Freie Universität Berlin

Prof. Dr. Agnès Voisard, Nicolas Lehmann

Datenbank Systeme, SoeSe 2017

Project 1

TutorIn: Hoffman Christian
Tutorium Tutorium 3

IngridTchilibou,Emil,Boyan Hristov

10. Mai 2017

---

1 Aufgabe: Projektdokumentation

Aufgabenstellung des Dozenten

2 Aufgabe:Explorative Datenanalyse

1) *Laden von der Datein herum*
2) angeschaut
3) Beschreibung von der Datensatz Wir haben einen Tabelle mit 10 Attributen von verschiedenen Tweets aus der America Election von 2016. Jede Tupel beschreibt genau von wem war den Tweet gepostet? was war die Nachricht von diesem Tweet war diesem Tweets weitergeleitet? war die Authenticität von dem Tweet gut(Das heißt ob der Tweet von diese Person ,der gepostet hat geschrieben war?

   Zusätzlich gibt es auch manche Tweets, die 'truncated' worden sind. Das heißt, dass die ein zitiertes Tweet oder irgendwelche Media enthalten, deswegen überschritten sie das Twitter Zeichenlimit und werden abgeschnitten.

3 Aufgabe: ER-Modelierung

Alle drei von uns haben dieses Modell gründlich besprochen und eine Entscheidung getroffen. Wir glauben, dass dieses Modell das beste für unsere Vorstellung des Projekts ist.

Probleme und Lösungen

- Um 'Welche Hashtags werden am meistens verwendet' zu beantworten, muss man für jeden Hashtag überprüfen, in welchen Tweets er vorkommt und dann diese aufzählen. Deswegen haben wir uns entschieden, in dem Hashtah Entitätstyp ein Attribut TotalCount zu speicher. So werden wir auch weniger Server-Anfragen bearbeiten müssen.

- Der Datenumfang ist über mehreren Monaten. Um die Fragen 'Wann traten insgesammt am meisten Hashtags auf?' und 'Wie hat sich die Häufigkeit der Verwendung eines speziellen Hashtags im Laufe der Zeit entwickelt?' zu beantworten müssen wir jeden Tweet zu einem zeitlichen Bereich zuordnen.

  Wir haben uns deswegen entschieden, einen Entitätstyp 'Week' zu haben, wo wir an einer Woche benutzten Hashtags speichern werden. (Natürlich auch einen Attribut "Count"für höhere effizienz verwenden)

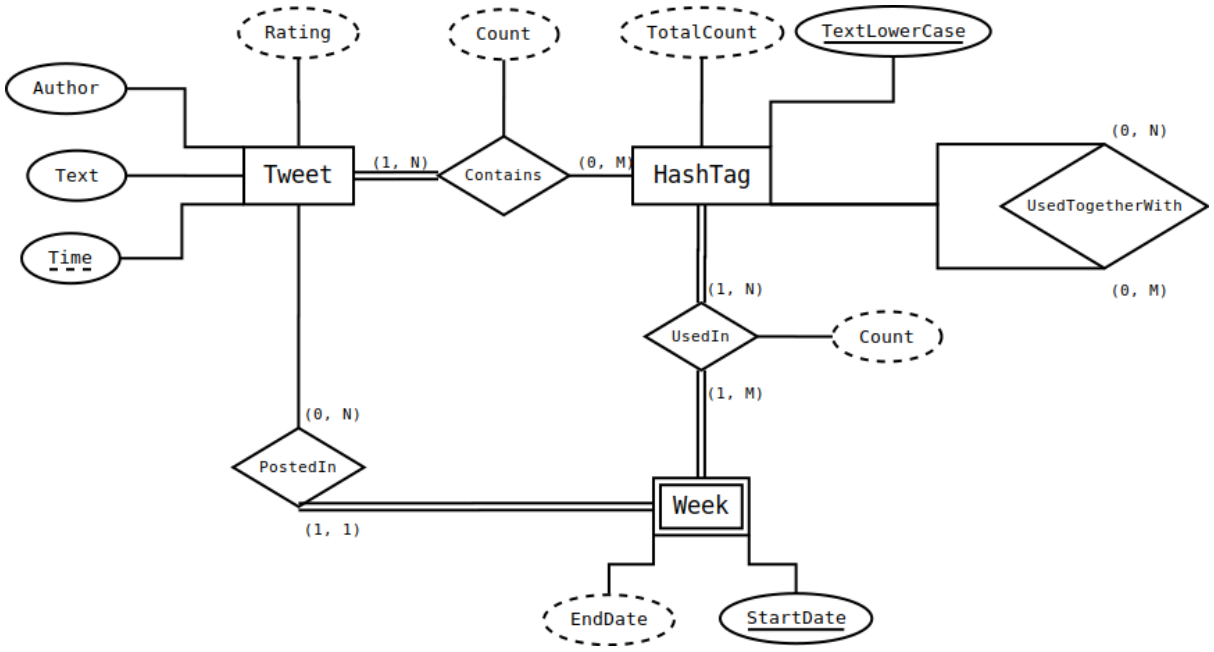
- Unserer Meinung nach, wäre eine Metrik für wichtige Tweets die Summe von Retweets und Favourites. So haben wir einen Attribut 'Rating', der nach der folgende Schema berechnet wird;

$$retweets * weight_1 + favourites * weight_2 = rating$$

  wobei $weight_1$ und $weight_2$ später genau bestimmt werden.

- Es wird ziemlich aufwändig herauszufinden, welche Paar von Hashtags am häufigsten gemeinsam auftritt, weil wir über alle Tweets iterieren müssen. Deswegen haben wir eine N:M rekursive Relation auf 'Hashtag' erstellt.

- Eigentlich haben wir zu fast jedem Entitätstyp einen Attribut 'count' hinzugefügt, weil die Berechnung von Anzahl viel einfacher zu Parse-Zeit wäre, als jedes mal eine Anfrage zu machen, und dann alle verschiedene Entitäten aufzuzählen.



4 Aufgabe: Relationales Modell

Nachdem wir die ER-Relation hatten, war es einfach das relationale Modell zu erstellen. Es kann sein, dass OriginalAuthor, ReplyTo und isQuote für uns nicht relevant sind.

```
Tweet(ID :: int, Handle :: char(20), Text :: char(200), Time :: Date, Favourites :: int,
    Retweets :: int, Rating :: int, Count :: int, OriginalAuthor :: char(20), replyTo :: char(20),
    isRetweet :: bool, isQuote :: bool, isTruncated :: bool, sourceUrl :: char(100) )
Hashtag(ID:: int, Text :: char(30), TotalCount :: int)
Week(ID:: int, StartDate :: Date, EndDate :: Date)

HashtagTweet(Tweet.ID, Hashtag.ID)
WeeklyHashtag(Week.ID, Hashtag.ID, Count :: int)
HashtagPairs(HT1.ID, HT2.ID)
```

5 Aufgabe: Datenbank erstellen