Learning Semantic Relationships between Entities in Twitter

ICWE, Cyprus, June 22, 2011

Ilknur Celik, Fabian Abel, Geert-Jan Houben
Web Information Systems, TU Delft
What we do: Science and Engineering for the Personal Web

domains: news social media cultural heritage public data e-learning

- Personalized Recommendations
- Personalized Search
- Adaptive Systems

Analysis and User Modeling

Semantic Enrichment, Linkage and Alignment

user/usage data

Social Web

Learning Semantic Relationships between Entities in Twitter
60,000,000

number of tweets published per day
number of tweets per day that are interesting for me
Searching on Twitter

Trending topics: #itmakesmesmilewhen, #scholes, #singleladies, LULUS RAME-RAME, NAIK KELAS RAME-RAME, Serious Question, Daesung, Jeremy Kyle, Sean Kingston, IKEA
Issues with Multiple Keywords Search

![Twitter search result](image)

Results for eindhoven entertainment event

No results for **eindhoven entertainment event**
Let’s try to search with One Keyword

Trending topics: #itmakesmesmilewhen, #scholes, #singleladies, Serious Question, NAIK KELAS RAME-RAME, LULUS RAME-RAME, Daesung, Sean Kingston, Jeremy Kyle, IKEA
Learning Semantic Relationships between Entities in Twitter
Next Saturday @thatsimpsonguy aka **Guilty Simpson** will be performing at **Area51** in my hometown **Eindhoven**. #realliveshit #iwillspinrecords
Is there an easier way? 
→ Faceted Search can help

Expand Query:
Locations more...
Events more...
Music Artists:
+ Guilty Simpson
+ Bryan Adams
+ Elton John
+ Golden Earring
+ Rihanna
+ The eagles
+ 3 Doors Down
more...

Current Query:

Current Query:
Eindhoven ✗ Music ✗

Results:
1. Yskiddd: Next saturday @thatsimpsonguy aka Guilty Simpson will be performing at Area51 in my homeytown Eindhoven. #realliveshit #iwillspinrecords2


3. sanmiquelmusic: This Saturday I'm joining @KrusadersMusic to Intents
Semantic relationships between entities are essential to realize such applications.
Relation Discovery Framework

Entity extraction & semantic enrichment

Relation discovery

Relation discovery framework

Typed relations

Applications
- Browsing support
- Query suggestions
- Schema enrichment

Learning Semantic Relationships between Entities in Twitter
Learning Semantic Relationships between Entities in Twitter

Entity Extraction and Semantic Enrichment

powered by

@bob: Julian Assange got arrested http://bit.ly/5d4r2t

Tweet-based enrichment

Julian Assange

Julian Assange, the founder of WikiLeaks, is under arrest in London...

News-based enrichment

Julian Assange

London

WikiLeaks
Relation Learning Strategies

• Relation:
  \[ \text{relation}(e_1, e_2, \text{type}, t_{\text{start}}, t_{\text{end}}, \text{weight}) \]

• Relation Learning strategy:
  • Input: entity \( e_1 \) and \( e_2 \), time period \((t_{\text{start}}, t_{\text{end}})\)
  • Challenge: infer \textit{weight} and \textit{type} of the relation for the given

• Weighting according to co-occurrence frequency:
  a) \textbf{Tweet-based:} count co-occurrence in tweets
  b) \textbf{News-based:} count co-occurrence in news
  c) \textbf{Tweet-News-based:} count co-occurrence in both tweets and news
Research Questions

1. Which strategy performs best in detecting relationships between entities?

2. Does the accuracy depend on the type of entities which are involved in a relation?

3. How do the strategies perform for discovering relationships which have temporal constraints (trending relationships)?
Dataset

more than:

20,000  Twitter users
2 months
10,000,000  tweets
75,000  news

WikiLeaks founder, Julian Assange, under arrest in London

Nov 15  Dec 15  Jan 15  time
Dataset Characteristics
Tweets and news articles per day

- **Twitter posts**: 50,000-400,000 tweets per day
- **News articles**: 100-1000 news articles per day

---

Learning Semantic Relationships between Entities in Twitter
Entities referenced per day

- ~40% tweets do not mention any (recognizable) entity
- 72.6% of the top 1000 mentioned entities in Twitter are also mentioned in the mainstream news media
- 99.3% of the news articles mention at least one (recognizable) entity

10,000-100,000 entity ref. in tweets per day
5,000-20,000 entity ref. in news per day
Number of Distinct Entities per Entity Types

39 types of entities

- URL
- Person
- Country
- Organization
- City
- Movie
- Product
- MusicAlbum
- MusicGroup
- SportsEvent
- PoliticalEvent
- City

Learning Semantic Relationships between Entities in Twitter
Performance of Relation Learning Strategies
Our Ground Truth of *true* relations

1. Based on DBpedia:
   - We mapped entities to their corresponding DBpedia resources
     - No appropriate DBpedia URIs for more than 35% of the entities
   - We analyzed whether there is a *direct relation* between two entities

2. Based on user study:
   - Participants judged whether two entities are really:
     a) related (62.6% were rated as *related*)
     b) related in the given time period (57.3% were rated as *related*)
   - Overall: 2588 judgments

Thank you!
1. Which strategy performs best in detecting relationships between entities?
Accuracy of relation discovery

Combining both tweet-based and news-based strategies allows for highest accuracy
Tweet-based strategy saturates quickly.

Combined strategy (and news-based) increase in performance.
2. Does the accuracy depend on the type of entities which are involved in a relation?
Does the accuracy depend on the type of entities?

Relationships which involve events can be discovered with high precision.

- Person/Group-Event: 87% precision, 92% precision
- Person/Group-Location: 26% precision
- Person/Group-Product: 23% precision
Does the accuracy depend on the type of entities? (cont.)

Relationships between persons/groups are difficult to detect.

Relationships between events can be detected with highest precision.
3. How do the strategies perform for discovering relationships which have temporal constraints?
Tweet-based strategy performs better in discovering relationships that are valid only for a specific period in time.
Where do relationships emerge faster?

- News is faster
- Twitter is faster

Speed of strategies is domain-dependent
Conclusions and Future Work

What we did: relation discovery framework based on Twitter

Findings:
1. Strategy that considers both tweets and (linked) news articles allows for highest accuracy
2. Performance varies for different domains (e.g. event-relationships can be detected with highest precision)
3. Tweet-based strategy allows for detecting relationships, which have a restricted temporal validity, with high precision (and fast)

Ongoing work: Adaptive Faceted Search on Twitter

http://wis.ewi.tudelft.nl/tweetum/
Relation Discovery for Adaptive Faceted Search

2. Analyze (temporal) relationships of entities that appear in the user profile to adapt facet ranking.

Music Artists:
+ Guilty Simpson
+ Bryan Adams
+ Elton John
+ Golden Earring
+ Rihanna
+ The eagles
+ 3 Doors Down
more...

Current Query:
Eindhoven Music

Results:
1. Yskiddd: Next saturday
2. Usee123: Cool #EV3door7980 !!!
3. sanmiquelmusic: This Saturday I'm joining @KrusadersMusic to Intents

1. Analyze (temporal) relationships of entities of the “current query” to adapt facet ranking.

3. sanmiquelmusic: This Saturday I'm joining @KrusadersMusic to Intents

Learning Semantic Relationships between Entities in Twitter
Thank you!

Ilknur Celik, Fabian Abel, Geert-Jan Houben

Twitter: @persweb
http://wis.ewi.tudelft.nl/tweetum/
Who is this? What are his personal demands? How can we make him happy?

Help me to tackle the information overload!

Recommend me news articles that now interest me!

Help me to find interesting (social) media!

Do not bother me with advertisements that are not interesting for me!

Personalize my Web experience!

Give me personalized support when I do my online training!