Informatiekunde, Rijksuniversiteit Groningen
17 maart 2015 *Begeleider: Johan Bos, Malvina Nissim*

Research proposal BA-scriptie Informatiekunde

Sub event detectie tijdens sportwedstrijden op basis van tweets

Mart Busger op Vollenbroek - *s2174634*

Probleemstelling

Onderscheid maken tussen main events en sub events kan moeilijk zijn. Door onderzoek te doen naar mogelijkheden om deze van elkaar te scheiden, wordt er geprobeerd een effectieve, breed inzetbare methode te ontwikkelen voor deze scheiding tussen main- en sub events. Van belang is hierbij dan het scheiden van hoofd- en bijzaken. In dit onderzoek zal nader gekeken worden naar het identificeren van sub events tijdens sportevenementen of wedstrijden. Bij een voetbalwedstrijd kan gedacht worden aan overtredingen en doelpunten die dan sub events zijn van de wedstrijd zelf. Het doel van dit onderzoek is om te beginnen met het detecteren van sub events tijdens voetbalwedstrijden, om daarna door middel van het veranderen van een paar waarden hetzelfde te bereiken bij andere sportwedstrijden.
De onderzoeksvraag die hierbij gesteld wordt, is als volgt:
Hoe kunnen met behulp van tweets sub events van een main event gescheiden worden tijdens sportwedstrijden?

De subvragen die gesteld worden zijn als volgt:
- *Wat zijn de beste indicators voor het detecteren van een sub event?*
- *Welke resultaten geeft het systeem op basis van unigrammen?*
- *Welke resultaten geeft het systeem op basis van bigrammen?*
- *Zijn tweets zonder hashtags betere documenten om sub event detectie uit te voeren?*

Dataset

Om deze onderzoeksvraag te testen zal er data verzameld moeten worden van voetbalwedstrijden. Op Twitter is het gebruikelijk om de twee afkortingen van de ploegen samen in een hashtag te gebruiken (#fcbrma staat voor FC Barcelona tegen Real Madrid), hierdoor zou er makkelijker onderscheid gemaakt kunnen worden om de juiste data te vinden voor het onderzoek. In eerste instantie wordt er alleen gekeken naar de tekstuele data van de tweets.

Methode

Het systeem zal op basis van de documenten die het aangeleverd krijgt, de documenten moeten kunnen onderscheiden in main events en sub events. Alle documenten die gebruikt worden zijn tweets. Om dit te doen zal eerst een strakke definitie van een main event en een sub event bedacht moeten worden. Deze kan gebaseerd worden op bestaande literatuur of gedefinieerd worden in het kader van dit onderzoek. Criteria voor deze scheiding kunnen onder anderen zijn: locatiegegevens, specifieke teksten of de hoeveelheid documenten per bepaalde tijdseenheid. Zoals vermeld bij de dataset zal er in eerste instantie alleen gekeken worden naar de tekstuele data van de tweets. Mocht het mogelijk zijn, dan wordt er gekeken naar de mogelijkheden van het gebruiken van locatiegegevens bij de tweets. Om de verschillende sub events te onderscheiden binnen een main event, zullen de sub events gecodeerd moeten worden. Bijvoorbeeld op basis van de rugnummers en het specifieke sub event dat gedetecteerd wordt. Als de nummer 10 van een thuisspelende ploeg een rode kaart krijgt, zou een voorbeeld van een codering “H10RED" kunnen zijn (H voor home playing team, 10 het rugnummer en RED voor de rode kaart). Daarnaast

zal er gekeken moeten worden naar het moment waarop de tweet geplaatst is, dit is altijd ná het event omdat de tweets reacties zijn op wat er in het veld gebeurt.

Evaluatie en uitkomsten

Het programma dat voor dit onderzoek ontwikkeld wordt, moet in staat zijn documenten te scheiden in een main event en sub events. Na iedere wedstrijd staat er vast in welke minuut er wat gebeurd is, dus aan de hand van die gegevens kan geëvalueerd worden hoe accuraat de detectie van van sub events tijdens een main event (wedstrijd) is.
Er kan gekeken worden of er een baseline vastgesteld kan worden en op basis daarvan kan gemeten en geëvalueerd worden of het programma goed werkt.
Ten slotte wordt er gekeken of de resultaten en het systeem te generaliseren zijn, zodat het systeem gebruikt kan worden voor andere sporten dan voetbal. Het idee is dan om een paar parameters aan te passen die voor iedere sport verschillend zijn. Bij hockey zijn er bijvoorbeeld strafcorners terwijl er bij voetbal penalty's zijn.

Relevante literatuur

Om beter begrip te krijgen van event detection in het algemeen en specifieker sub event detection, is er gezocht naar relevante literatuur op basis waarvan hypothesen gesteld kunnen en een onderzoeksvraag geformuleerd kan worden. In het artikel van Pohl et al. (2012) wordt gekeken naar het gebruik van sub event detection in noodsituaties. Zij hebben onderzoek gedaan met de metadata van Flickr en Youtube en scheiden met behulp van die data main events van sub events.
Het artikel van Del Fabro & Böszörmenyi wordt een algoritme beschreven waarmee verschillende soorten data samengevoegd worden om een samenvatting van een bepaald event te geven. Dit kan een goede basis zijn voor het scheiden van events.
Weng et al. beschrijven methoden van event detection op Twitter, ze ontwikkelen een systeem dat ieder woord apart behandelt (dus alleen unigrammen) en op basis van de woorden in de tweets maken ze analyses van events. Zij hebben dit getest tijdens de verkiezingen in Singapore in 2011.
Corney et al. hebben een soortgelijk onderzoek uitgevoerd als wat dit onderzoek beaamd. Hier kan verder op worden gebouwd, vooral op het gebied met betrekking tot de generalisering van de systeem voor gebruik bij andere sporten.

Bibliografie

Corney, D., Martin, C., & Göker, A. (2014). Spot the ball: Detecting sports events on Twitter. In *Advances in Information Retrieval* (pp. 449-454). Springer International Publishing.

Del Fabro, M., & Böszörmenyi, L. (2012). Summarization and presentation of real-life events using community-contributed content (pp. 630-632). Springer Berlin Heidelberg.

Pohl, D., Bouchachia, A., & Hellwagner, H. (2012, April). Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference companion on World Wide Web* (pp. 683-686). ACM.

Weng, J., & Lee, B. S. (2011). Event Detection in Twitter. *ICWSM*, *11*, 401-408.