Das Bayes Locality sensitive Hashing (Bayes-LSH) ist ein Algorithmus, der die Ähnlichkeit schätzt und führt Kandidatenschnitt durch. Bei Locality sensitive Hashing geht es um die Findung einer Familie von Hash Funktionen unter der Bedingung, dass für eine zufällige Hash-Funktion aus dieser Familie das Zerhacken von zwei Objekte mit hoher Ähnlichkeit aus dieser Familie erfolgt sehr wahrscheinlich mit dem gleichen Eimer. Der BayesLSH-Algorithmus erfolgt in zwei Phasen, nämlich die Kandidatengenerierung und Kandidat Überprüfung[SAP12].

Bei der Kandidatengenerierung werden l Unterschriften für jedes Objekt im Datensatz gebildet. Die Signaturen dienen als eine Dateienverknüpfung von k Hashes. Es erfolgt die Generierung von Kandidatenpaar, für alle Objektpaare, die mindestens eine der Signatur gemeinsam haben. Der effiziente Abruf von jedem signaturteilenden Objektpaar ist mit einer Hashtabelle möglich.

Bei der Ähnlichkeitsschätzung für LSH ist es schwierig die Anzahl der Hashes abzustimmen und das potenzial für eine frühzeitige Beschneidung von Kandidatenpaaren wird nicht ausgenutzt.

Trotz die Nachteile der Ähnlichkeitsschätzung für LSH, es ist möglich mit BayesLSH Kandidatenschnitt und Ähnlichkeitsschätzung durchzuführen. Bei der Bayessche Inferenz wird erstmal die Prior Verteilung über die Parameter bestimmt und danach die Posterior Verteilung über den Parameter wird berechnet unter der Bedingung, dass Daten beobachtet wurden. [SAP12] besagt, dass Charikar definiert lokalisierungssensitives Hashing-Schema als eine Verteilung auf eine Familie von Hash-Funktionen F, die auf einer Sammlung von Objekten arbeitet, so dass für zwei beliebige Objekte x, y, die folgende Wahrscheinlichkeit besteht:

[h(x) = h(y)] = sim (x, y) (1)

Für ein Paar (x, y) wird eine Aussage gemacht, dass es eine Übereinstimmung gibt bei m von den ersten n Hashes für dieses Paar. Es wird die Bezeichnung M(m,n) zu diesem Ereignis gegeben. Aus der Gleichung 1, die Binomialverteilung mit n Versuchen, wo ein eigenständiger Erfolg besteht, liefert die bedingte Wahrscheinlichkeit des Ereignisses M (m, n) bei der Ähnlichkeit S (Zufallsvariable). m und n sind in diesem Fall keine Variablen, sondern daten:

(2)

Wichtig zu finden ist die Wahrscheinlichkeitsverteilung für die Ähnlichkeit S, unter der Annahme, dass es eine Abstimmung für m von n Hashes gibt. Eine Schreibweise der Posterior Verteilung für S kann mithilfe des Satzes von Bayes wie folgt gegeben sein:

p(S |M(m, n)) = = =

Die Posterior Verteilung für jeden Wert von n und m ist erhalten durch den Zusatz von Ausdrücke für p(M(m, n) | S) aus der Gleichung 2 und eine geeignete Prior Verteilung p(S). Die Berechnung folgender Größen kann in Bezug auf der Posterior Verteilung:

1. Die Wahrscheinlichkeit, dass die Ähnlichkeit großer ist als den Schwellenwert t für m zustimmende Überstimmungen nach dem Vergleich von n Hashwerten, lautet:

Pr [S ≥ t |M (m, n)] = (3)

1. Die maximale a posteriori Schätzung für die Ähnlichkeit für m zustimmende Überstimmungen nach dem Vergleich von n Hashwerten, lautet:

= p (s |M (m, n)) (4)

1. Unter der Annahme, dass nachdem n Hashes verglichen wurden m Überstimmungen zustimmen und die geschätzte Ähnlichkeit ist, die Konzentrationswahrscheinlichkeit von , bzw. die Wahrscheinlichkeit, dass diese Schätzung ist enthalten in δ von der wahren Ähnlichkeit lautet:

Pr [|S − , | < δ |M (m, n)] = Pr [, − δ < S < , + δ |M (m, n)] (5)

= (6)

Mit der Annahme, dass die Durchführung der drei vorher genannten Arten von Inferenzen möglich ist, [SAP12] entwirft 2 Algorithmen: Das BayesLSH und das BayesLSH-Lite.

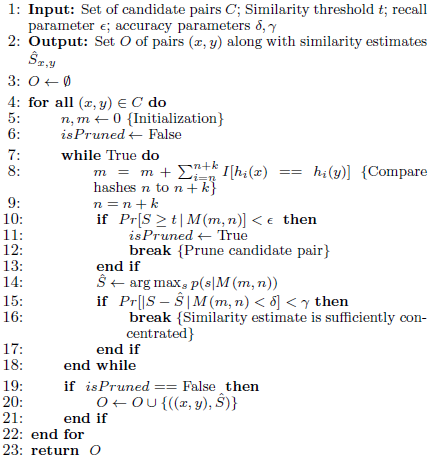


Abbildung 1: Der BayesLSH Algorithmus

Bei dem BayesLSH Algorithmus (Siehe Abbildung 1) erfolgt den Vergleich von den jeweiligen Hashes (Zeile 8, Angabe der Anzahl an Hashes, die zur gleichen Zeit verglichen werden durch den Parameter k) für ein Paar (x, y) bis zum Geschehen von einem diesen beiden Ereignisse. Es bestehen zwei Möglichkeiten, nämlich die erste bei der der Abschnitt des Kandidatenpaares erfolgt wegen der ziemlich geringen Wahrscheinlichkeit, dass das Paar echt positiv ist (Zeilen 10,11, und 12). Die Gleichung 3 wird angewendet zur Berechnung dieser Wahrscheinlichkeit. Bei der anderen Möglichkeit erfolgt den Schnitt des Kandidatenpaares nicht und die Hashes werden weiter verglichen bis zur genügenden Konzentration der Ähnlichkeitsschätzung entsprechend der Genauigkeitsanforderung (Zeilen 15 und 16). Es erfolgt zuletzt die Einfügung von jedem derartigen Paar zusammen mit der Ähnlichkeitsschätzung (Zeilen 19 und 20).

Der zweite Algorithmus BayesLSH-Lite ist eine einfachere Variante vom BayesLSH Algorithmus, die eine genaue Berechnung der Ähnlichkeiten ermöglicht. Die Abbildung 2 präsentiert dieser Algorithmus. Die Parameter δ und γ sind nicht erforderlich wegen dem genauer Aspekt der Ähnlichkeitsberechnungen. Dieser Algorithmus kann schneller sein als BayesLSH für Datensätze mit billige exakte Ähnlichkeitsberechnungen.

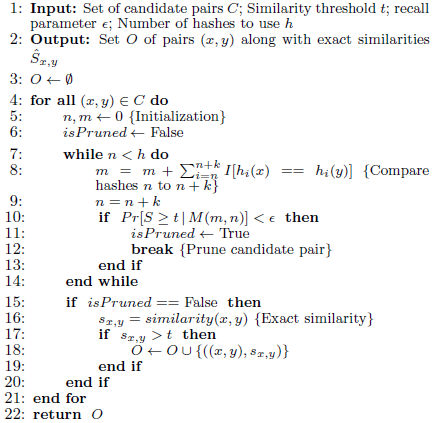


Abbildung 2:Der BayesLSH-Lite Algorithmus

Die Spezifikation von Aspekte wie die LSH-Familie von Hash-Funktionen, die Wahl der Prior Wahrscheinlichkeit und wie die lenkbare Inferenz durchgeführt sein kann, ermöglichen die konkrete Instanziierung von BayesLSH Algorithmus kann sowohl für Ähnlichkeitsmaße wie die Jaccard Ähnlichkeit und die Cosinus Ähnlichkeit Instanziierung.

Jaccard Ähnlichkeit: **die LSH Familie** entspricht eine Familie von unabhängigen MinHash Permutationen auf einem Universum, aus dem die Zeichnung der Sammlung von Mengen herausgenommen wird. Die Rückgabe von minimalem Element des Eingabe-Sets bei der Permutation von Elemente der Menge wie durch die Hash-Funktion spezifiziert. Die Familie von Hash-Funktionen gibt deswegen eine integer darstellende Zahl vom ganzen minimalen Element der permutierten Menge aus.

**Wahl des Priors**: Die Wahl von Prior aus einer zur Wahrscheinlichkeitsverteilung konjugierte Familie von Distributionen und zur Erhaltung folgerichtiger Schluss, sodass das die Posterior Verteilung sowie der Prior Teil der gleichen Verteilungsfamilie sind, ist gängige Praxis [SAP12]. Die Wahrscheinlichkeit wird durch die Binomialverteilung gegeben. Die Betaverteilung, Konjugiert für das Binom, besitzt zwei Parameter α> 0, β> 0 und wird auf der Domäne (0,1) spezifiziert. Der definierte pdf für Beta (α, β) ist der folgende:

p(s) =

B (α, β) repräsentiert die Beta Funktion oder eine Normalisierung konstante zur Sicherstellung, dass die Integration der gesamten Verteilung zu 1 erfolgt in diesem Fall. Für die Wahl von Parameter können folgende Werte gegeben sein: α = 1 und β = 1. Das Lernen des Fakts, dass α, β am besten geeignet sind für ein zufälliges Ähnlichkeitsbeispiel von Ausgabe Kandidaten Paare bei Kandidaten Generierungsalgorithmus, ist möglich. Die Methode der Momenten Schätzung ist eine einfache und effektive Methode zum Lernen den Parametern für die Beta Verteilung. Bei dieser Methode erfolgt die Berechnung der Stichprobenmomente (Stichprobenmittelwert und Stichprobenvarianz) unter der Annahme, dass diese entsprechen die wahren Momente der Verteilung und abhängig von den zu den erhaltenen Momenten führenden Parameterwerten. α, β haben folgende Schätzungen:

;

Wobei ist das Stichprobenmittel und die Varianz mit den folgenden Werte:

;

Für eine gegebene Prior Beta (α, β) Verteilung auf die Ähnlichkeit und ein betrachtetes Ereignis M (m, n) die Posterior-Verteilung der Ähnlichkeit ist die folgende:

p(s| M(m, n)) =

=

**Inferenz:** durch die Berechnung der Gleichungen 3, 4, und 6 kann die Durchführung der Inferenz gezeigt werden. Die Wahrscheinlichkeit, dass die Ähnlichkeit größer als den Schwellenwert ist, ist wie folgend gegeben durch:

=

(.,.) entspricht die regulierte unvollständige Beta-Funktion verantwortlich für die Angabe des cdf für die Beta-Verteilung. Die Ähnlichkeitsschätzung nach der Beobachtung von m Übereinstimmungen in n Hashes entspricht die Art der Posterior Verteilung p (s | M (m, n)). Die abgeleitete Konzentrationswahrscheinlichkeit der Ähnlichkeitsschätzung sieht wie folgend aus:

= -

Wenn diese abgeleitete Konzentrationswahrscheinlichkeit der Ähnlichkeitsschätzung in der entsprechenden Stelle des Algorithmus 1 ersetzt wird, dann entsteht eine speziell auf Jaccard-Ähnlichkeit angepasste Version von Bayes-LSH.

**Die Cosinus Ähnlichkeit:**

**LSH Familie:** Für die Cosinus Ähnlichkeit ist es so, dass es eine Zuordnung jeder Hash-Funktion mit einem Zufallsvektor gibt. Die Komponente von entsprechen eine Stichprobe aus dem Standard-Gauß-Operator (μ = 0, σ = 1). Bei der BayesLSH für die Jaccard-Ähnlichkeit fehlte, den Fakt, dass anstatt für einen Cosinus Ähnlichkeitsmaß, ist die LSH-Familie für mit . Dieser Ähnlichkeitsfunktion wird hier r (x, y) benannt. Ausdrücklich,  
 Pr[hi(x) == hi(y)] = r (x, y)

Pr [M (m, n) |r] =

Bei probabilistischen Garantien für die Qualität der Ausgabe ist das recherchierte ist cos (x, y) und nicht r (x, y). Der Ausdruck der Posterior Wahrscheinlichkeit muss in Form s = cos (x, y) sein. Der Ausdruck der Wahrscheinlichkeit könnte auch s = cos (x, y) sein anstatt r, was cos () - Begriffe einbezieht. Die Berechnung der Posterior-Verteilung und einschließlich die Transformation von r in der Posterior-Verteilung wird durchgeführt da die Einbeziehung von cos () - Begriffe der Fund des   
Priors erschwert.

**Die Wahl des Priors:** gewählt muss die Prior-Verteilung für r, dessen Bereich (0.5, 1) ist, wobei die unterstützte Domäne der Beta-Verteilung (0,1) ist. Die Abbildung der Standard-Beta-Verteilung ist dennoch möglich auf die Domäne (0.5,1), wobei die Konjugation der Verteilung zur binomiale Wahrscheinlichkeit nicht möglich ist. Die Nutzung einer einfachen Gleichverteilung auf [0.5, 1] ​​als Prior für t dient als Lösung. Das Prior pdf ist dann:

Das posteriore pdf nach der beobachteten Zustimmung von m aus dem ersten

n Hashes ist:

p (r| M (m, n)) =

=

=

entspricht die unvollständige Beta-Funktion und hat den folgenden Wert:

=

**Die Inferenz:** Die Konvertierung von r nach s und umgekehrt ist wichtig um die Berechnung der Gleichung 3, 6 und 4 zu ermöglichen. Sei eine 1-zu1-Funktion r2c : [0.5, 1] → [0, 1], die die Abbildung von r (x, y) nach cos (x, y) gibt, r2c (r) = cos(π ∗

(1 − r)). Sei c2r, die dieselbe Karte umgekehrt ausführende 1-zu-1-Funktion, c2r = .

Sei R eine Zufallsvariable mit R = c2r(S) und = c2r(t). Nach beobachteter Abstimmung von m aus der ersten n Hashes, die Wahrscheinlichkeit die Cosinus Ähnlichkeit größer als der Schwellen t ist, ist die folgende:

Pr[S t|M(m, n)] = Pr[c2r(S) c2r(t)|M(m, n)]

= Pr[R |M(m, n)]

=

=

=

Sei , die Erhaltung eines geschlossenen Ausdrucks für ist durch die nach = 0 möglich. r = wird erhalten, deswegen = . Nun , deswegen . Dieser Ausdruck hilft dabei bei der Berechnung von der Gleichung 4.

Pr[| − S| < δ|M(m, n)]

= Pr[ − δ < S < + δ|M(m, n)]

= Pr[c2r( − δ) < c2r(S) < c2r( + δ)|M(m, n)]

= Pr[c2r( − δ) < R < c2r( + δ)|M(m, n)]

=

=

Konkrete Ausdrücke für die Gleichungen 3, 4 und 6 werden erhalten und dadurch wird ein an Cosine Ähnlichkeit angepasster BayesLSH instanziiert .

[SAP12] V. Satuluri, S. Parthasarathy, Bayesian Locality Sensitive Hashing for Fast Similarity Search, 2012 <https://arxiv.org/pdf/1110.1328.pdf> (Letzter Zugriff:20.03.2018)

[JWC07] <http://www.cs.princeton.edu/cass/papers/mplsh_vldb07.pdf>,

William Josephson Zhe Wang Moses Charikar Kai Li, Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search, 2007 (Letzter Zugriff: 27.05.2018)