CAPTURE USER INTEREST USING SEMANTIC WEB

AAKASH GOEL (DATA SCIENTIST, FRACTAL ANALYTICS)

NITIN AGARWAL (SENIOR DATA SCIENTIST, FRACTAL ANALYTICS)

OUTLINE

- Abstract

- Problem Definition

- Terminologies

- Proposed Solution

- Application

- Challenges

- References

ABSTRACT

Today more than 3 billion people are using social media and using it as a medium to express their real feelings which makes different social media platform like Facebook, Twitter etc. an ideal source for capturing interest of users. Obviously, data mined from social media alone cannot be used to achieve target i.e. predict user's Interest, it needs some form of supervision.

Our talk propose how Semantic web a.k.a Knowledge bases add supervision into system and can prove helpful to predict user's Interest given social media data. Once, User's Interest is captured, it can be widely used for many purposes like Recommendation system, campaigning, analytics over user interests etc.

Keywords: Knowledge systems, linked data, OpenIE, NLP, Semantic Web, User Interest, SPARQL.

PROBLEM DEFINITION

- How to identify user's Interest from social media (Twitter, Facebook) posts ?

- According to Bakalov et al., interest is defined as "a fact indicating that a given user is interested in a certain term with a certain degree of interest", in which they use "term" to refer to a concept denoting either a real world object or an abstract notion [1].

- Are social media posts good source of finding interest?

  - More than 3 billion people are using social media to express their feelings.

  - Many linguistic challenges are there like slang words, incorrect usage of grammar.

TERMINOLOGIES

- **WWW (world wide web) -** Information space where documents and other web resources are identified by URL's and these documents are interlinked by hypertext links and can be accessed via internet.

- **Semantic Web** – Extension of www in which data in web pages is structured which creates web of knowledge and provide us with a new frontier in information retrieval and intelligent agents. It is about linked data which can be searched for patterns and queried for information, to develop intelligent agents.
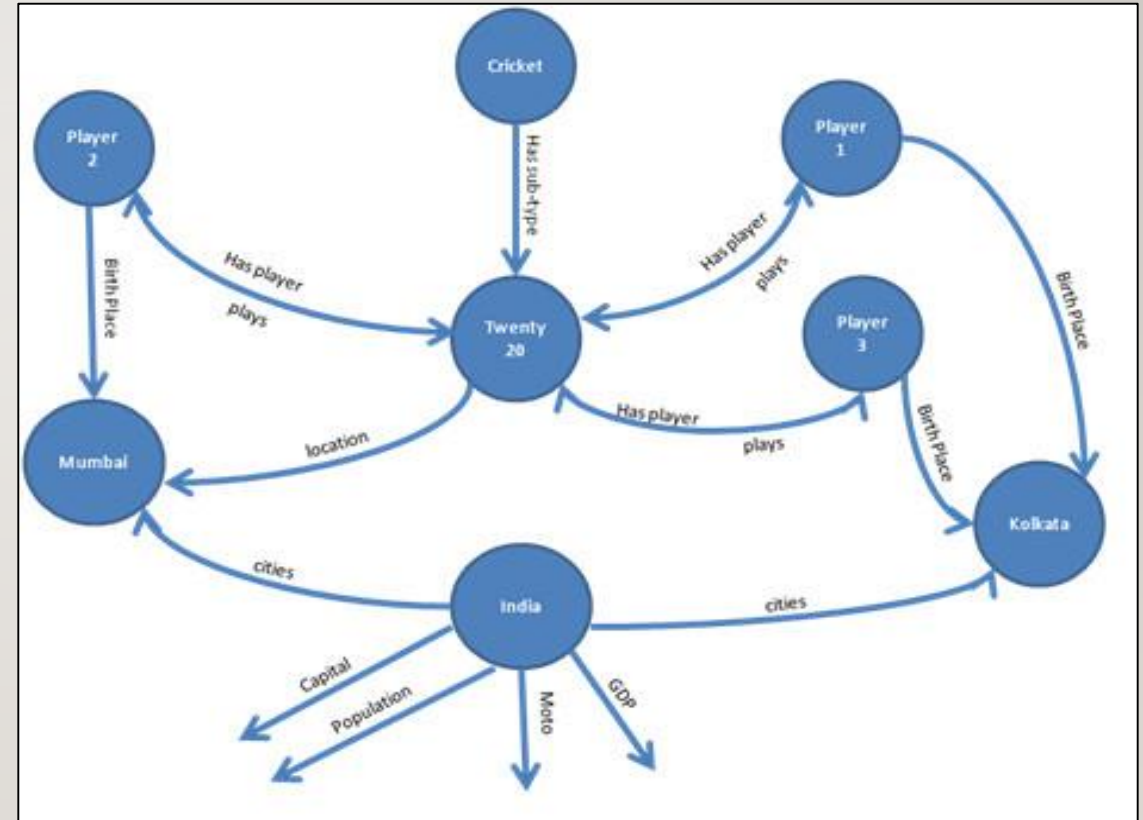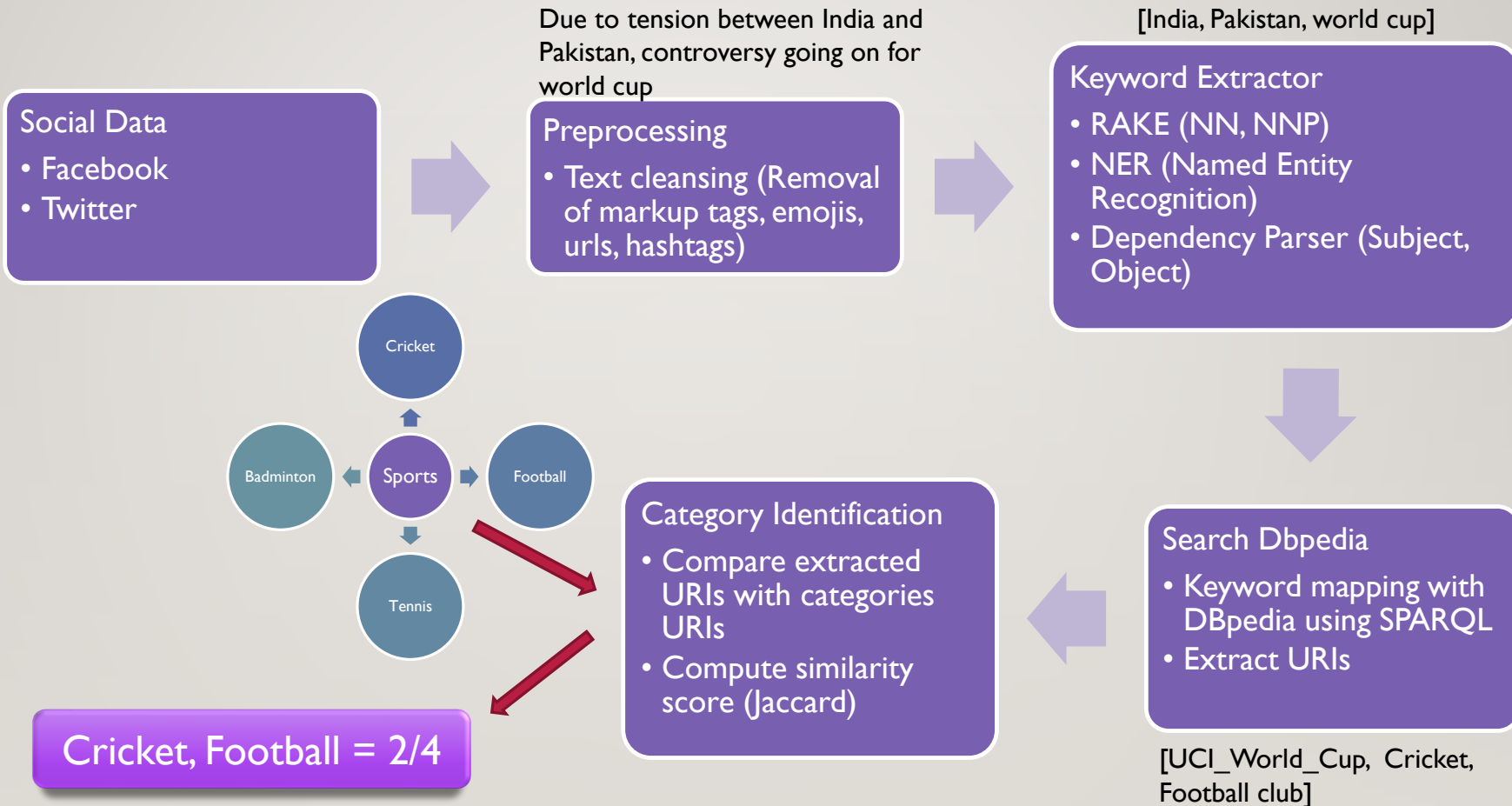
TERMINOLOGIES CONTD..

- **RDF (Resource Description Framework)** – It is an infrastructure which provide information or metadata for data available on the Internet and methodology for specifying, structuring and transferring metadata, and provides the basic XML syntax for software applications to exchange or use that information.

- **SPARQL (SPARQL Protocol and RDF Query Language) -** RDF query language to retrieve and manipulate data stored in RDF format.

- **OWL (Web Ontology Language) -** A computational logic based language to represent rich and complex knowledge about things, group of things, and relations between things, such that the expressed knowledge in OWL can be exploited by computer programs. OWL documents are called Ontologies.

PROPOSED SOLUTION

Due to tension between India and Pakistan, controversy going on for world cup ☹ #Cricket https://www.msn.com/en-xl/asia/asia-sport/india-urged-to-boycott-world-cup-match-against-pakistan/ar-BBTMs3C?li=BBJE6P9

Social Data
• Facebook
• Twitter

Due to tension between India and Pakistan, controversy going on for world cup

Preprocessing
• Text cleansing (Removal of markup tags, emojis, urls, hashtags)

[India, Pakistan, world cup]

Keyword Extractor
• RAKE (NN, NNP)
• NER (Named Entity Recognition)
• Dependency Parser (Subject, Object)

Cricket

Badminton   Sports   Football

Tennis

Category Identification
• Compare extracted URIs with categories URIs
• Compute similarity score (Jaccard)

Search Dbpedia
• Keyword mapping with DBpedia using SPARQL
• Extract URIs

[UCI_World_Cup, Cricket, Football club]

Cricket, Football = 2/4

APPLICATION

- Interest level information of user can be used to create the personalized interest model.

- Interest based campaigning/ Advertising/ Recommendation System: Example - Company looking for target users whose interest lies in 'healthcare' domain.

- Business Analytics over User's Interest: Interest over time, Interest by regions. Example – Google Trend [7].

CHALLENGES

- It may happen that keywords aren't present at all in DBpedia.

- Difficult to extract relevant URI's from the results returned by DBpedia.

- Needs learning a new query language.

- Performance depends on the created ontology.

REFERENCES

1. F. Bakalov, B. König-Ries, A. Nauerz, and M. Welsch, "A Hybrid Approach to Identifying User Interests in Web Portals," Proc. The 9th International Conference on Innovative Internet Community Systems, Jena, Germany, 2009, pp. 123-134.

2. https://dbpedia.org/fct/

3. http://www.w3.org/RDF/FAQ

4. http://www.w3.org/2001/sw/sweo/public/UseCases/ClevelandClinic/

5. https://en.wikipedia.org/wiki/SPARQL

6. https://en.wikipedia.org/wiki/Resource_Description_Framework

7. https://trends.google.com/trends/?geo=US