

Das Scatter/Gather- Verfahren

Seminarvortrag zum Studienprojekt
Invisible Web
an der Universität Duisburg-Essen

André Nurzenski
Duisburg, den 06.02.2004

Inhaltsverzeichnis

1	Einleitung	2
2	Das Prinzip von Scatter/Gather	3
2.1	Anforderungen	4
2.2	Probleme.....	4
3	Teilschritte von Scatter/Gather.....	5
3.1	Partitionierendes Clustering	5
3.1.1	Finden von k Startzentren	5
3.1.2	Zuweisen der Dokumente zu den Zentren.....	5
3.1.3	Verfeinern der erstellten Partition	5
3.2	Beschreiben der Gruppen/Cluster	5
4	Algorithmen	6
4.1	Definitionen.....	6
4.2	Algorithmen zum Finden von Startzentren	6
4.2.1	Buckshot.....	6
4.2.2	Fractionation.....	6
4.2.3	Bewertung	6
4.3	Assign-to-Nearest.....	6
4.4	Algorithmen zur Verfeinerung der Partition	7
4.4.1	Iterated Assign-to-Nearest.....	7
4.4.2	Split	7
4.4.3	Join	7
4.4.4	Bewertung	7
4.5	Cluster Digest	7
5	Beispiel.....	8
6	Fazit.....	10
	Literaturverzeichnis.....	11

1 Einleitung

Wir betrachten zunächst die klassische Vorgehensweise, in einer Dokumentenkollektion relevante Dokumente für ein bestimmtes Thema zu suchen. In eine Suchmaschine wird eine spezifische Suchanfrage (Query) eingegeben, mit deren Hilfe das System dann unter Verwendung verschiedenster Retrieval-Modelle relevante Dokumente auswählt und dem Anwender diese nach diversen Kriterien gewichtet zurückliefert.

Da eine Suchanfrage das Bedürfnis des Anwenders nach einer bestimmten Information darstellt, muss diese dementsprechend exakt formuliert werden. Nun kann es allerdings vorkommen, dass eben diese exakte Formulierung nicht oder nur schwer möglich ist, sei es durch Unkenntnis des benötigten Vokabulars oder durch das Bedürfnis, sich erstmal einen Überblick über die zur Verfügung stehenden Themengebiete einer Kollektion zu verschaffen. Außerdem können selbst gut formulierte Suchanfragen eine Vielzahl von Ergebnissen zurückliefern, die es dem Anwender sehr schwer machen sich einen geeigneten Überblick über die einzelnen Themen zu verschaffen. Sind diese Dokumente auch noch themenübergreifend verfasst, ist es fast unmöglich auf diesem Weg relevante Dokumente zu finden. Somit ist es leicht einzusehen, dass hier die klassische Suchanfrage nicht oder nur sehr langsam zum gewünschten Ergebnis führt. Aus diesem Grund betrachten wir nun einen anderen Ansatz, und zwar das „Browsen“ bzw. „Stöbern“ in einer Dokumentenkollektion. Beim „Browsen“ ist in erster Linie kein festes Ziel vorgegeben, sondern es geht vielmehr darum, mehr über die Dokumente der Kollektion zu erfahren und die so gewonnenen Informationen zur Formulierung einer korrekten Suchanfrage zu verwenden. Natürlich ist es auch möglich, durch das Browsen selbst interessante Dokumente zu finden und sich diese auf ihre Relevanz hin anzuschauen.

Das Scatter/Gather-Verfahren beschreibt nun eine solche Browsing-Komponente, die auf Cluster-Algorithmen basiert. Hierbei werden ähnliche Dokumente in Gruppen/Cluster aufgeteilt, deren Inhalt durch eine automatisch generierte Übersicht beschrieben wird. Der Anwender hat jetzt die Möglichkeit, eine oder mehrere Gruppen auszuwählen und diese erneut aufzuteilen. So wird die Sicht auf die Kollektion bei jedem Schritt genauer und Themenbereiche grenzen sich immer mehr voneinander ab.

Der Name des Verfahrens setzt sich aus den beiden englischen Wörtern *Scatter* und *Gather* zusammen, deren Bedeutung hier wie folgt zu interpretieren ist:

Scatter:

- streuen; zerstreuen
- Verteilen von Dokumenten in Gruppen/Cluster

Gather:

- sammeln; erfassen
- Auswahl einer Teilgruppe und erneute Verteilung um neue Gruppen/Cluster zu bilden

Im nachfolgenden kann der Begriff „Scatter/Gather“ auch mit „S/G“ abgekürzt werden.

2 Das Prinzip von Scatter/Gather

Zu Beginn der Browsing-Session werden alle Dokumente aus einer Kollektion durch Clustering in eine kleine Anzahl von Gruppen aufgeteilt, deren Inhalt durch kurze Zusammenfassungen beschrieben ist.

Aus diesen Gruppen wählt der Anwender, basierend auf den kurzen Beschreibungen, eine oder mehrere für ihn interessante Gruppen aus. Die ausgewählten Gruppen werden zusammengefasst und bilden somit eine Sub-Kollektion, aus der mittels Clustering erneut eine kleine Anzahl von Gruppen gebildet wird. Hieraus kann nun der Anwender erneut für ihn interessante Gruppen auswählen.

Dieses Verfahren kann nun beliebig oft wiederholt werden. Es ist leicht einzusehen, dass sich nach jedem dieser Schritte weniger Dokumente in den einzelnen Gruppen befinden und diese somit immer detaillierter werden. Idealerweise befinden sich am Ende der Browsing-Session nur noch themenspezifische Dokumente in den einzelnen Gruppen, die der Anwender sich als für ihn relevante Dokumente nun genauer ansehen kann.

Dieses Clustering-Verfahren liefert im Gegensatz zum hierarchischen, partitionierenden Clustering keine Baumstruktur mit genau einem Elternknoten pro Cluster, sondern einen gerichteten Grafen, bei dem die Dokumente in einem Knoten aus mehreren disjunkten Elternknoten stammen können.

Dies ist möglich, weil der Anwender mehrere Cluster auswählen kann, die zu einer Sub-Kollektion zusammengefasst und dann wieder in Gruppen unterteilt werden. Somit können die Dokumente in einer Gruppe aus mehreren, voneinander verschiedenen Clustern stammen. Die gesamte Struktur des Grafen hängt also von der Auswahl des Benutzers ab.

Im Folgenden wird nun eine solche Browsing-Session aus [1] beschrieben. Als Grundlage dient hier eine Dokumentenkollektion, die ca. 5000 vom *New York Times News Service* veröffentlichte Artikel vom August 1990 enthält. Diese Session ist durch Abbildung 1 illustriert. Um die Abbildung zu vereinfachen, wurden hier manuell einzelne Wörter, basierend auf der kompletten Cluster-Beschreibung, den einzelnen Gruppen zugewiesen. Die gesamte Session mit allen Ausgaben und Beschreibungen ist in Kapitel 5 zu finden.

Angenommen, der Anwender möchte mehr über die Ereignisse in diesem Monat erfahren, wird aber durch folgende Problematik von der Verwendung konventioneller Suchtechniken abgehalten:

- Die gesuchte Information kann nicht immer in ein einzelnes Thema gefasst werden.
- Selbst wenn ein Thema bekannt wäre, kann es sein, dass dem Anwender die Wörter zur Beschreibung des Themas nicht bekannt sind.
- Die Wörter, die ein Thema beschreiben, müssen nicht zwingend die Wörter sein, die in diesem Thema vorkommen, und tauchen deshalb vielleicht gar nicht im Artikel auf.
- Für viele Wörter existieren Synonyme, die stattdessen verwendet werden können.

Beim Scatter/Gather-Verfahren muss der Anwender nun keine Terme angeben, um seine Suche zu beginnen, sondern kann direkt mit dem „Stöbern“ in der gesamten Kollektion beginnen. Dieses „Stöbern“ wird in Abbildung 1 aus [1] veranschaulicht.

Nach Durchführung des ersten Clusterings werden uns 8 Gruppen präsentiert. Wir wollen nun unser Augenmerk auf internationale Ereignisse legen und wählen deshalb die Gruppen *Iraq*, *Oil* und *Germany*. Diese 3 Gruppen werden nun zusammengefasst und neu geclustert. Wir erhalten also 8 neuen Gruppen, dieses Mal mit dem Schwerpunkt *internationale Ereignisse*. Da uns die wichtigen Ereignisse dieses Monats bereits hinreichend bekannt sind, wollen wir mehr über andere Ereignisse in diesem Monat herausfinden. Also wählen wir nun die Gruppen *Pakistan* und *Africa*. Nun erhalten wir erneut 8 Gruppen, die sich mit sehr speziellen internationalen Ereignissen beschäftigen, z.B. einer Geiselnahme in Trinidad oder einem Putsch in Pakistan. Hier könnten wir nun einzelne Dokumente aus den Gruppen auswählen, um mehr über diese Themen zu erfahren

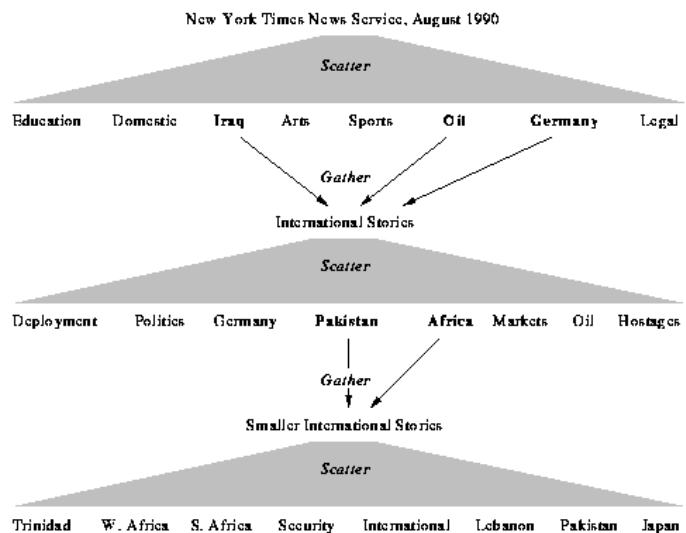


Abbildung 1: Veranschaulichung des Scatter/Gather-Verfahrens

Aus dem hier erläuterten Prinzip ergeben sich mehrere spezielle Anforderungen an das Verfahren bzw. entstehen einige Probleme.

2.1 Anforderungen

Scatter/Gather benötigt zwei grundlegende Verfahren, um ein adäquates Arbeiten zu erlauben:

- Da Clustering einen essentiellen Vorgang in diesem Verfahren darstellt, benötigen wir einen Algorithmus, der in der Lage ist, eine große Anzahl von Dokumenten in einer akzeptablen Zeit zu gruppieren, und somit interaktives Arbeiten ermöglicht.
- Ein weiterer essentieller Vorgang ist die Auswahl von relevanten Gruppen durch den Benutzer. Somit benötigen wir eine Methode, die automatisch eine Beschreibung der Gruppen generiert, und unter deren Verwendung ausreichend auf das Thema/den Inhalt der Dokumente in einer Gruppe geschlossen werden kann.

2.2 Probleme

Bei der Verwendung von Scatter/Gather ergeben sich aber auch einige Probleme, die nicht ohne weiteres lösbar sind bzw. nur von der Ausgangskollektion abhängig sind:

- Wie viele Gruppen/Cluster sollen am Anfang gebildet werden, um eine gute Übersicht über die darunter liegenden Themenbereiche zu erhalten?
- Geben die automatisch generierten Clusterbeschreibungen hinreichend Aufschluss über den Inhalt der Dokumente im Cluster?
- Wie viel Wartezeit zwischen den einzelnen Iterationsschritten kann dem Benutzer zugemutet werden, um noch ein interaktives Arbeiten zu ermöglichen (5-30 Sek.)?

3 Teilschritte von Scatter/Gather

In diesem Kapitel wird nun eine konkrete Umsetzung von Scatter/Gather aus [1] beschrieben und die einzelnen Teilschritte einer S/G-Iteration näher erläutert. Diese Teilschritte werden immer zwischen den Benutzerinteraktionen (nämlich dem Auswählen interessanter Gruppen durch den Benutzer und dem Präsentieren dieser neu geordneten Gruppen durch das System) durchgeführt. Die verschiedenen Algorithmen, die den Teilschritten zu Grunde liegen, werden in Kapitel 4 (Algorithmen) genauer beschrieben.

3.1 Partitionierendes Clustering

In diesem Teilschritt werden für eine Menge von Dokumenten Gruppen bestimmt, und die Dokumente nachher diesen Gruppen zugeordnet. Wir erhalten hier also eine Menge von k disjunkten Dokumentengruppen. Dieses Verfahren wird *Partitionierendes Clustering* genannt und besteht aus drei Einzelschritten, die wir nun im nachfolgenden genauer betrachten werden.

3.1.1 Finden von k Startzentren

Um die Startzentren zu definieren, können zwei Algorithmen verwendet werden, nämlich *Buckshot* und *Fractionation*. Sie können als „grobe“ Cluster-Algorithmen betrachtet werden, deren Ausgabe allerdings nur zur Definition der Clusterzentren verwendet wird. Beide Algorithmen benötigen eine *Cluster Subroutine*, die das eigentliche Clustern durchführt und hierbei gute Ergebnisse erzielen sollte, aber eventuell sehr langsam arbeitet. Diese Subroutine wird von beiden Algorithmen auf kleinen Mengen (Stichprobe) der gesamten Kollektion angewendet.

3.1.2 Zuweisen der Dokumente zu den Zentren

Hierbei wird ein sehr einfaches Verfahren angewendet, nämlich die Zuweisung jedes Dokuments zu dem ihm am nächst gelegenen Zentrum einer Gruppe. Dieser Algorithmus wird im weitem *Assign-to-Nearest* genannt und in Kapitel 4 näher erläutert.

3.1.3 Verfeinern der erstellten Partition

Bei diesem Schritt stehen uns wieder zwei Verfahren zur Verfügung. Die Entscheidung, welches der beiden Verfahren hier angewendet werden soll, ist abhängig von der zur Verfügung stehenden Zeit und der gewünschten Genauigkeit.

Die schnellste, aber in ihren Möglichkeiten begrenzte Methode ist die iterierte Anwendung von *Assign-to-Nearest*, wobei die Anzahl der Iterationen sehr klein sein sollte (2 Iterationen haben in der Praxis gute Ergebnisse geliefert).

Eine genauere, aber langsamere Methode ist die Verwendung der Prozeduren *Split* und *Join*. Dabei werden schlecht definierte Gruppen in zwei inhaltlich gut getrennte Teilgruppen aufgeteilt und zu ähnliche Gruppen werden zu einer Gruppe zusammengefasst. Auch dieses Verfahren kann mehrfach durchgeführt werden.

3.2 Beschreiben der Gruppen/Cluster

Nachdem die einzelnen Dokumente zu Gruppen zusammengefasst wurden, muss noch eine Beschreibung für die einzelnen Gruppen generiert werden. Aus dieser Beschreibung muss für den Anwender der Inhalt der Gruppen, also die Themen die die Dokumente jeder Gruppe behandeln, klar ersichtlich sein.

Dies wird nun durch das sog. *Cluster Digest* ermöglicht, ein Verfahren das die häufigsten, am höchsten gewichteten Terme einer Gruppe als Beschreibung dem Anwender ausgibt. Wie so eine Beschreibung aussieht, wird in Kapitel 5 (Beispiel) deutlich.

4 Algorithmen

Nachdem wir uns im vorhergehenden Kapitel mit den einzelnen Teilschritten im Scatter/Gather-Verfahren beschäftigt haben, liegt der Schwerpunkt nun auf der Beschreibung der für jeden Teilschritt verfügbaren Algorithmen. Hier wird nun erläutert, wie die Algorithmen im Detail arbeiten und welche Vor- und Nachteile sich bei ihrer Verwendung ergeben können. Dazu sind allerdings erst einmal ein paar allgemeine Definitionen nötig.

4.1 Definitionen

C = Kollektion von Dokumenten

V = Menge von eindeutigen Wörtern in C

Γ = Gruppe/Cluster von Dokumenten

P = Partition (Menge von Clustern/Gruppen)

α = Individuelles Dokument/Einzernes Dokument

n = Gesamtanzahl der Dokumente in einer Kollektion

k = gewünschte Anzahl von Gruppen/Clustern beim Scatter-Schritt

w = Anzahl häufigster eindeutiger/themenspezifischer Wörter in einem Cluster

4.2 Algorithmen zum Finden von Startzentren

4.2.1 Buckshot

Beim Buckshot-Algorithmus wird zuerst eine zufällige Stichprobe vom Umfang \sqrt{kn} aus der Dokumentenkollektion ausgewählt, auf die nun die *Cluster Subroutine* angewendet wird. Als Rückgabe liefert Buckshot die Zentren der gefundenen Cluster. Die Laufzeit des Algorithmus ist $O(kn)$.

Da der Algorithmus mit zufällig erzeugten Stichproben arbeitet, ist er nicht deterministisch, was zu verschiedenen Partitionen bei der gleichen Dokumentenkollektion führen kann. In der Praxis hat sich allerdings gezeigt, dass diese verschiedenen Partitionen von vergleichbarer Qualität sind.

4.2.2 Fractionation

Bei Fractionation wird zunächst die Kollektion C in N/m „Behälter“ mit fester Größe $m > k$ aufgeteilt. Die *Cluster Subroutine* wird dann auf jeden „Behälter“ einzeln angewendet, um individuelle Dokumente zu Gruppen zusammenzufassen. Diese Gruppen werden nun wie einzelne Dokumente behandelt und der gesamte Prozess wird wiederholt. Diese Iteration terminiert, wenn nur noch k Gruppen übrig bleiben. Bei diesem Vorgang wird also quasi ein Baum von unten nach oben aufgebaut, bei dem die Blätter einzelne Dokumente sind und der fertig ist, wenn nur noch k Wurzeln übrig sind. Die Laufzeit beträgt $O(mn)$.

4.2.3 Bewertung

Buckshot ist nicht so genau wie *Fractionation*, arbeitet aber schneller und ist deshalb besser für die häufig auftretenden Iterationen beim Scatter/Gather-Verfahren geeignet. Die höhere Genauigkeit von *Fractionation* kann dazu verwendet werden, eine Initialpartition aus einer Dokumentenkollektion zu generieren, die dann dem Benutzer beim Starten des Programms präsentiert wird. Für die weiteren Iterationen ist wieder *Buckshot* zu empfehlen.

4.3 Assign-to-Nearest

Nachdem die Zentren der Gruppen gefunden wurden, muss nun jedes Dokument einem Zentrum zugewiesen werden. Dazu wird ein Ähnlichkeitsmaß zwischen dem jeweiligen

Dokument und allen Zentren berechnet. Das Dokument wird dann dem Zentrum zugewiesen, bei dem das Ähnlichkeitsmaß maximiert wird. Die Kosten hierfür sind proportional zu kn .

4.4 Algorithmen zur Verfeinerung der Partition

4.4.1 Iterated Assign-to-Nearest

Dieses Verfahren ist die iterierte Anwendung des unter Punkt 4.3 beschriebenen *Assign-to-Nearest* Verfahrens. Hierzu werden aus einer gegebenen Menge von Clustern neue Clusterzentren berechnet, um dann jedes Dokument einem der neuen Zentren zuzuweisen. Dieser Prozess kann beliebig oft wiederholt werden, erzielt allerdings während der ersten Schritte die größten Erfolge.

4.4.2 Split

Beim *Split*-Verfahren wird jeder Cluster I in einer Partition in zwei neue Cluster aufgeteilt. Dies wird durch Verwendung von *Buckshot* erreicht, indem als Eingabe $C = I$ und $k = 2$ festgelegt werden.

Bei einer modifizierten Variante dieses Verfahrens werden nur Gruppen aufgeteilt, die nach einem bestimmten Kriterium schlecht bewertet werden. Ein solches Kriterium ist z.B. die „Selbstähnlichkeit“ eines Clusters. Wird nun ein bestimmter, vorher festgelegter Toleranzwert unterschritten, wird die betreffende Gruppe unter Verwendung von *Buckshot* in zwei Gruppen aufgeteilt.

4.4.3 Join

Unter Verwendung des vorhergehenden Verfahrens können wir nun Cluster aufteilen, deren Inhalte aus unserer Sicht zu verschieden voneinander sind. Nun ist es aber auch sinnvoll, Cluster, die sich sehr ähnlich sind, zu einem Cluster zu verbinden. Da nach Definition alle Dokumente in der Partition voneinander verschieden sind, können Cluster niemals „typische“ Dokumente gemeinsam enthalten. Allerdings kann die Liste der „typischen“ Wörter in Clustern sich sehr stark überlappen.

Es wird also zunächst überprüft, wie viele themenspezifische Wörter zweier Cluster übereinstimmen. Wird nun eine vorher festgelegte Anzahl p mit $0 < p \leq w$ erreicht bzw. überschritten, so werden beide Cluster vereinigt.

Die Laufzeit dieses Algorithmus beträgt $O(kn)$.

4.4.4 Bewertung

Wie schon bei der Bewertung unter Punkt 4.2.3, findet auch hier eine Abwägung zwischen Qualität und Schnelligkeit der einzelnen Verfahren statt. Die iterierte Anwendung von *Assign-to-Nearest* ist schnell, liefert akzeptable Ergebnisse und ist somit besser für das interaktive Arbeiten mit Scatter/Gather geeignet.

Spilt und *Join* arbeiten genauer, benötigen aber auch mehr Zeit und sind deshalb bei großen Kollektionen nicht zu empfehlen.

4.5 Cluster Digest

Das sog. *Cluster Digest* liefert, wie schon unter 3.2 angesprochen, die Beschreibung eines Clusters/einer Gruppe. Dazu werden anstelle der Dokumente nahe am Zentrum die häufigsten „typischen“ Wörter eines Clusters bestimmt und dem Benutzer präsentiert. Das *Cluster Digest* eines Clusters kann in Laufzeit $O(|I| + |V|)$ berechnet werden.

5 Beispiel

Im Nachfolgenden ist die komplette Ausgabe einer Scatter/Gather-Session abgebildet. Dieses Beispiel wurde unverändert aus [1] übernommen. Die Dokumentensammlung besteht aus Artikeln, die vom *New York Times News Service* während des Monats August im Jahre 1990 veröffentlicht wurden. Die Kollektion beinhaltet ungefähr 30 Megabyte ASCII-Text in ca. 5000 Artikeln. Manche Artikel wiederholen sich, um neue Entwicklungen in den Geschehnissen aufzuzeigen.

Unser Ziel ist es nun, mehr über internationale politische Ereignisse im Verlauf dieses Monats zu erfahren. Zur Erstellung der Initial-Partition wurde der *Buckshot*-Clustering-Algorithmus verwendet (Abbildung 2). Für diesen Schritt wird allerdings *Fractionation* empfohlen, falls genügend Zeit zur Verfügung steht.

Jeder Cluster wird durch zwei Zeilen beschrieben, die durch seinen *Cluster Digest* dargestellt werden. Die erste Zeile enthält die Nummer des Clusters, die Anzahl der Dokumente im Cluster und Titel von Dokumenten nahe dem Zentrum. Die zweite Zeile enthält häufige Wörter im Cluster.

Es werden die Cluster 2 (Iraks Invasion in Kuwait), 5 (Märkte, inklusive Öl) und 6 (Meldungen über Deutschland und Möglicherweise andere internationale Ereignisse) ausgewählt, da diese anscheinend für uns interessante Artikel enthalten. Diese Cluster werden zusammengefügt, neu geclustert und ein neuer *Cluster Digest* wird angezeigt (Abbildung 3).

Im nächsten Schritt (Abbildung 4) werden nun die Cluster 3 (Meldungen über Ereignisse im Bezug zu Pakistan) und 4 (Meldungen zu Ereignissen im Bezug zu Afrika) ausgewählt. Somit wurden nun spezielle Ereignisse angezeigt, wie zum Beispiel Artikel über eine Geiselnahme in Trinidad, Krieg in Liberia, Polizeiaktionen in Süd-Afrika und so weiter.

Man erhält eine detailliertere Sicht auf die Ereignisse in Liberia, wenn man sich die Titel der Artikel im entsprechenden Cluster anschaut (Abbildung 5).

```
> (time (setq first (outline (all-docs tdb))))
cluster 4970 times
global cluster 199 items...sizes: 18 24 53 5 25 47 13 14
move to nearest...sizes: 517 1293 835 86 677 1020 273 269
move to nearest...sizes: 287 1731 749 275 481 844 310 293
0 (287) CRITICS URGE NEW METHODS; PROGRAMS FOR PARENTS THE; TEACHING SUBJECTS T
school, year, student, child, university, state, program, percent, study, educ
1 (1731) FEDERAL WORK PROGRAMS HE; RESORT TAKES STEPS TO PR; AMERICANS CUT BACK
year, state, york, city, million, day, service, company, week, official, house
2 (749) PENTAGON SAYS 60,000 IRA, BUSH ``DRAWS A LINE`` IN; BUSH SAYS FOREIGNER
iraq, iraqi, kuwait, american, state, unite, saudi, official, military, presid
3 (275) TRILLIN'S MANY HATS; NEW MUSICAL FROM THE CRE; AFTER NASTY TEEN-AGERS I
film, year, music, play, company, movie, art, angeles, york, american, directo
4 (481) TWISTS AND TURNS MAY MEA; SAX LOOKING FOR RELIEF I; PAINTING THE DODGER
game, year, play, team, season, win, player, day, league, hit, right, coach, l
5 (844) CRISIS PUSHES OIL PRICES; WHY MAJOR PANIC OVER A M; OIL PRICES RISE AS
price, oil, percent, market, company, year, million, stock, day, rate, week, s
6 (310) LEADERS OF TWO GERMANYS ; REPRESENTATIVES OF TWO G; SECURITY COUNCIL RE
government, year, state, party, political, country, official, leader, presiden
7 (293) U.S. APPEALS ORDER FREEI; DID JUDGE MOVE TOO HASTI; MAYOR BARRY CONVICT
case, court, charge, year, judge, lawyer, attorney, trial, jury, federal, dist
real time 131258 msec
```

Abbildung 2: Initial-Partition

```

> (time (setq second (outline first 2 5 6)))
cluster 1903 times
global cluster 123 items...sizes: 51 8 5 5 4 7 28 15
move to nearest...sizes: 730 67 65 62 56 99 714 110
move to nearest...sizes: 650 66 57 117 59 242 586 126
0 (650) PENTAGON SAYS 60,000 IRA; BUSH SAYS FOREIGNERS DET; BUSH ``DRAWS A LINE
  iraq, iraqi, american, kuwait, state, unite, military, official, president, sa
1 (66) LEGISLATIVE LEADERS BACK; THE PROBLEM WITH AN EARL; ROAD STILL TOUGH FOR
  party, state, election, year, political, candidate, vote, campaign, democratic
2 (57) IN PUSH FOR UNIFICATION,; IN PUSH FOR UNIFICATION,; LEADERS OF TWO GERMA
  german, east, germany, west, year, government, soviet, union, state, unificati
3 (117) BHUTTO GOVERNMENT DISMIS; IN FRACTIOUS PAKISTAN, G; PAKISTANIS FEEL LET
  government, minister, year, party, political, military, country, official, sta
4 (59) DEATH TOLL EXCEEDS 500 I; DE KLERK, MANDELA HOLD U; NEGOTIATIONS TO SETT
  african, government, south, leader, police, national, fight, group, official,
5 (242) WEST GERMANS TO BUY FIRE; FIRST EXECUTIVE CORP. EA; FARM BANK, MERRILL
  company, million, percent, share, year, corp, stock, market, sell, price, pres
6 (586) OIL PRICES RISE AS STOCK; MIDEAST CRISIS PUSHES OI; WHY MAYOR PANIC OVE
  oil, price, percent, market, year, company, day, stock ,million, rate, week, f
7 (126) IRAQ GRANTS 237 FOREIGN ; WOMAN TELLS OF 12 DAYS I; CONCERN HEIGHTENS F
  kuwait, iraqi, american, iraq, saudi, day, year, invasion, country, state, ara
real time 54184 msec

```

Abbildung 3: Zweiter Scatter-Schritt

```

> (time (setq third (outline second 3 4)))
cluster 176 times
global cluster 37 items...sizes: 1 4 12 1 5 3 8 3
move to nearest...sizes: 4 16 44 1 23 7 71 10
move to nearest...sizes: 5 16 28 1 51 7 55 13
0 (5) MUSLIM MILITANTS LAY DOW; MUSLIM MILITANTS LAY DOW; DRAMA IS OVER BUT BOO
  government, trinidad, minister, parliament, wednesday, bakr, hostage, robinson
1 (16) NEGOTIATIONS TO SETTLE L; NEGOTIATIONS TO SETTLE L; WEST AFRICAN FORCE S
  rebel, african, taylor, west, liberia, troop, group, liberian, leader, officia
2 (28) DEATH TOLL EXCEEDS 500 I; DE KLERK, MANDELA HOLD U; COMPETING FACTIONS T
  south, police, african, black, mandela, africa, congress, anc, political, gove
3 (1) SHIFT IN U.S. COMPUTER S;
  security, agency, computer, technology, national, center, communication, milit
4 (51) SECURITY COUNCIL REACHES; @SECURITY COUNCIL REACHE; NEW U.S. POLICY IS W
  government, year, state, official, army, country, group, guerrilla, war, natio
5 (7) CLASHES BETWEEN RIVAL SH; MUSLIM FRATIONS BATTLE I; BOMBINGS IN SOUTHERN
  lebanon, muslim, christian, al, party, kill, god, lebanese, aoun, beirut, amal
6 (55) BHUTTO GOVERNMENT DISMIS; MS. BHUTTO CALLS HER OUS; MS. BHUTTO CALLS HER
  government, minister, party, political, military, prime, pakistan, president,
7 (13) SHEVARDNADZE TO VISIT TO; 45 YEARS AFTER WAR'S END; JAPAN'S ROLE IN WORL
  japan, soviet, war, korean, japanese, year, tokyo, government, south, korea, w
real time 11140 msec

```

Abbildung 4: Dritter Scatter-Schritt

```

> (print-titles (nth 1 third))
3720 REBEL LEADER SEIZES ABOUT A DOZEN FOREIGNERS
4804 WEST AFRICAN FORCE SENT TO LIBERIA AS TALKS REMAIN DEADLOCKED
4778 WAR THREATENS TO WIDEN AS NEIGHBORING COUNTRIES TAKE SIDES
3719 REBEL LEADER AGREES TO HOLD CEASE-FIRE TALKS
3409 OUSTER OF LIBERIAN PRESIDENT NOW SEEMS INEVITABLE
3114 NEGOTIATIONS TO SETTLE LIBERIAN WAR END IN FAILURE
3113 NEGOTIATIONS TO SETTLE LIBERIAN WAR END IN FAILURE
2785 LIBERIANS IN U.S. CRITICAL OF ADMINISTRATION POLICY
2784 LIBERIANS IN U.S. CRITICAL OF ADMINISTRATION POLICY
2783 LIBERIAN REBEL LEADER CHARLES TAYLOR HURT EN ROUTE TO CEASE-FIRE
2782 LIBERIA LEADER, REJECTING TRUCE OFFER, WON'T QUIT
1801 FIVE WEST AFRICAN NATIONS MOVING TROOPS TOWARD LIBERIA
1685 FACES OF DEATH IN LIBERIA
1684 FACES OF DEATH IN LIBERIA
248 OUSTER OF LIBERIAN PRESIDENT NOW SEEMS INEVITABLE

```

Abbildung 5: Titel der Artikel in Gruppe 1 aus Abbildung 4

6 Fazit

Scatter/Gather zeigt, dass Dokumenten-Clustering ein effektives Werkzeug für den Zugriff auf Informationen sein kann. Dies ist unter anderem der Fall, wenn es nicht möglich ist eine genaue Suchanfrage zu formulieren, oder man sich einen Überblick über eine große Sammlung von Dokumenten verschaffen möchte. In diesen Fällen ist das „Browsen“ in der Kollektion effizienter als das direkte „Suchen“ nach relevanten Dokumenten. Durch die Sicht auf die verfügbaren Themenbereiche, vergleichbar mit einem Inhaltsverzeichnis, ist mit jedem Iterationsschritt ein genauerer und tieferer Einblick in die darunter liegenden Themen gegeben. Aufgrund dieser Sichtweise ist eine einfache und intuitive Bedienung durch den Benutzer möglich.

Um Scatter/Gather effizient nutzen zu können, müssen schnelle Cluster-Algorithmen verwendet werden. Das Clustering kann schnell durchgeführt werden, wenn man mit kleinen Gruppen anstelle der gesamten Kollektion arbeitet.

Bei sehr großen Kollektionen kann aber selbst das von *Buckshot* und *Fractionation* in linearer Zeit durchgeführte Clustering zu langsam sein. Aus diesem Grund müssen noch andere Ansätze und Methoden betrachtet werden, die auch bei diesen Fällen ein effizientes und genaues Clustering ermöglichen.

Literaturverzeichnis

- [1] **Titel:** Scatter/Gather: A cluster-based Approach to Browsing Large Document Collection
Autoren: Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey
Verlag: ACM
In: Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seiten 318-329)
Jahr: 1992
Web: <http://citeseer.nj.nec.com/cutting92scattergather.html>
- [2] **Titel:** About Scatter/Gather
Veröffentlicht von: Xerox PARC (Palo Alto Research Center)
Jahr: 1997
Web: <http://www2.parc.com/istl/projects/ia/sg-overview.html>