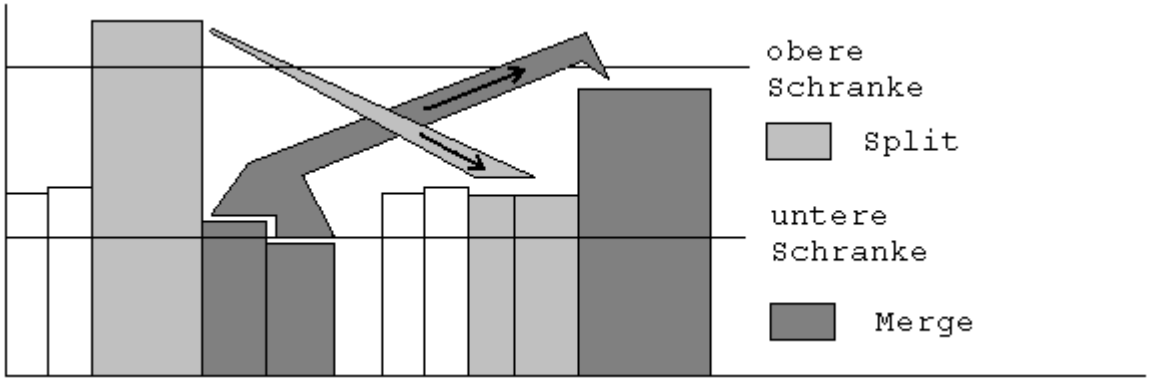
2.1.4 Der teilen&zusammenfassen Algorithmus (split&merge)

Wird die obere Schranke von einem *Korb* durchbrochen, wird dieser in zwei gleich große aufgeteilt. Durch diesen Vorgang gibt es natürlich einen *Korb* mehr, als vorgeschrieben. Um wieder auf die vorgegebene Anzahl zu kommen, werden einfach zwei nebeneinanderstehende *Körbe*, die in ihrer Summe weniger Werte haben, als es die obere Schranke vorschreibt, zusammengeworfen. Dieser Algorithmus kann natürlich nur angewendet werden, wenn die obere Schranke nicht zu „streng“ gezogen wird, der Parameter γ also nahe bei der –1 liegt. Mit der Wahl dieses Parameters kann man also bestimmen, wie oft es notwendig wird, daß man das Histogramm neu aus der Ersatz Stichprobe berechnen muß. Je größer er wird, um so weniger ist dies notwendig, aber dafür steigt mit ihm, wie oben beschrieben, das Fehlermaß $\mu_{ed.}$ Der Parameter $\gamma=0.5$ hat sich als sinnvoll erwiesen.

2.1.5 Verwaltung eines Histogrammes beim löschen von Daten

Genau wie man eine obere Schranke einführen kann, kann man auch eine untere Schranke einführen, die dafür sorgt, daß es keine *Körbe* gibt, die eine bestimmt Größe unterschreiten. Die Formel für die untere Schranke lautet $T_u = N/(\beta(2+\gamma_u))$, wobei $\gamma_u > -1$. Wird die untere Schranke durch einen *Korb* unterschritten, wird dieser mit einem seiner „Nachbarn" zusammengefügt. Danach wird der größte *Korb* gesucht, der mindestens doppelt so groß sein muß, wie die unterste Schranke, und in zwei gleich große *Körbe* geteilt. Dies kann sowohl der neu entstandene, als auch ein anderer sein. Gibt es keinen solchen, wird das Histogramm neu aus dem gespeicherten Sample berechnet.

Skizze des Split- und Merge-Vorganges



2.2 Kompakte Histogramme (Compressed histograms)

Kompakte Histogramme sind Histogramme, bei denen eine bestimmte Anzahl von Werten, die am häufigsten vorkommen, direkt mit ihrer Anzahl gespeichert werden. Die restlichen Werte, die weniger oft vorkommen, werden wie bei *gleich hohen Histogrammen* auf die restlichen *Körbe* verteilt. Die Anzahl der gleich hohen *Körbe* wird mit β' bezeichnet, die Anzahl aller Werte, die in diesen *Körben* sind, mit N'. Für diese gilt die gleiche Formel für das Fehlermaß μ_{ed}, bloß mit β' und N'.

2.2.1 Besonderheiten bei kompakten Histogrammen

Bei *kompakten Histogrammen* gilt im Prinzip das gleiche, wie *bei gleich hohen Histogrammen*, nur daß es noch einige Besonderheiten zu beachten gibt.

- Durch einfügen neuer Werte kann ein Wert, der noch nicht als einzelner Wert aufgeführt wird, so oft vorkommen, daß er nun doch als einzeln aufgeführt werden müßte. In diesem Fall ist der *Korb*, in dem dieser Wert vorkommt, aber ziemlich groß und es kann wie beim überschreiten der oberen Schranke in zwei *Körbe* aufgeteilt werden, von denen einer nur den oft vorkommenden Wert enthält.
- Wenn viele Werte in *Körbe* eingefügt werden, die für mehrere Werte sind, kann es dazu kommen, daß die kleinsten Einzelwerte nicht mehr zu den Einzelwerten zählen. In diesem Fall werden diese „falschen“ Einzelwerte in einem *Korb* zusammengefaßt.
- Wenn Werte, die in einen einzelnen *Korb* gehören, oft gelöscht werden, kann dieser so klein werden, daß er nicht mehr als einzelner Wert geführt wird. Dann muß er mit einem anderen zusammengefaßt werden.

Die Anzahl der Werte, die häufig vorkommen, und daher als Einzelwerte gespeichert werden, kann sich bei einem *kompakten Histogramm* beim einfügen und entfernen von Werten aus der Datenmenge ändern.

Zusammenfassung

Zuerst haben wir die Ersatzstichprobe kennengelernt, die sowohl beim Löschen als auch beim Einfügen von Daten angepaßt werden konnte. Danach haben wir Kompakte Stichproben eingeführt, die es erlaubten, mehr Stichprobenwerte bei gleichem Speicherbedarf zu speichern. Löschungen in der Datenmenge konnten aber nur noch schwer durchgeführt werden. Aus diesem Grund wurden Zählende Stichproben eingeführt, die auch bei Löschungen aus der Datenmenge angepaßt werden können, aber keine gleichmäßige zufällige Stichprobe mehr darstellen. Sie können aber leicht in eine solche umgewandelt werden.
Von den Stichproben kamen wir zu den Histogrammen, die weniger Speicher brauchen, die aber schwieriger zu verwalten sind, wenn sich die zugrundeliegende Datenmenge ändert. Wir haben auch gesehen, daß man Histogramme auch aus Stichproben erzeugen kann und nicht immer die komplette Datenmenge betrachten muß, wenn man Histogramme neu erstellt.

Literaturverzeichnis

[GMP97] Phillip B. Gibbons, Yossi Matias, Viswanath Poosala
Fast Incremental Maintenance of Approximate Histograms, 23rd VLDB Conference, 1997

[GM98] Phillip B. Gibbons, Yossi Matias
New Sampling-Based Summary Statistics for Improving Approximate Query Answers, ACM SIGMOND Conference, 1998

[BDF97] Daniel Barbara, William DuMouchel, Christos Faloutsos, et al.
The New Jersey Data Reduction Report, IEEE Bulletin of the Technical Committee of Data Engineering, 12/1997