Ausarbeitung

zum Vortrag

Ein Vergleich verschiedener Histogramm-Techniken

im Rahmen des Proseminars

Datenreduktionstechniken

Sommersemester 1999

Universität des Saarlandes

Martin Wanke

Mai 1999

Inhaltsverzeichnis

1 Einleitung **4**

2 Vorgehen **4**

3 Beispiel **5**

3.1 Gleiche Breite 5

3.2 Gleiche Tiefe 6

4 Kategorisierung von Histogrammen **7**

4.1 Kriterien 7

4.1.1 Klasse 7

4.1.2 Bedingung 8

4.1.3 Sortier-Parameter 8

4.1.4 Quell-Parameter 9

4.2 Annahmen über Werte und deren Verteilung 9

4.3 Schreibweisen 9

5 Bisherige Histogramme **9**

5.1 Triviales Histogramm 10

5.2 gleiche Summe(V,S) (= „gleiche Breite“) 10

5.3 gleiche Summe(V,F) (= „gleiche Höhe bzw. Tiefe“) 10

5.4 Spline-basiert(V,C) 10

5.5 Varianz-optimiert(F,F) 10

6 Neue Herangehensweisen **11**

6.1 Kriterien 11

6.1.1 Bedingung 11

6.1.2 Parameter 12

6.2 Beispiel für Annahmen über Werteverteilungen 12

7 Neue Histogramme **13**

7.1 Varianz-optimiert(V,F), (V,A), (A,A), (V,C) 13

7.2 Maximale Differenz(V,F), (V,A) 13

7.3 Komprimiert(V,F), (V,A) 13

8 Techniken zur Erstellung von Histogrammen **14**

8.1 Bereichsgrenzen (Quantilen) 14

8.1.1 Reservoir sampling 14

8.2 Weitere Parameter 15

9 Experimente **15**

9.1 Erstellungszeiten 15

9.2 Fehleranfälligkeit 15

9.3 Größe der Stichproben 16

9.4 Verteilungsannahmen 16

10 Fazit **16**

11 Zusammenfassung **17**

1 Einleitung

Um effizient Anfragen in Datenbanksystem bearbeiten zu können, benötigt man zusammenfassende Statistiken über die Datenverteilung in den zugrundeliegenden Relationen des Datenbanksystems. Damit lassen sich z.B. Schätzungen über die Anzahl oder Anteile von Tupeln, die bestimmte Anfragen erfüllen, treffen. Hierzu ist im Vergleich zu der Größe der gesamten Datenbank relativ wenig Platz notwendig; außerdem stehen die Daten im Gegensatz zum Sampling nicht nur temporär zur Verfügung. Aus diesem Grund sind Histogramme die meistbenutzte Technik bei Datenbanksystemen für zusammenfassende Informationen.
Die gewonnenen Informationen kann man nutzen, um Ausführungzeiten für Anfragen zu schätzen, da die Anzahl der betrachteten Tupel (z.B. bei Joins) eine entscheidende Rolle spielt. Auch Ausführungspläne für Anfragen lassen sich so entwickeln, da hierbei die Reihenfolge der Operationen bzw. die Größe der Zwischenresultate relevant ist. Weiterhin können schon Ergebnisse zurückgegeben werden, falls nur die Größenordnung und keine exaktes Ergebnis von Interesse ist.
Somit können Histogramme zu wesentlichen Laufzeiteinsparungen beitragen.

2 Vorgehen

Definition: Sei R eine Relation. Dann nennt man die Menge
$T = \{(v_1, f_1), (v_2, f_2), \ldots, (v_{|D|}, f_{|D|})\}$ die „Datenverteilung auf R“,
wobei die $v_i \in V$ die verschiedenen Werte und die f_i die Häufigkeiten ihres Auftretens darstellen.
D ist die „Wertemenge“, also die Menge aller möglichen Werte, die ein bestimmtes Attribut annehmen kann; V ist Menge der in R tatsächlich vorhandenen Werte.

Ferner definiert man T^c als die Menge $\{(v_1, c_1), (v_2, c_2), \ldots, (v_{|D|}, c_{|D|})\}$, wobei die c_i für die „Gesamthäufigkeit“ der Tupel steht, d.h. die Mächtigkeit der Menge $|\{t : t \in R; t.X \leq v_i\}|$.

Ein Histogramm stellt eine Einteilung in β Bereiche ($N \ni \beta \geq 1$) bezüglich eines bestimmten Sortierschlüssels dar, um die Verteilung der Werte v_i und die dazugehörigen Häufigkeiten f_i wiederzugeben.
Ein Extremfall für β ist 1 einerseits (das sogenannte „triviale Histogramm“, bei dem alle Tupel einem Bereich zugeordnet werden); dies entspricht der Gleichverteilungsannahme für die gesamte Relation. Andererseits wäre ein Einteilung in N Bereiche denkbar, wobei N die Anzahl der Tupel in R ist.
Je größer β ist, desto größer ist auch die Genauigkeit der aus dem Histogramm ableitbaren Informationen; jedoch wird auch mehr Speichplatz

benötigt (dieser aber ist eine wertvolle Ressource, da man Histogramme wegen des schnellen Zugriffs im allgemeinen gerne im Hauptspeicher eines Datenbanksystems speichert).
Pro Bereich speichert man normalerweise den jeweils kleinsten und größten vorkommenden Wert sowie die Anzahl und die durchschnittliche Häufigkeit der Tupel.

Histogramme müssen regelmäßig bei Änderungen des Datenbestandes aktualisiert werden; das gilt nicht für jede Veränderung, aber es sollte gewährleistet sein, daß die Histogramme noch ihren Zweck erfüllen können, nämlich den gesamten Datenbestand der Relation zu repräsentieren.

Bei Anfragen (z.B. der Form $a \leq X \leq b$) kann man für bestimmte Intervalle $B = [v_*(B), wobei v^*(B)]$, $v_*(B), v^*(B)$ der niedrigste bzw. höchste Wert im betrachteten Intervall B ist, prüfen, ob sich das Anfrageintervall und bestimmte Abschnitte des Histogramms überschneiden, also ob $[a, b] \cap [v_*(B), v^*(B)] \neq \emptyset$ gilt.

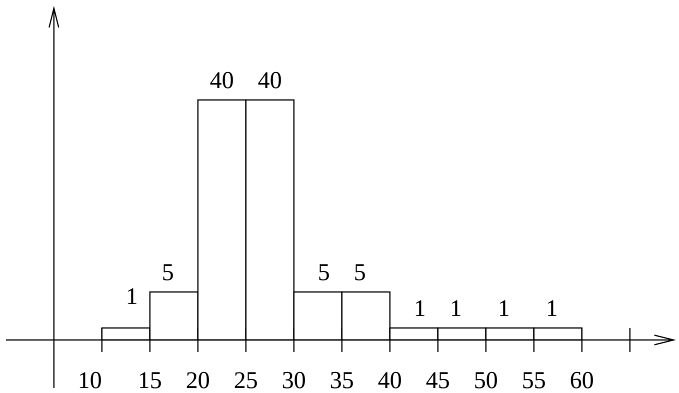
3 Beispiel

3.1 Gleiche Breite

Gegeben sei eine Relation, die das Alter der 100 Angestellten einer Firma wiedergibt.

Betrachte zunächst ein Histogramm, bei dem alle Bereich Tupel aus jeweils gleich großen Teilmengen des Wertebereichs enthalten; man spricht hierbei auch von „gleicher Breite" der Bereiche (vgl. Abbildung 1).

Abbildung 1: „Gleiche-Breite"-Histogramm



Will man nun z.B. die Anzahl der Leute bestimmen, deren Alter geringer als 29 ist, so gilt folgende Abschätzung:

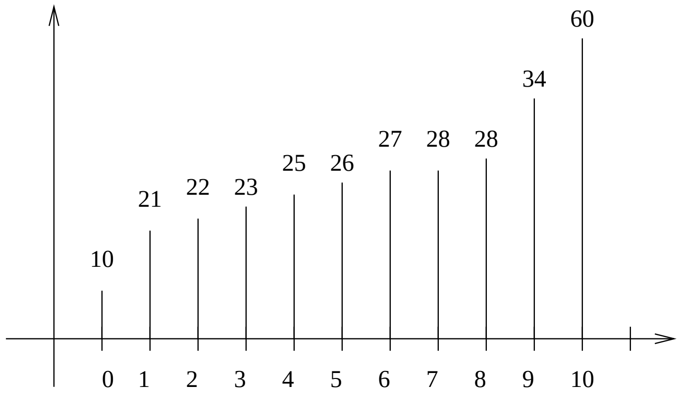
(Erinnerung: die Selektivität SEL stellt den Anteil der Tupel, die die betrachte Anfrage erfüllen, an der Anzahl aller Tupel der entsprechenden Relation dar).

$0,46 \leq \text{SEL}\ (< 29) \leq 0,86$, d.h. mindestens 46 der 100 Leute sind auf jeden Fall jünger als 29 (nämlich genau die aus den ersten drei Bereichen des Histogramms); 14 Leute sind auf jeden Fall älter (nämlich genau die aus den letzten sechs Bereichen des Histogramms); für die Tupel aus dem Bereich „25 - 30 Jahre alt“ läßt sich keine Aussage treffen, da diese Leute sowohl jünger als auch älter als 29 sein könnten (natürlich auch genau 29).
Geht man davon aus, daß sich die gesuchte Anzahl genau in der Mitte zwischen den beiden Grenzen befindet, so beträgt dieser 0,66; dies bedeutet also, daß man einen maximalen Fehler von 0,20 hat.

3.2 Gleiche Tiefe

Das Problem beim gerade vorgestellten Ansatz ist, daß die „Ausreißer“ am Rand des Wertebereichs die Mehrzahl der Werte zusammendrängen, so daß sich für den entsprechenden Abschnitt nur noch recht ungenaue Aussagen treffen lassen.
Daher erscheint es plausibel, die Histogrammbereiche nicht nach den Werten selbst, sondern deren Anzahl pro Bereich einzuteilen, so daß sich in jedem Bereich in etwa gleich viele Tupel befinden. Bei diesem Verfahren spricht man auch von „gleicher Tiefe“ oder „gleicher Höhe“ der Bereiche.
Für das betrachtete Beispiel hat ein entsprechendes Histogramm ein Aussehen gemäß Abbildung 2.

Abbildung 2: „Gleiche-Höhe“-Histogramm



Die Selektivität der Leute, die jünger als 29 sind, läßt sich nun wie folgt abschätzen:
$0.80 \leq \text{SEL}\ (< 29) \leq 0.90$, da die Leute aus den ersten acht Bereichen auf jeden Fall jünger und die Leute aus dem letzten Bereich auf jeden Fall älter

als 29 sind.
Analog zu obiger Annahme kann man davon ausgehen, daß das gesuchte Ergebnis 0,85 beträgt; der maximale Fehler dabei beträgt also 0,05.

Bei diesem Ansatz beträgt der maximale Fehler also 25% des maximalen Fehlers der vorigen Methode.

4 Kategorisierung von Histogrammen

4.1 Kriterien

Offensichtlich gibt es also mehrere Arten von Histogrammen. Im vorigen Beispiel wurde die Bereichseinteilung zunächst nach den Werten der Tupel, dann nach deren Häufigkeit vorgenommen. Außerdem gibt es noch weitere Kriterien für den Aufbau von Histogrammen, nach denen sich diese in Kategorien einteilen lassen:

- Klasse
  die grundsätzliche Methode, nach der ein Histogramm aufgebaut ist (im vorigen Beispiel seriell, d.h. die Histogrammbereiche entsprechen Gruppen von Elementen aus T in fortlaufender Reihenfolge bezüglich der Sortierung)

- Bedingung
  Die Bedingung, nach der man die Bereichseinteilung vornimmt (im vorigen Beispiel sollte die Summe von Elementen, d.h. die Werte der Tupel bzw. deren Häufigkeit in jedem Bereich gleich sein)

- Sortier-Parameter
  Das Kriterium, nach dem die Elemente von T sortiert werden (im vorigen Beispiel ein nummerische Sortierung der Werte der Tupel), also die Festlegung einer Reihenfolge der Tupel

- Quell-Parameter
  Das Kriterium, auf das die Bedingung angewendet wird (im Beispiel die Werte der Tupel bzw. deren Häufigkeit)

In der Literatur sind für die einzelnen Punkte folgende Möglichkeiten bekannt:

4.1.1 Klasse

- „seriell“, d.h. die Histogrammbereiche entsprechen Gruppen von Elementen aus T in fortlaufender Reihenfolge bezüglich der Sortierung; dies ist wohl die intuitivste Herangehensweise.

- „End-ausgerichtet“, d.h. jeder bis auf einen Bereich enthält nur ein Element. Hierbei brauchen keine Bereichsgrenzen, sondern nur der einzelne Wert gespeichert zu werden.
  Mit Hilfe dieser Methode lassen sich Extremwerte besonders gut abschätzen, allerdings zu Ungunsten der Genauigkeit an anderer Stelle.

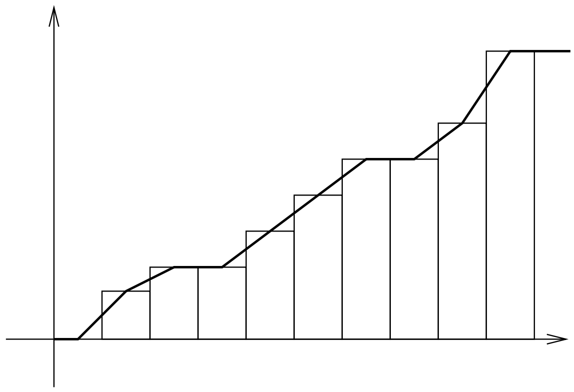
4.1.2 Bedingung

- gleiche Summe: Die Summe der Quell-Parameter pro Bereich beträgt jeweils $\frac{1}{\beta}\cdot$ Summe aller Werte im Histogramm.

- Varianz-optimiert: die gewichtete Gesamt-Varianz $\sum_{i=1}^{\beta} n_i \cdot V_i$ wird minimiert, wobei n_i die Anzahl der Tupel im Bereich i und V_i die Varianz der Werte im Bereich i ist.

- spline-basiert: Hier liegt die Idee zugrunde, die Datenverteilung durch eine abschnittsweise definierte Funktion zu approximieren; normalerweise wird der Einfachheit halber eine lineare Funktion betrachtet, die Abschnitte des Definitionsbereichs entsprechen den Bereichen des Histogramms.
  Mathematisch ausgedrückt wird bei dieser Methode die maximale Differenz aller einzelnen Werte zum Bereichs-Durchschnitt minimiert.
  Ein Beispiel ist ein Abbildung 3 zu sehen.

Abbildung 3: Beispiel einer Näheringskurve



4.1.3 Sortier-Parameter

Mögliche Sortier-Kriterien sind z.B. die Werte der Tupel und deren Häufigkeit.

4.1.4 Quell-Parameter

Die Bedingung bezieht sich im allgemeinen auf die Häufigkeit f_i, die Gesamthäufigkeit c_i oder den Abstand s_i (zum nächsten Element bzgl. V). Dabei ist der Abstand s_i eines Elements v_i definiert als $s_i := v_{i+1} - v_i$.

4.2 Annahmen über Werte und deren Verteilung

Bei allen Histogrammen geht man davon aus, daß alle Werte in einem Bereich gleich häufig vorkommen.

Außerdem gibt es mehrere mögliche Annahmen über die Verteilung der Daten innerhalb eines Bereichs:

- <u>fortlaufende Werte:</u> Alle Werte des entsprechenden Intervalls sind vorhanden. Falls es sich bei dem betrachteten Intervall um eine unendliche Menge handelt, so müssen die Werte gegebenenfalls interpoliert werden (z.B. ein Rundung bei einem Intervall von reellen Zahlen).
- <u>punktweise Betrachtung:</u> Hierbei geht man davon aus, daß nur ein Wert pro Bereich vorhanden ist; normalerweise geht man hierbei vom niedrigst-möglichen Wert aus.
- <u>keine expliziten Annahmen:</u> Hierbei muß jeder Wert gespeichert werden.

4.3 Schreibweisen

Für ein Histogramm mit Bedingung p, Sortier-Parameter s und Quell-Parameter u schreibt man $p(s, u)$.

Dabei werden für s und u folgende Abkürzungen verwendet:
S der Abstand s_i zum nächsten Element bzgl. V
V die Werte der Tupel
F die Häufigkeit der Tupel
C die Gesamthäufigkeit der Tupel.

Außerdem geht man normalerweise von fortlaufenden Werten aus.

5 Bisherige Histogramme

Bisher wurden in in der Literatur folgende Arten von Histogrammen betrachtet:

5.1 Triviales Histogramm

Das triviale Histogramm besteht nur aus einem Bereich und entspricht somit der Gleichverteilungsannahme für die gesamte Relation R.

5.2 gleiche Summe(V,S) (= „gleiche Breite“)

Hierbei besteht jeder Bereich aus einem zusammenhängenden Intervall von Attributwerten, und zwar jeweils einem Intervall von $\frac{1}{\beta}\cdot$ Gesamtintervall (bzgl. V).

5.3 gleiche Summe(V,F) (= „gleiche Höhe bzw. Tiefe“)

Bei diesem Histogrammtyp enthält jeder Bereich gleich viele Tupel aus R. Das bedeutet insbesondere, daß ein Wert in mehreren Bereichen auftreten kann, sofern er z.B. häufiger auftritt, als überhaupt Elemente in jedem Bereich gespeichert werden.

5.4 Spline-basiert(V,C)

Hierbei versucht man, die Datenverteilung T^c durch eine Nährungskurve zu approximieren; hierbei verwendet man im allgemeinen abschnittsweise definierte lineare Funktionen (vgl. Abschnitt 4.1.2). Problematisch dabei ist die Frage, an welchen Stellen die Stützstellen der Nährungskurve liegen sollten (das sogenannte „optimal knot placement problem“, für das keine effizenten Algorithmen, sondern nur Heuristiken existieren, die ein approximierte Lösung liefern).
Da bei dieser Art von Histogrammen nur der approximierte Wert für jeden Bereich (und nicht ein kleinster und größter Wert) gespeichert werden muß, steht mehr Platz für die Bereiche selber zur Verfügung.

5.5 Varianz-optimiert(F,F)

Hierbei werden die Tupel nach ihrer Häufigkeit angeordnet; die Varianz der Gesamthäufigkeit wird minimiert. Werte, die in etwa gleich häufig auftreten, liegen also im Histogramm normalerweise in gleichen Bereichen.
Allerdings hat die Konstruktion derartiger Histogramme exponentielle Komplexität. In der Praxis verwendet man daher randomisierte Algorithmen, die approximative Lösungen liefern.
Außerdem ist hierbei ein Index über alle Werte erforderlich, um feststellen zu können, welchem Bereich ein bestimmter Wert zugeordnet ist.

Bemerkung

Sofern V als Sortier-Parameter verwendet wird, kann man beim Abspeichern der einzelnen Bereich z.B. den kleinsten Wert eines Intervalls weglassen, da dieser der Nachfolger (bezüglich V) des höchsten Werts des vorherigen Bereichs ist.

6 Neue Herangehensweisen

Neben den oben vorgestellten Methoden gibt es noch weitere Herangehensweisen zur Konstruktion von Histogrammen, die in [1] vorgestellt werden.

6.1 Kriterien

Analog zu Abschnitt 4.1 gibt es wiederum verschiedene Kriterien, die für neue Ansätze in Frage kommen:

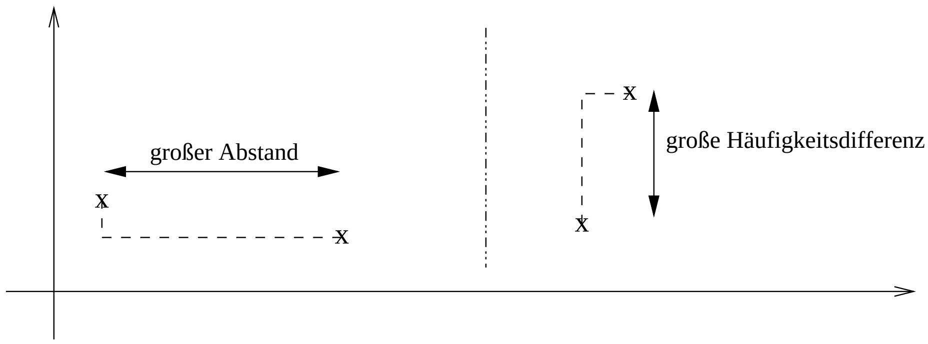
6.1.1 Bedingung

- „Maximale Differenz“: Die Bereichsgrenzen werden an den Stellen mit den größten Differenzen zwischen den bezüglich des Quell-Parameters sortierten Werten angeordnet. Diese Methode trägt der möglichen Forderung Rechnung, daß deutlich unterschiedliche Werte unterschiedlichen Histogramm-Bereichen zugeordnet werden sollen.
- „komprimiert“: Hierbei werden die n höchsten Werte in n einzelnen Bereichen (d.h. in solchen, die nur aus einem Element bestehen) gespeichert; der Rest wie bei „gleiche Summe“ behandelt.
  Dabei müssen die n Einzelwerte größer als der Quotient aus der Summe aller Werte durch die Anzahl der Bereiche sein (man bezieht sich also auf „überdurchschnittliche“ Werte), und die Werte müssen sich in den β Bereichen unterbringen lassen (andernfalls müssen doch wieder mehrere dieser Werte in einem Bereich gespeichert werden).
- „gleicher Abstand“: Hier werden die Tupel so angeordnet, daß daß die Werte in einem Histogramm-Bereich den gleichen Abstand s_i (vom Bereichsdurchschnitt) haben. Hierbei läßt sich auch bei einer sehr ungleichmäßigen Verteilung der Werte eine hohe Genauigkeit erzielen.

6.1.2 Parameter

In früheren Histogrammen wurde entweder der Abstand eines Tupels *oder* dessen Häufigkeit betrachtet, also nur ein Parameter. Da das Ziel von Histogrammen aber ist, die Datenverteilung T, also beides, zu approximieren, wäre es wünschenswert, beides gleichzeitig betrachten zu können.
Aufgrund dieser Überlegungen führt man das als „Fläche“ (A) bezeichnete Produkt des Abstandes und der Häufigkeit eines Tupels als neuen möglichen Parameter ein: $a_i = f_i \cdot s_i$.

Abbildung 4: Motivation der „Fläche“ als Parameter



6.2 Beispiel für Annahmen über Werteverteilungen

Seien die Bereichsbreiten jeweils gleich 100, $D = \{0, 1, 2, \ldots\}$,
insgesamt gebe es 200 Werte pro Bereich, davon jeweils 10 verschiedene.
Die gestellte Anfrage sei $10 \leq X \leq 25$.

Dann gibt es je nach Annahme über die Werteverteilung innerhalb eines Histogramm-Bereichs unterschiedliche Resultate bezüglich der Anzahl der gefunden Tupel:

- <u>gleicher Abstand:</u>
  Dann geht man von folgenden Werten innerhalb des Intervalls [1,100] aus:
  1, **12, 23**, 34, 45, 56, 67, 78, 89, 100.
  Jeder dieser Werte ist 20 mal vertreten, 2 Werte erfüllen die Anfragebedingung.
  Also beträgt das Resultat $20 \cdot 2 = 40$.

- <u>fortlaufende Werte:</u>
  Hier ist jeder Wert im Intervall vorhanden, also:
  1, 2, ..., 9, **10, ..., 25**, 26, ..., 100.

Jeder dieser Werte ist 2 mal vertreten, 16 Werte erfüllen die Anfragbedingung.
Also beträgt das Resultat $2 \cdot 16 = 32$.

- punktweise Betrachtung:
  Nun liegt nur ein Wert im betrachteten Intervall, nämlich:
  1.
  Dieser Werte ist 200 mal vertreten, aber erfüllt nicht die Anfragbedingung.
  Also beträgt das Resultat $200 \cdot 0 = 0$.

7 Neue Histogramme

Vorbemerkung: im folgenden geht man von den Annahmen des gleichen Abstands und der gleichen Häufigkeit der Werte pro Bereich aus.

7.1 Varianz-optimiert(V,F), (V,A), (A,A), (V,C)

Eine nicht gleichmäßige Datenverteilung bezüglich des Quell-Parameters wird bei Bereichseinteilung berücksichtigt. Der Histogramm-Typ „Varianz-optimiert(A,A)" sollte sowohl die Werte als auch deren Häufigkeit gut wiedergeben.
Bei der Verwendung von V als Sortier-Parameter ist außerdem kein Index erforderlich.

7.2 Maximale Differenz(V,F), (V,A)

Bei diesem Histogramm-Typ werden bezüglich des Quell-Parameters deutlich verschiedene Werte unterschiedlichen Bereichen zugeordnet (vgl. Abschnitt 6.1.1). Diese Histogramme können effizient konstruiert werden, indem man die die Differenzen der aufeinanderfolgenden Quell-Paramteter berechnet; die Bereichsgrenzen werden an den Stellen gesetzt, an denen die $\beta - 1$ größten Differenzen aufgetreten sind.

7.3 Komprimiert(V,F), (V,A)

Da die Werte mit den größten Häufigkeiten bzw. Flächen allein einzelnen Bereichen zugeordnet werden, ergibt sich eine hohe Genauigkeit auch bei einer unregelmäßigen Verteilung.

8 Techniken zur Erstellung von Histogrammen

Zur Erstellung von Histogrammen benötigt man unter anderem eine Berechnung von Bereichsgrenzen, Häufigkeiten f_i und c_i für jeden Wert, Anzahl verschiedener Werte in einem gegebenenen Intervall sowie Abstände der Werte voneinander, jeweils bezüglich verschiedener Kriterien. Dabei ist es im allgemeinen nicht erforderlich, alle genannten Daten zu berechnen; vielmehr kann man einen Teil der Daten aus bereits bekannten Ergebnissen ableiten, z.B. lassen sich die c_i aus den f_i berechnen, sofern diese alle bekannt sind. Da jedoch gleichzeitig die Prozessor- und I/O-Last sowie der Speicherbedarf möglichst gering gehalten werden sollen, kann man diese in der Praxis nicht alle durch jeweils komplettes Durchsuchen und der gesamten Relation und gewohnte Berechnungsverfahren erstellen.

8.1 Bereichsgrenzen (Quantilen)

Unter einem Quantil versteht man eine Grenze, die den entsprechenden Anteil von Tupeln vom Rest der Tupel trennt, z.B. sind 40% von Tupeln kleiner bezüglich der entsprechenden Sortierung als der entsprechende Wert des 0.4-Quantils, 60% sind dementsprechend größer.

Quantilen lassen sich z.B. durch den P^2-Algorithmus oder Stichproben-Verfahren berechnen. Da die Verwendung des P^2-Algorithmus im Vergleich zu Sampling zu teuer und zu ungenau ist, greift man in der Praxis auf Stichproben-Verfahren zurück. Hierbei bietet sich z.B. das sogenannte „reservoir sampling“ an, das die Möglichkeit bietet, auch auf Datenströmen vorher unbekannter Gesamtgröße arbeiten zu können, was z.B. von Vorteil ist, wenn weitere Tupel zur Relation hinzugefügt werden und die Histogramme entsprechend aktualisiert werden müssen.

8.1.1 Reservoir sampling

Dieser Algorithmus funktioniert wie folgt (dabei wird die Relation in einer festgelegten Reihenfolge durchlaufen):

1. Nehme eine Stichprobe von n Elementen und speichere sie ab.
2. Lasse zufällig viele Elemente aus.
3. Ersetzte ein zufällig ausgewähltes Element aus dem Speicher durch das nächste ausgewählte Tupel.
4. Wiederhole die Schritte 2 und 3, bis das Ende der Relation erreicht ist.

Dieses Verfahren spart Systemressourcen im Vergleich zu einem vollständigen Durchgehen der Relation; diesen Vorteil erkauft man aber durch eventuelle Ungenauigkeiten.

8.2 Weitere Parameter

Auch die anderen oben genannten Parameter lassen sich durch Sampling bestimmen; die meisten Daten lassen sich einfach von dem per Stichprobe festgestelltem Ergebnis auf die gesamte Relation übertragen.

9 Experimente

9.1 Erstellungszeiten

Hier werden jeweils die (V,A)-Formen der jeweiligen Histogramme betrachtet, aber alle anderen Formen liefern ähnliche Resultate.
Die Zeit zum Erstellen der jeweiligen Stichproben wurden hierbei nicht berücksichtigt.
Insgesamt hatte die betrachtete Relation 100.000 Tupel, davon 200 verschiedene; die Größe des reservoir sampling betrug 2.000 Tupel.

Histogramm-	benötigte Zeit:	
Typ:	160 b [/sec]	400 b[/sec]
Komprimiert	5.9	9.3
gleiche Summe:	6.2	10.9
maximale Differenz:	7.0	12.8
End-ausgreichted Varianz-optimiert:	7.2	10.9
Spline-basiert:	20.3	41.7
Varianz-optimiert:	42.9	67.0
Gleiche Tiefe (mit P^2):	4992	10524

9.2 Fehleranfälligkeit

Der Fehler E wird berechnet als $E = \frac{100}{N} \sum_{q \in Q} \frac{|S_q - S'_q|}{S_q}$, wobei S_q die wirkliche Größe des Anfrageresultats, S'_q die geschätzte Größe des Anfrageresultats und Q eine Menge von N Anfragen ist.

Die Größe der betrachteten Relationen betrug 100.000 - 500.000 Tupel, davon 200-1000 verschiedene; die Stichprobengröße umfaßte 2.000 Tupel, die verschieden gleich bzw. „schräg“ verteilt waren.
Für alle Histogrammarten wurde der gleiche Speicherplatz zur Verfügung gestellt.
Dann ergaben sich folgende Fehlerwerte:

Histogrammtyp:	Fehler [/%]
trivial	60.84
Gleiche Tiefe (P^2)	17.87
Varianz-optimiert(A,A)	15.28
Varianz-optimiert(V,C)	14.62
Gleiche Breite	14.01
Varianz-optimiert(F,F)	13.40
End-ausgerichtet Varianz-optimiert(A,A)	12.84
End-ausgerichtet Varianz-optimiert(F,F)	11.67
Gleiche Tiefe (genau)	10.92
Spline-basiert(V,C)	10.55
Komprimiert(V,A)	3.76
Komprimiert(V,F)	3.45
Maximale Differenz	3.26
Varianz-optimiert(V,F)	3.26
Maximale Differenz	0.77
Varianz-optimiert(V,A)	0.77

9.3 Größe der Stichproben

Geringfügige Änderungen der Größe der Stichproben sowie der Anzahl der Attributwerte haben keine signifikanten Auswirkungen auf die Genauigkeit, ebenso das zur Verfügungstellen von Speicherplatz über 160 Bytes.

9.4 Verteilungsannahmen

Die Annahme der gleichen Abstände führt zu etwas genaueren Resultaten als die Annahme der fortlaufende Werte oder die punktweise Betrachtung.

10 Fazit

Zusammenfassend läßt sich sagen, daß man auf jeden Fall V als Sortier-Parameter verwenden sollte, da in diesem Fall kein Index zur Zuordnung von Werten zu einem Histogramm-Bereich notwendig ist; außerdem reicht es dann aus, nur eine Bereichsgrenze abzuspeichern.

Das „Maximale Differenz(V,A)“-Histogramm erscheint in den meisten Situationen als beste Wahl, wenn man sowohl niedrige Erzeugungskosten als auch eine niedrige Fehlerrate benötigt.

11 Zusammenfassung

Histogramme werden zum effizienten Umgang mit großen Datenmengen benötigt. Dabei ordnet man die Tupel nach gemeinsamen Kriterien in Bereiche an. Es existieren verschiedene Arten von Histogrammen, die sich nach Klasse, Bedingung sowie Sortier- und Quell-Parameter in verschiedene Kategorien einteilen lassen. Zur effizienten Berechnung von Histogrammen benötigt man Stichproben und randomisierte Algorithmen. Experimente zeigen, daß ein Histogramm vom Typ „Maximale Differenz“ die beste Wahl bezüglich möglichst geringer Erzeugungskosten und Fehlerrate ist.

Literatur

[1] Viswa Poosala, Yannis E. Ioannidis, Peter J. Haas, Eugene J. Shekita: Improved Histograms for Selectivity Estimation of Range Predicates, ACM SIGMOD Conference, 1996

[2] Gregory Piatetsky-Shapiro, Charles Connell: Accurate Estimation of the Number of Tuples satisfying a Condition, ACM SIGMOD Conference, 1984