User Interests Identification From Twitter using Hierarchical Knowledge Base

Pavan Kapanipathi*, Prateek Jain^,
Chitra Venkataramani^, Amit Sheth*

*Kno.e.sis Center, Wright State University
^IBM TJ Watson Research Center

#eswc2014Kapanipathi

IBM Research

Agenda

- Motivation
- Background
- Approach
- Evaluation
- Conclusion & Future Work

Agenda

- **Motivation**
- Approach
- Evaluation
- Conclusion & Future Work

Twitter

- Tapping into Social Networks to identify interests is not new (2006+). It works!!
  - Google, Bing, Samsung TV etc.

- Twitter Content
  - 500M+ Users generating 500M+ tweets per day.
  - Public and useful for research

What's there?

- Interests with lesser or no semantics
  - Bag of Words [1]
  - Bag of Concepts

- Some Semantics
  - Bag of Linked Entities with intentions of using Knowledge Bases. [2, 3]

1. Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. WSDM '10.
2. Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. UMAP '11
3. Fabrizio Orlandi, John Breslin, and Alexandre Passant. Aggregated, Interoperable and Multi-domain User Profiles for the Social Web. I-SEMANTICS '12.

#semtech talk "How To Make Linked Data More than Data" http://bit.ly/cILgV5 #Semtech #semanticweb #linkeddata federated #Sparql

Reply Retweet Favorite More

@google & #semanticweb: juxtapo kno.e.sis Norvig views/counter views with schema.org: http://bit.ly/google-on-SW #semtech #schema #semantics

View translation

Reply Retweet Favorite More

RT @EdGedvilla Great report from The Pew Research Center on "The Fate of the Semantic Web" http://bit.ly/d9ifNw

Reply Retweet Favorite More

Later today 4:45p at #semtech on Semantic Social Networking panel: http://bit.ly/bp81jl

Reply Retweet Favorite More

5 Real world app demoed of semantic sensor web/nw at #ssn2011 #iswc2011 at 3pm

Reply Retweet Favorite More

@novaspivak #iswc2009 semantic search w/ ontology driven human assisted site extraction on verts reinvents 2002 http://tinyurl.com/ykg9eqe

Reply Retweet Favorite More

A Visualization of the Metadata Universe pdf : http://bit.ly/bOIfCo via @addthis

View translation

Reply Retweet Favorite More

What's new?

- How can Semantics/Knowledge Bases be utilized to infer interests?
  - Extensive use of Knowledge Bases to infer user interests from Tweets is yet to be explored.

- First we started with utilizing Hierarchical Relationships

#semtech talk "How To Make Linked Data More than Data" http://bit.ly/cILgV5 #Semtech #semanticweb #linkeddata federated #Sparql

Reply Retweet Favorite More

@novaspivak #iswc2009 semantic search w/ ontology driven human assisted site extraction on verts reinvents 2002 http://tinyurl.com/ykg9eqe

Reply Retweet Favorite More

A Visualization of the Metadata Universe pdf : http://bit.ly/bOIfCo via @addthis

View translation

Reply Retweet Favorite More

#semtech talk "How To Make Linked Data More than Data" http://bit.ly/cILgV5 #Semtech #semanticweb #linkeddata federated #Sparql

Reply Retweet Favorite More

@novaspivak #iswc2009 semantic search w/ ontology driven human assisted site extraction on verts reinvents 2002 http://tinyurl.com/ykg9eqe

Reply Retweet Favorite More

Semantic Search

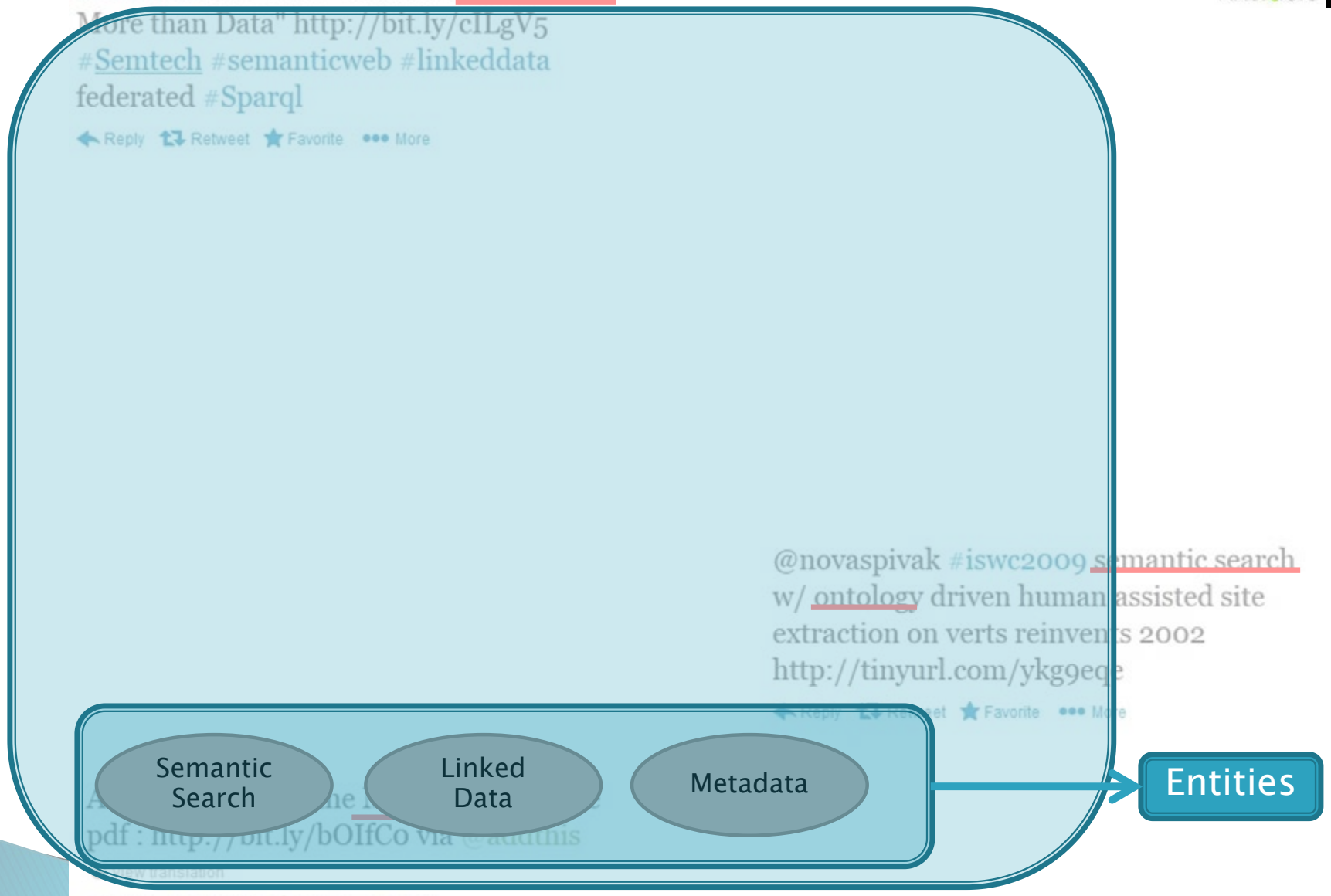
Linked Data

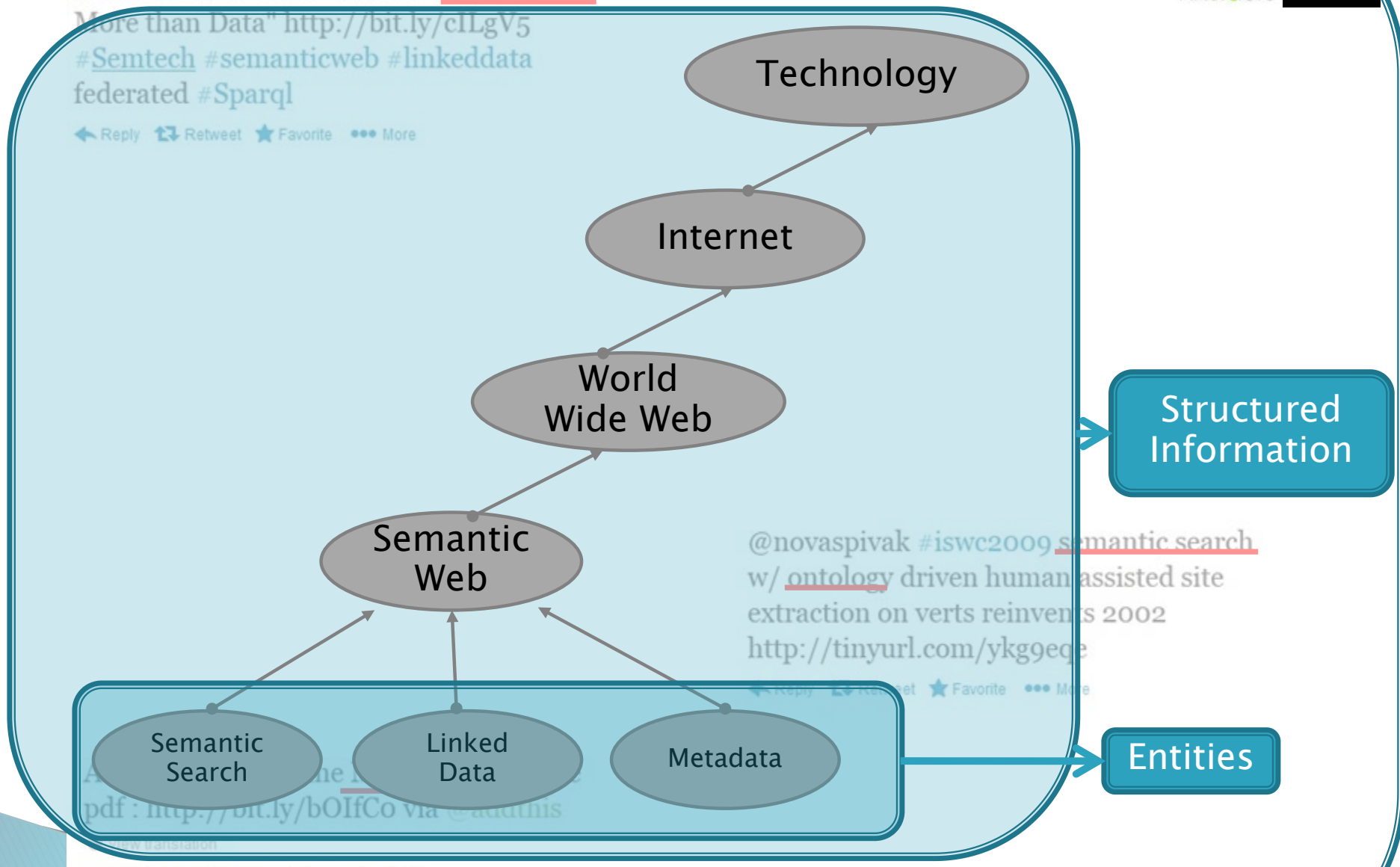
Metadata

pdf : http://bit.ly/bOIfCo via @addthis

View translation

Reply Retweet Favorite More


Semantic Search
Linked Data
Metadata
Entities


Technology
Internet
World Wide Web
Semantic Web
Semantic Search
Linked Data
Metadata
Structured Information
Entities

Advantages of Hierarchical Interests

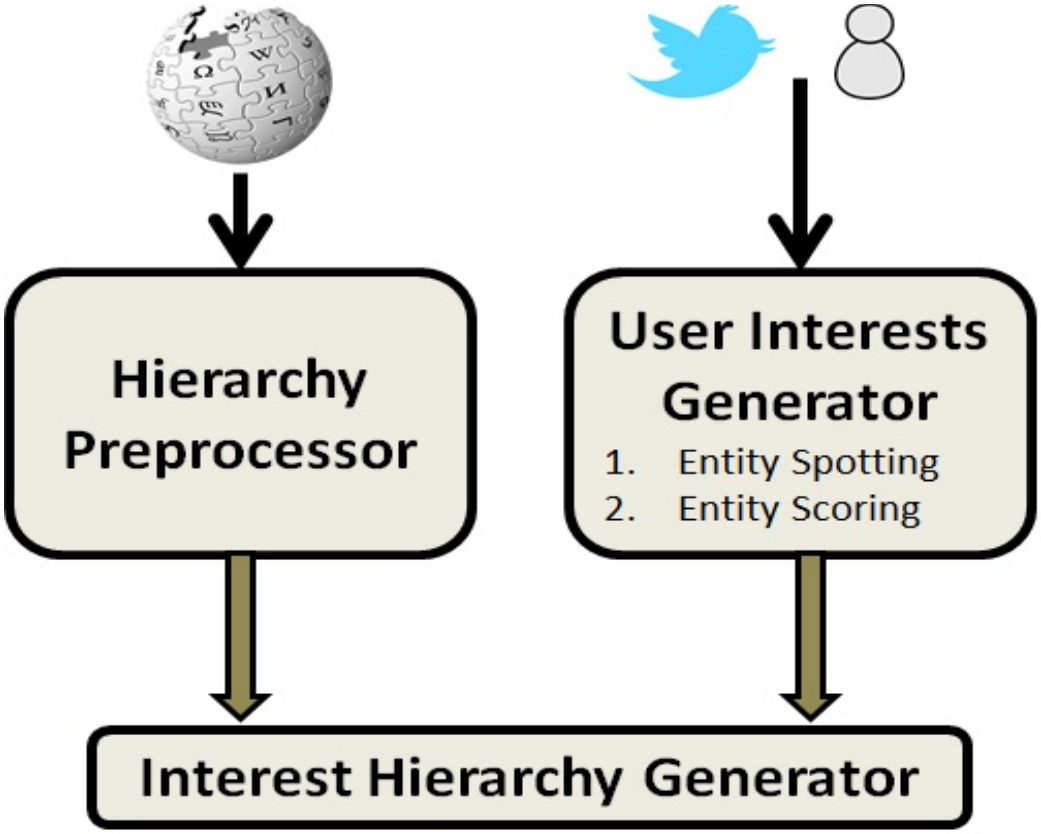
- Addressing Data Sparcity Problem
  - Infer more interests of the users with lesser data.

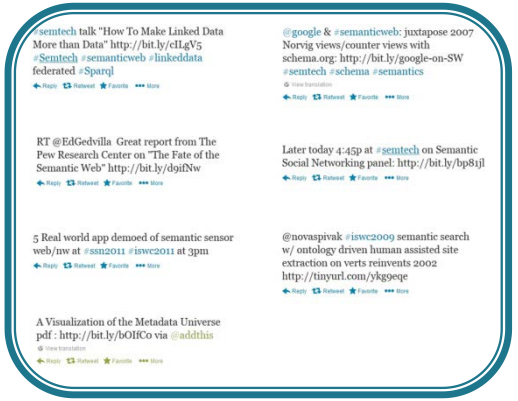
- Flexibility for Recommendations
  - Recommend about Sports or Football
    - KB knows that Football is a sub-category of Sports
  - Resource Description Framework and Semantic Web
    - RDF has lesser data online to recommend.

Agenda

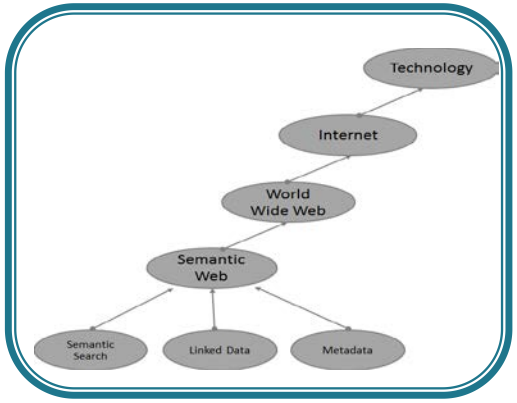
- Motivation
- **Approach**
- Evaluation
- Conclusion & Future Work

Approach



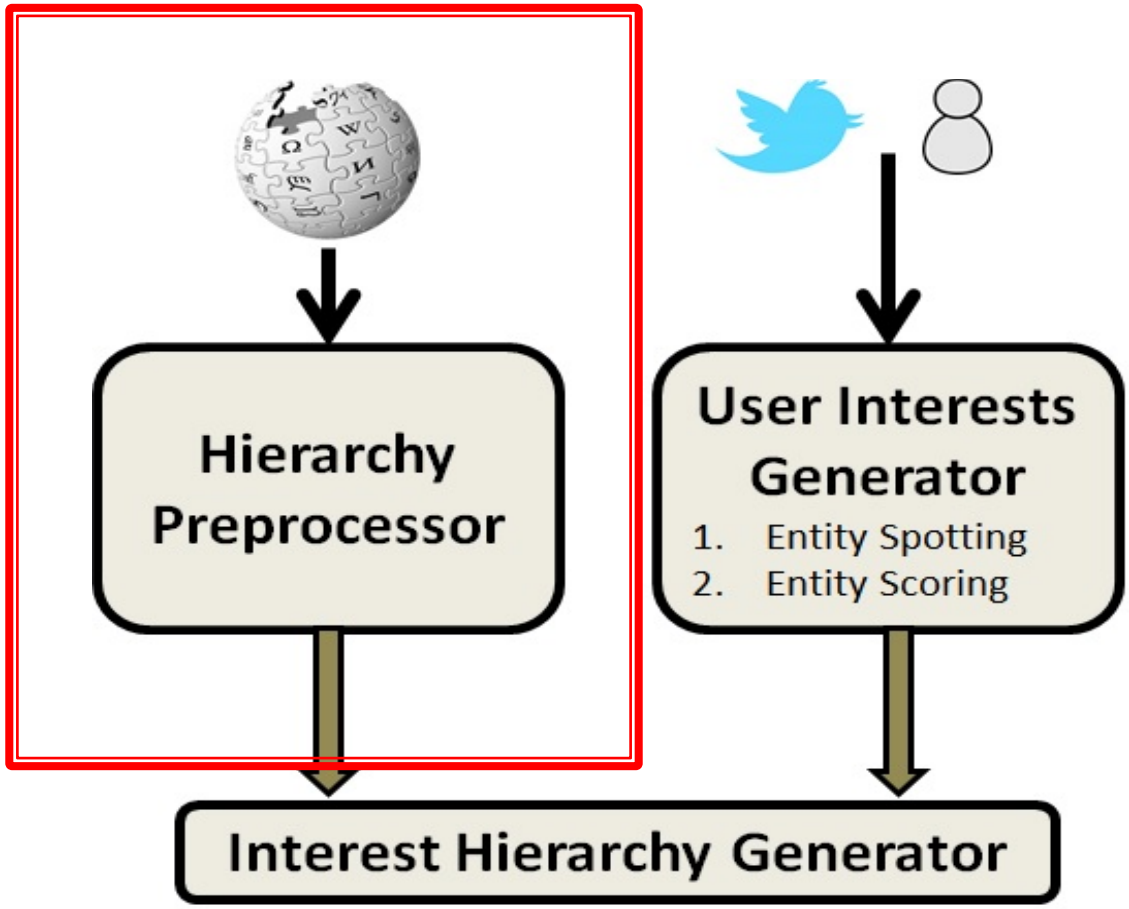


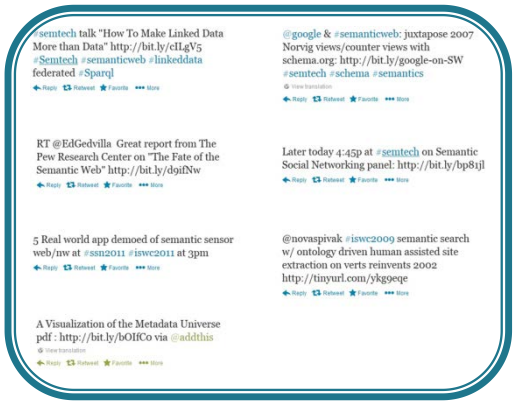
Tweets



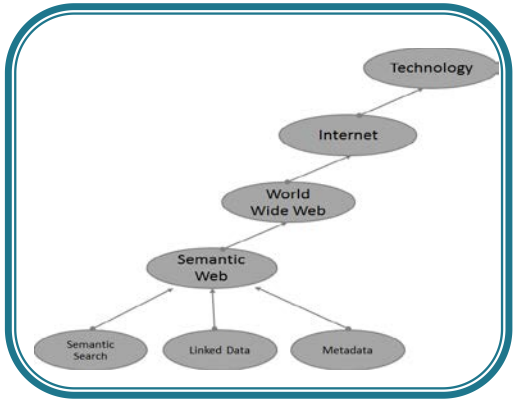
Interest Hierarchy

Approach





Tweets



Interest Hierarchy

Hierarchy Preprocessing

- Selecting an Ontology
  - Available: Wikipedia, Dmoz, OpenCyc, Freebase
  - Our framework can adapt to any ontology

- Wikipedia
  - Diverse Domains & Coverage
  - Resemblance to a Taxonomy
  - Extracted Structured Wikipedia – Dbpedia
  - Existing entity recognition techniques (Explained further)

Wikipedia Category Graph

- 4.2 Million Articles
- 0.8 Million Wikipedia Categories
- 2.0 Million Category-Subcategory relationships

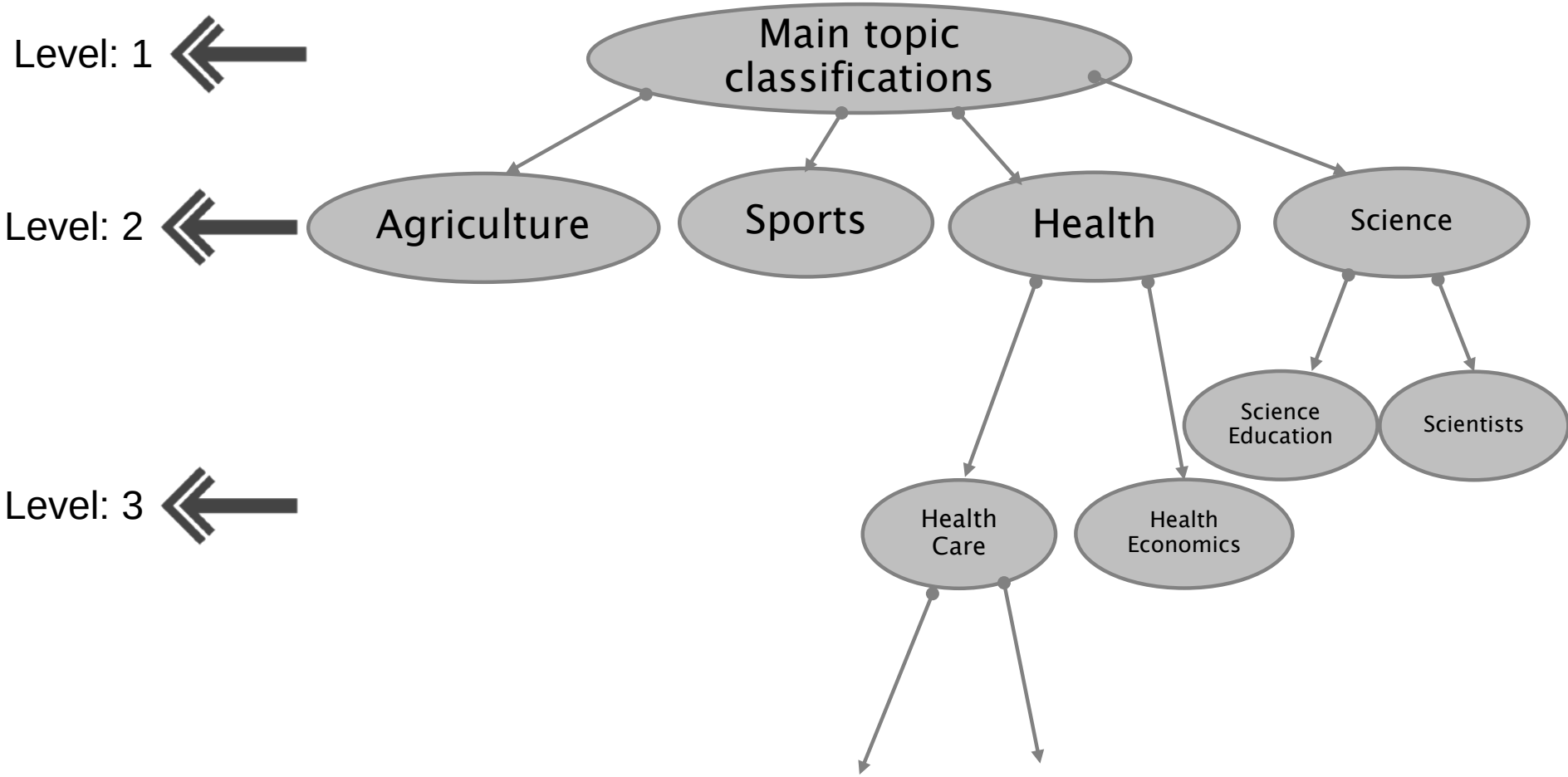
- Challenges
  - Since crowd-sourced – Noisy
  - Not a hierarchy/taxonomy
    - It is a graph
    - It has cycles

Wikipedia Hierarchy

- Clean up -- Removed Wiki Admin Categories

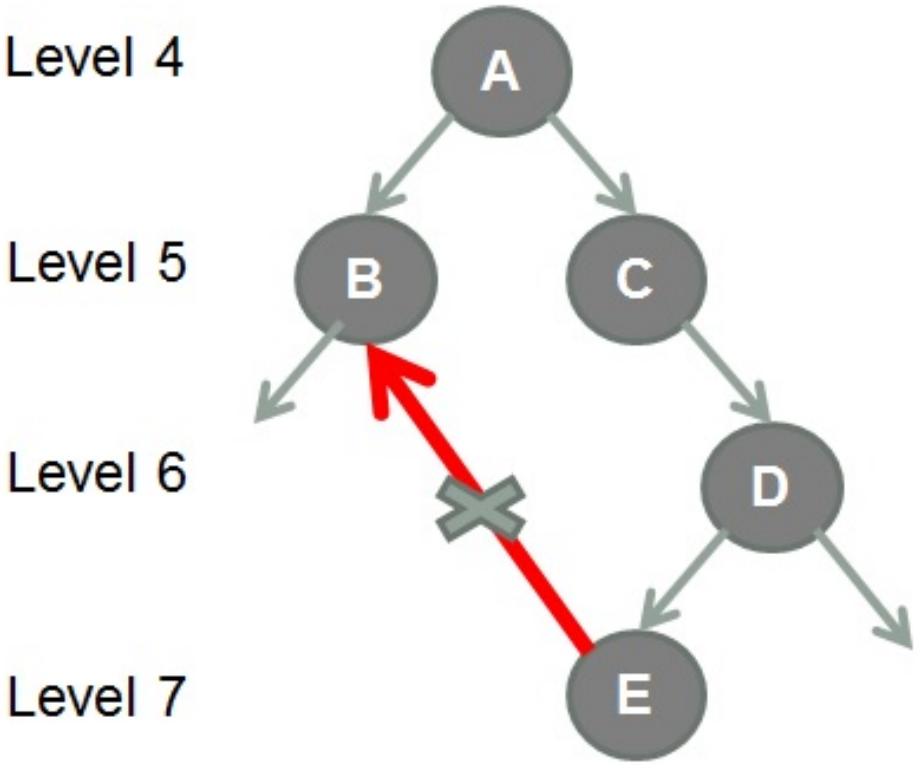
- Hierarchical Interest Graph needs a Base Hierarchy
  - ***Shortest Path from the root node***
    - Root Node: *Category:Main Topic Classifications*
    - Assumption – Hops to the root node determines the level of abstraction of the category.

Determining the Hierarchical Level

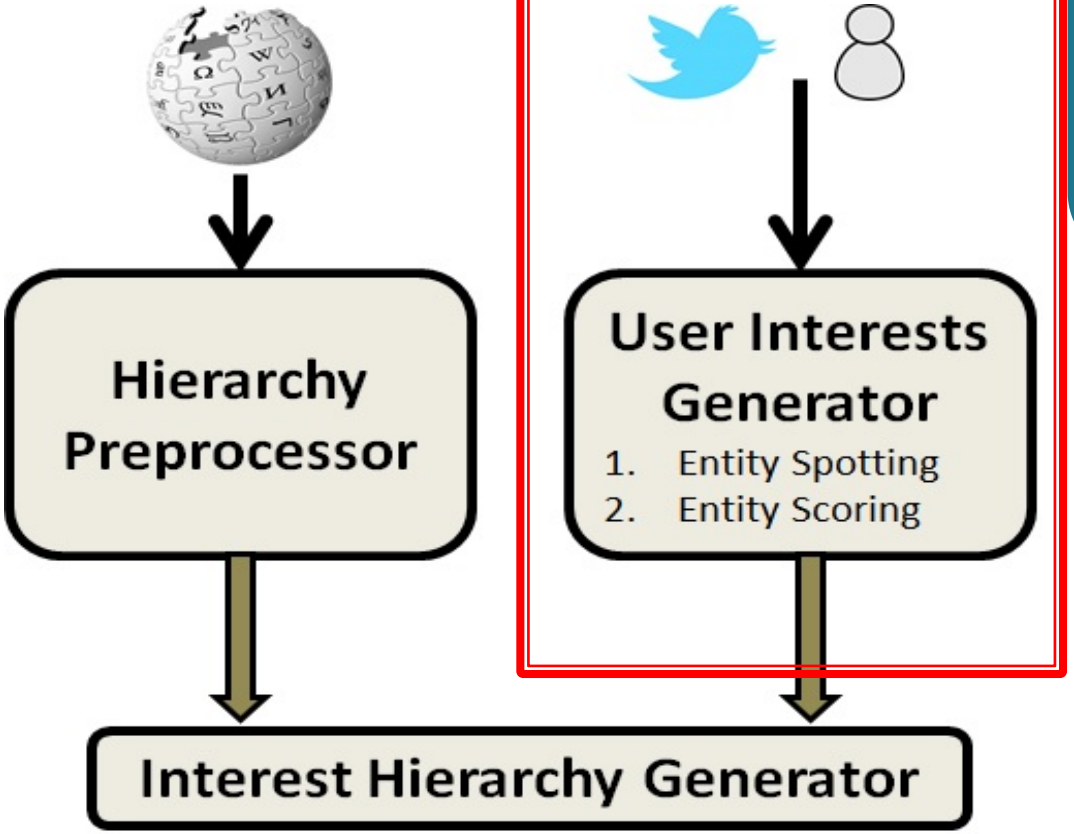


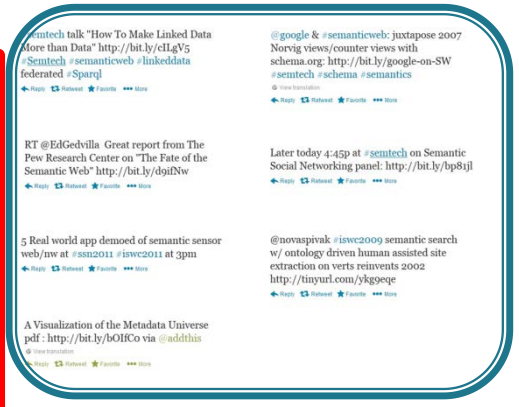
Links concurring to a Hierarchy

- Removing Links that does not concur to a hierarchy

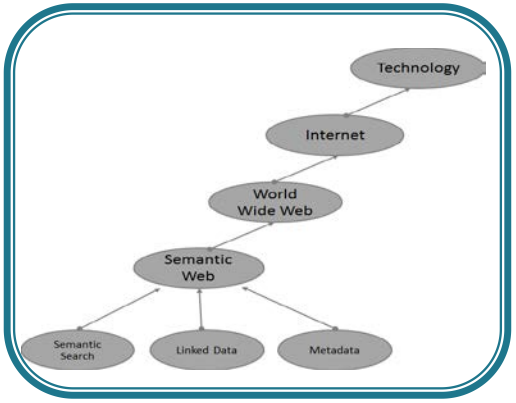


Approach





Tweets



Interest Hierarchy

User Interests Generator

- Extracting Wikipedia concepts from Tweets

- Interests Scoring

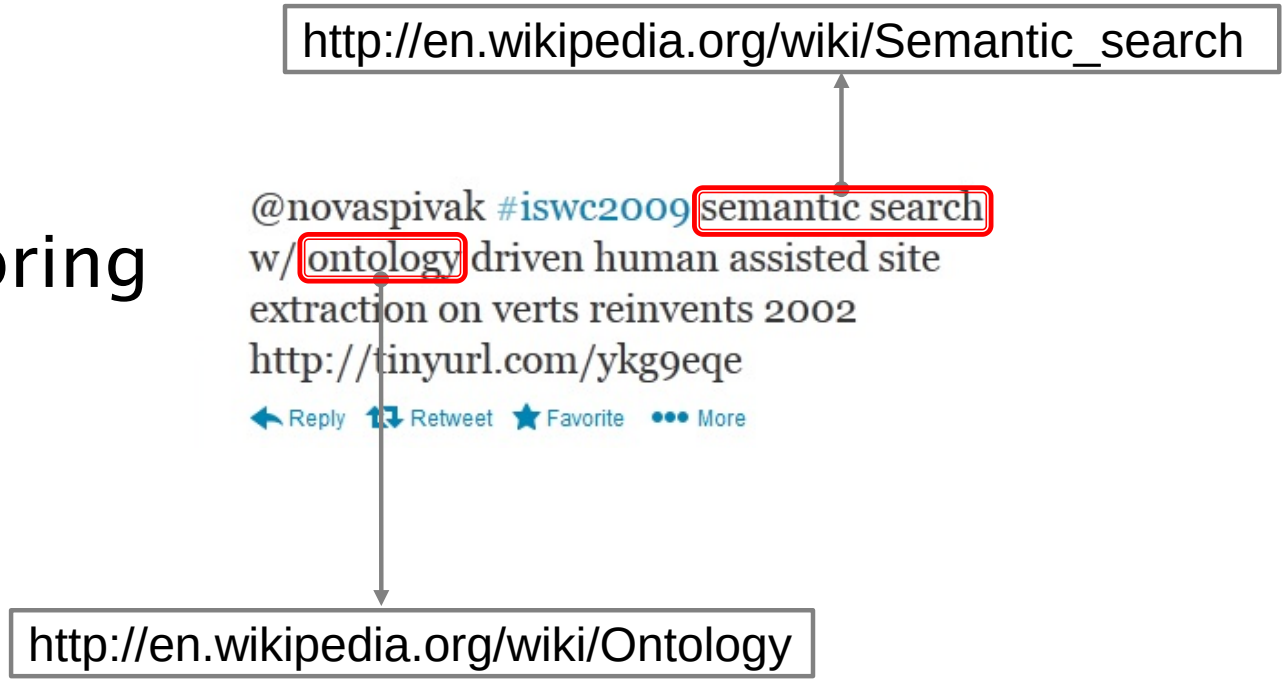
User Interests Generator

- Extracting Wikipedia concepts from Tweets

- Interests Scoring

User Interests Generator

- Extracting Wikipedia concepts from Tweets
- Interests Scoring



Entity Extraction on Tweets*

	Precision	Recall	F-measure	Usability	Rate Limit License
Dbpedia Spotlight	20.1	47.5	28.3	Inhouse+Web Service	N/A Apache 2.0
Text Razor	**64.6**	26.9	38.0	Web Service	500/day
Zemanta	57.7	31.8	**41.0**	Web Service	**10000/day**

- Issues relevant to entity extraction are handled by the web services
  - Stop words removal, URLs, Disambiguation etc.

*L. Derczynski, D. Maynard, N. Aswani and K. Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT '13.

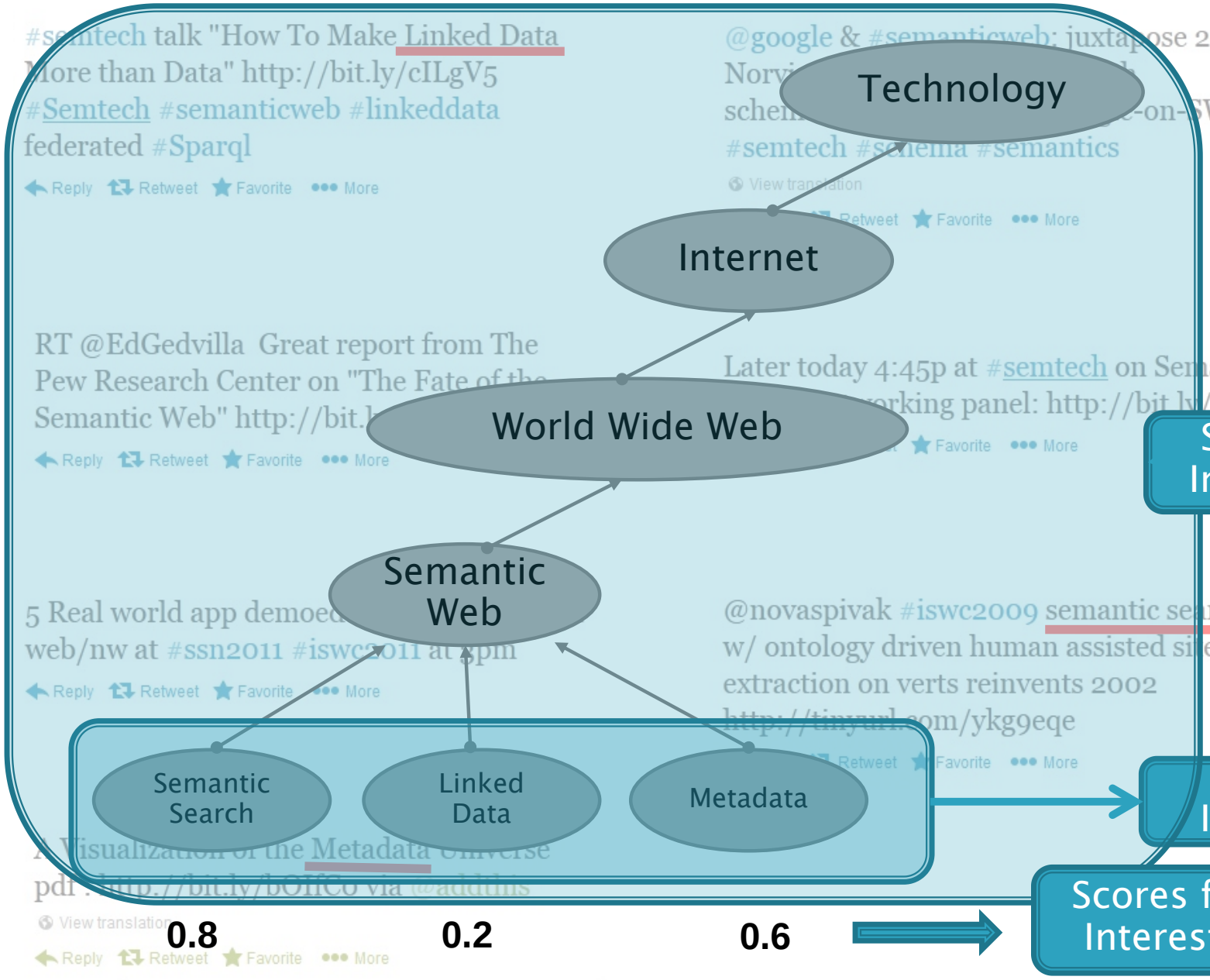
Scoring Interests

- Scoring Wikipedia concepts

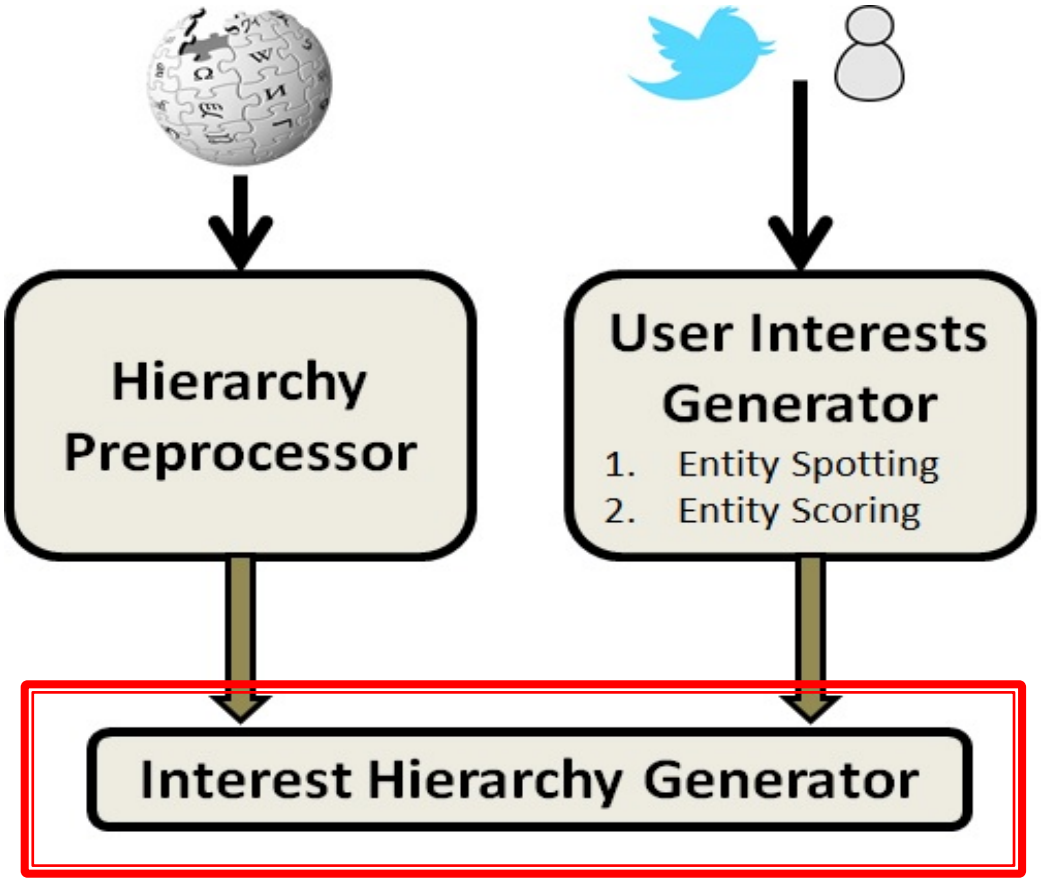
$$nf_i = \frac{frequency(e_i)}{frequency(e_{max})}$$

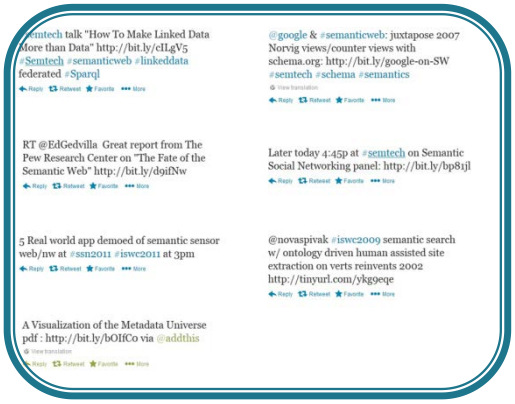
e_i -- entity to be scored

e_{max} -- entity mentioned most number of times by the user

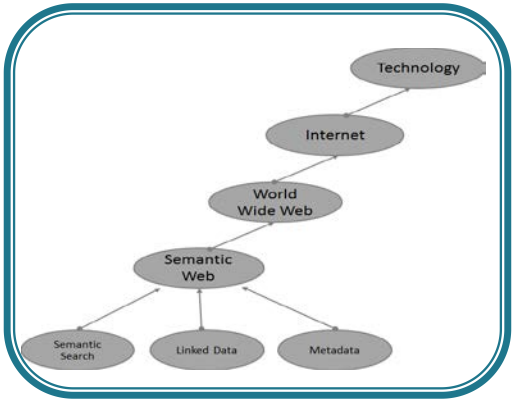

Technology
Internet
World Wide Web
Semantic Web
Semantic Search
Linked Data
Metadata
0.8
0.2
0.6
Structured Information
User Interests
Scores for Interests

Approach



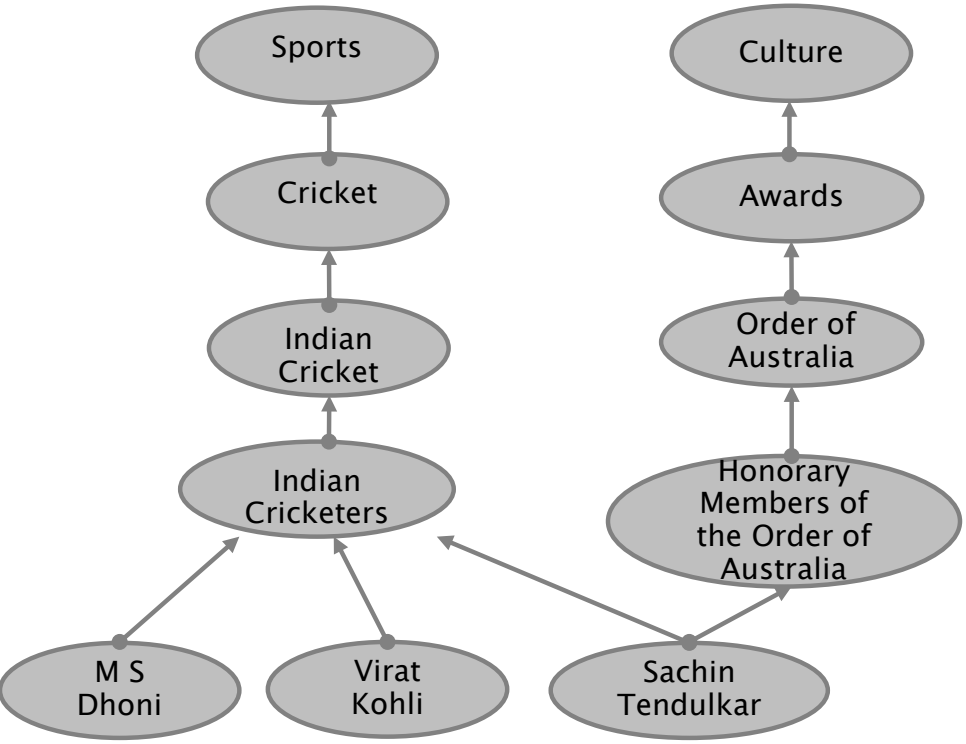


Tweets



Interest Hierarchy

Naïve Strategy – Inferring every Hierarchical Interest

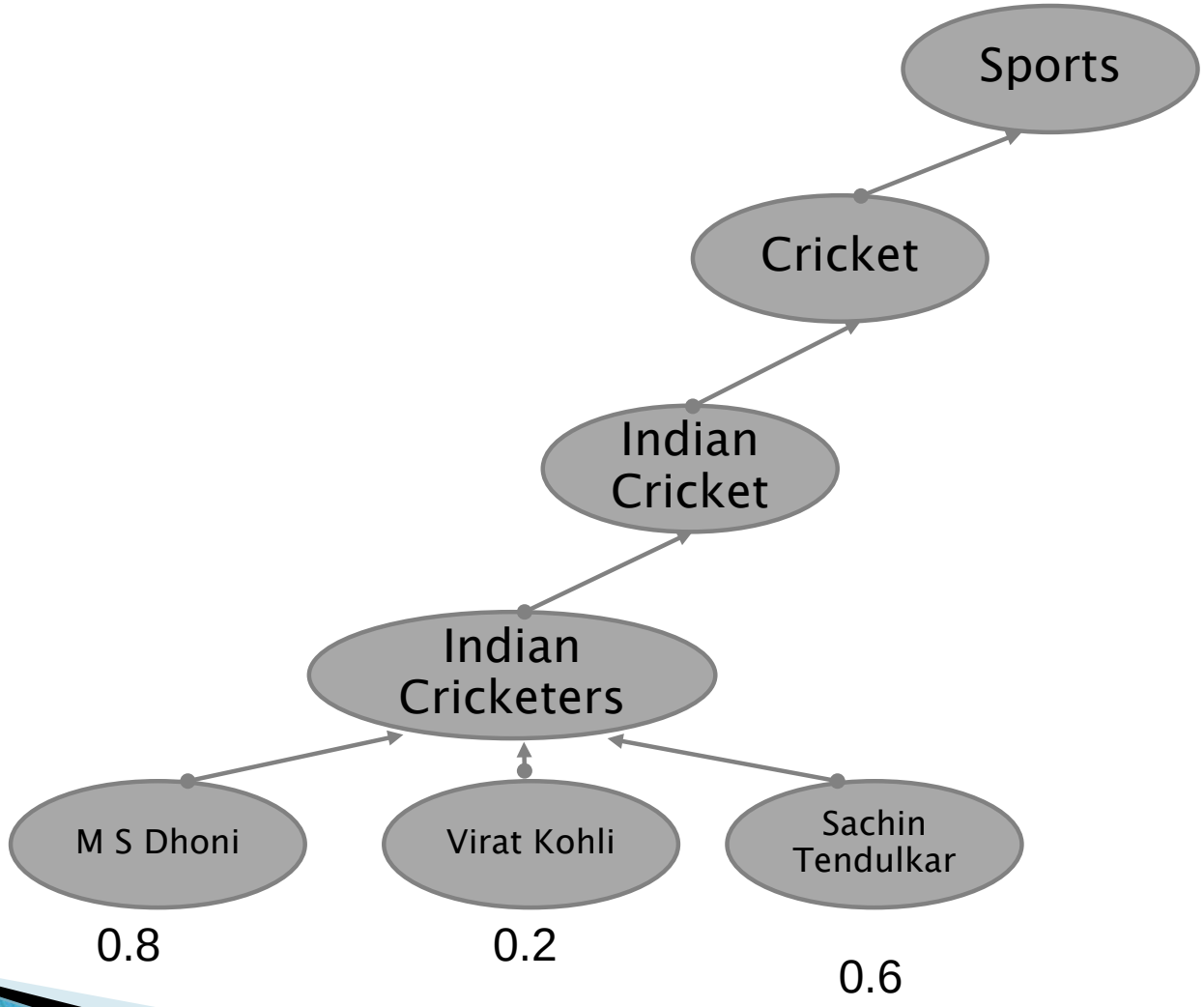


- Result (Challenges)
  - Infer more categories without context
  - Equal weights regardless Interest Score
  - Cannot rank categories of Interest for a user
  - *We use **Spreading Activation***

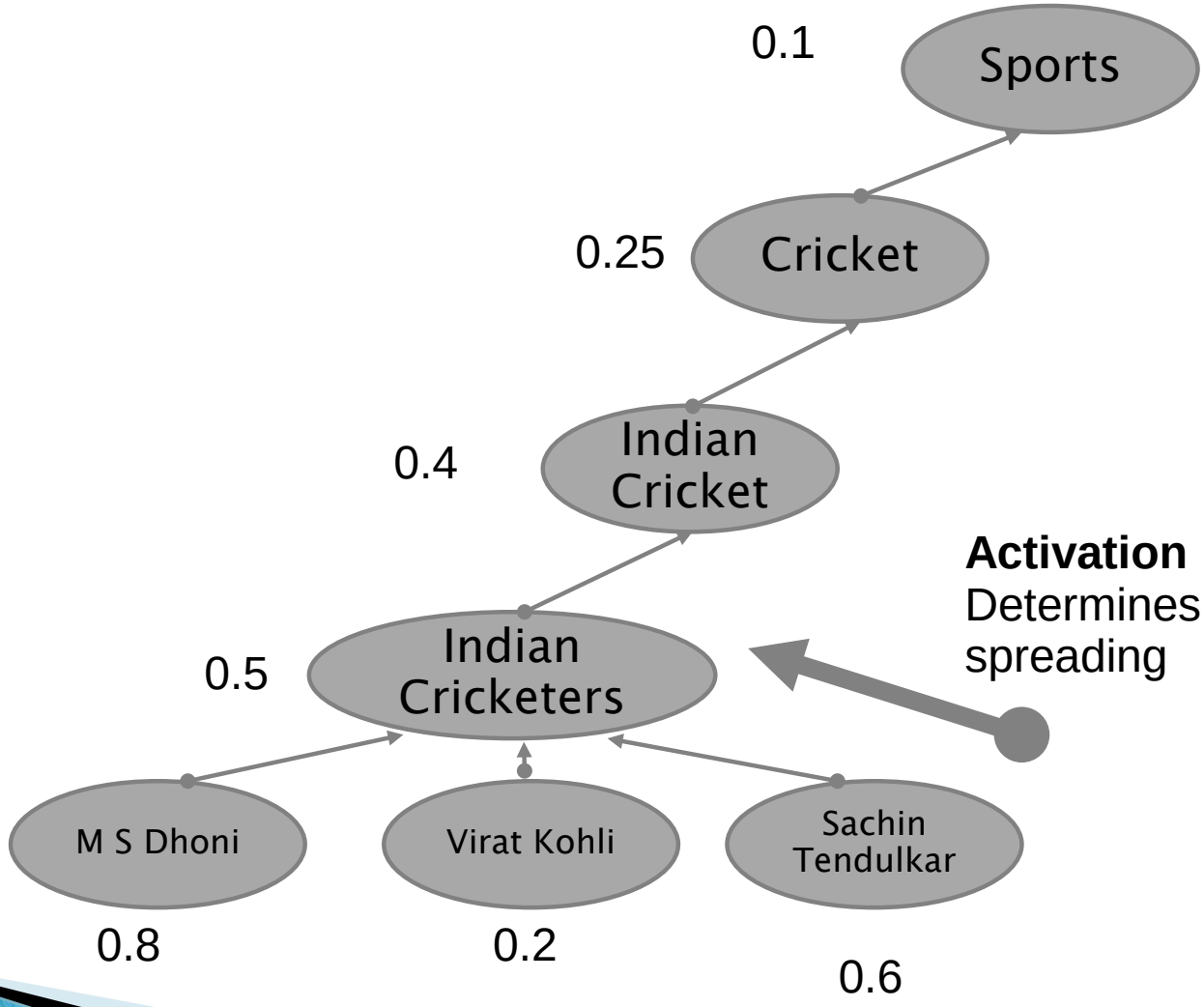
Spreading Activation

- Graph Algorithm to find contextual nodes
  - Cognitive Sciences
  - Neural Networks
  - Information Retrieval
    - Associative, Semantic Networks
  - Semantic Web
    - Context Generation

Spreading Activation Example



Spreading Activation Example



Activation Function
Determines the extent of spreading

Activation Function

- Simple Activation Function

$$A_j = \sum_{i=1}^{n} A_i \times W_{ij} \times D$$

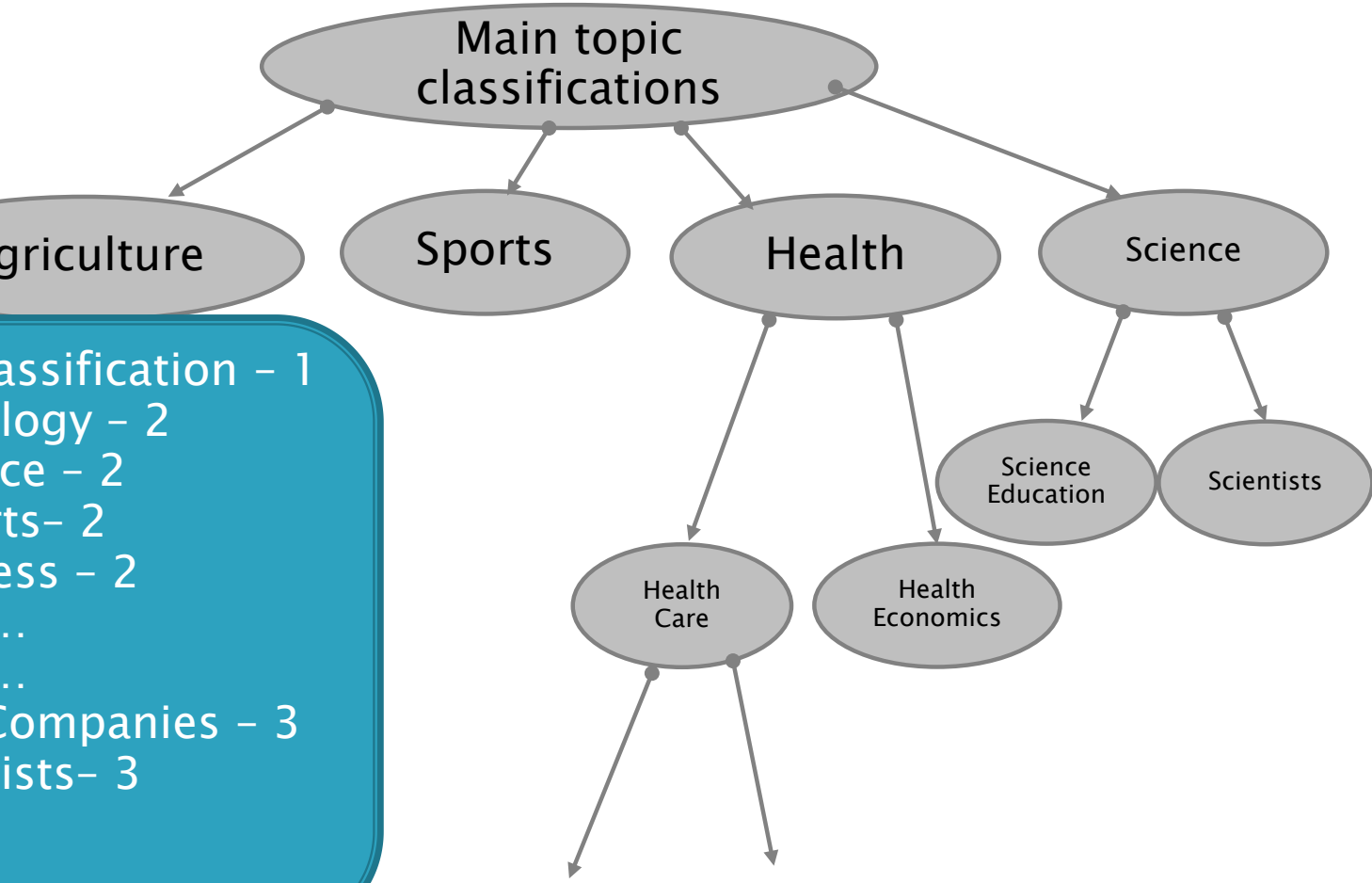
i *is the child or subcategory of* j $(Activated)$.
j *is the category to be activated.*
W_{ij} *is the edge weight between* j *and* i.
D *is the decay factor.*

Initial Experiments with Decay & Weighted Edge

- No Decay – No Weighted Edge
  - Result: Most generic categories ranked higher
- Decays over the hops of the activation
  - 0.4, 0.6, 0.8
  - Result: Same as above

Results: Constant Decay

Level: 1



Main Topic Classification – 1
Technology – 2
Science – 2
Sports– 2
Business – 2
…
…
Technology Companies – 3
Scientists– 3

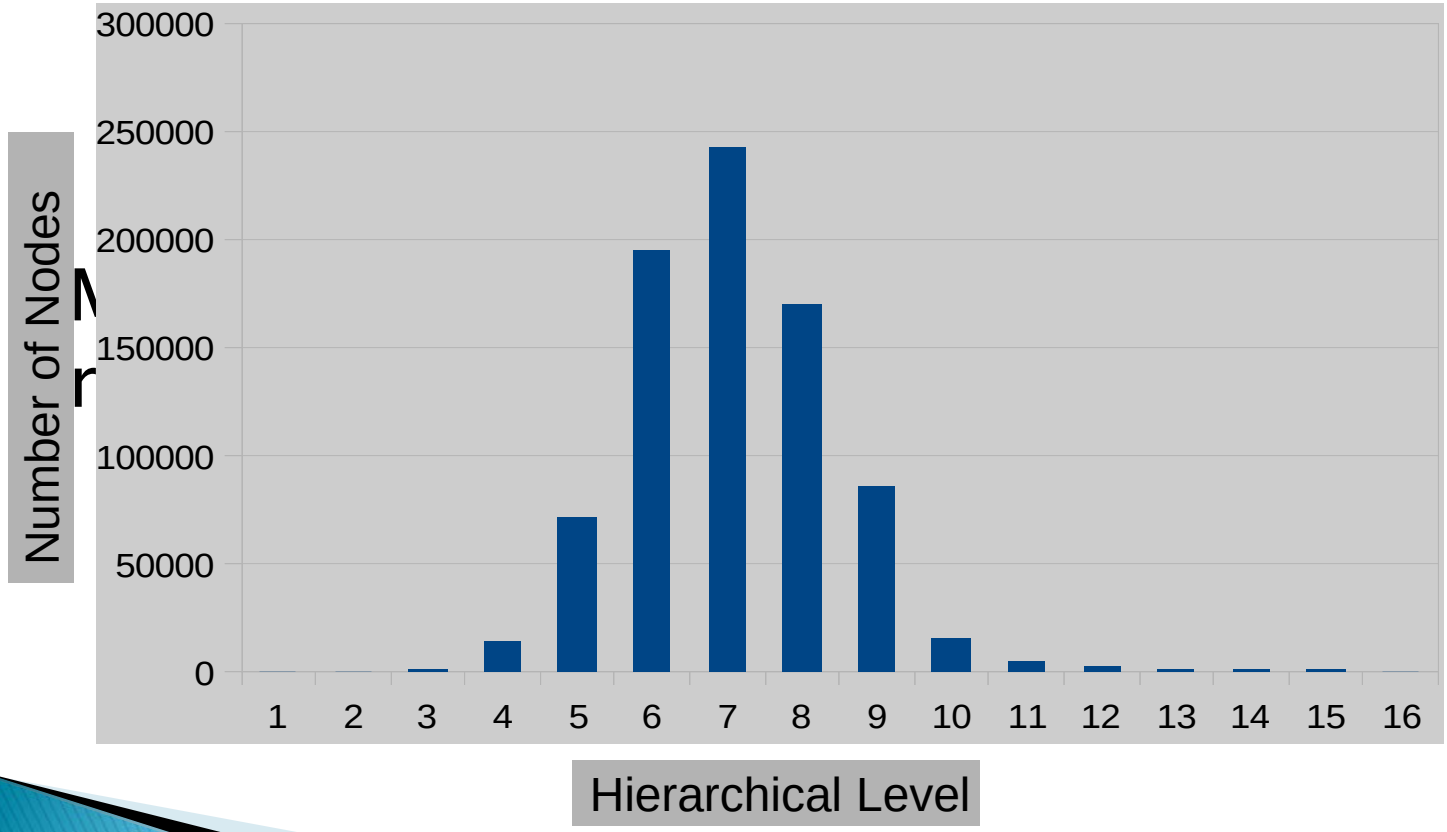
Wikipedia Challenges to Find Relevant Nodes in the Hierarchy

- Uneven distribution of nodes in the hierarchy

- Many-many for category-subcategory relationships

Wikipedia Challenges to Find Relevant Nodes in the Hierarchy

- Uneven distribution of nodes in the hierarchy



Wikipedia Challenges to Find Relevant Nodes in the Hierarchy

- Uneven distribution of nodes in the hierarchy

- Many-many for category-subcategory relationships

Wikipedia Challenges to Find Relevant Nodes in the Hierarchy

- Uneven distribution of nodes in the hierarchy

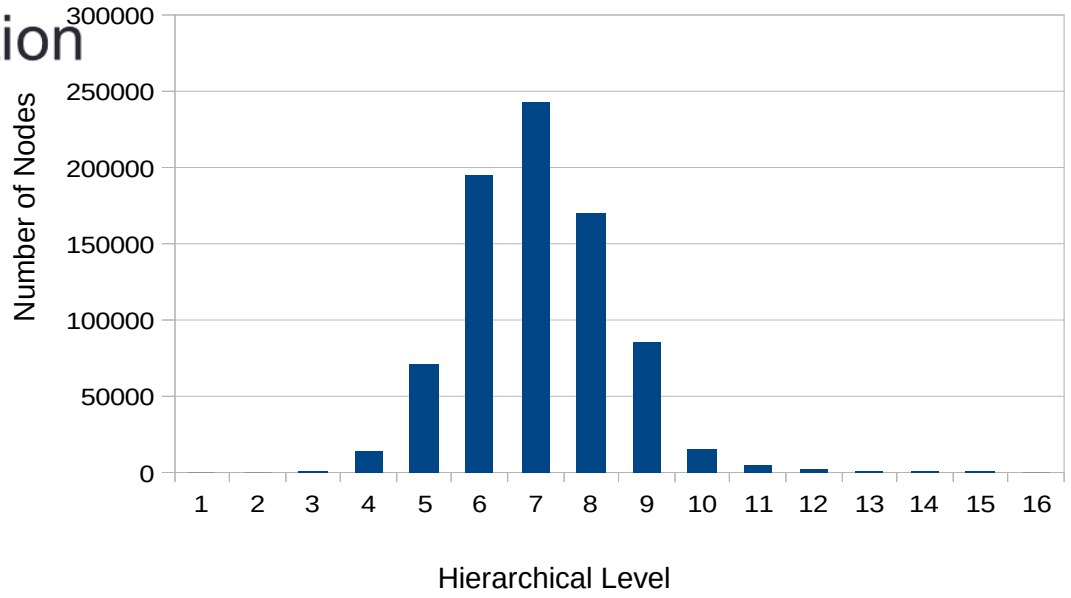
Addressing Uneven Nodes Distribution

- Bell Curve Normalization
  - Linear Fanout

$$F_{n_i} = {}^{1}/_{nodes_{(i+1)}}$$

  - Log Fanout

$$FL_{n_i} = {}^{1}/_{\log_{10} nodes_{(i+1)}}$$



Where $nodes_i$ is the number of nodes at hierarchical level i

Many to Many – Preferential Path

Article | Talk | Read | View sou

Football

From Wikipedia, the free encyclopedia

For other uses, see Football (disambiguation).

Football refers to a number of sports that involve, to varying degrees, kicking a ball with the foot to score a goal. The most popular of these sports worldwide is association footba commonly known as just "football" or "soccer". Unqualified, the word *football* applies to whichever form of football is the most popular in the regional context in which the word app including association football, as well as American football, Australian rules football, Canadian football, Gaelic football, rugby league, rugby union,[1] and other related games. The variations of football are known as football codes.

Various forms of football can be identified in history, often as popular peasant games. Contemporary codes of football can be traced back to the codification of these games at Eng public schools in the eighteenth and nineteenth century.[2][3] The influence and power of the British Empire allowed these rules of football to spread to areas of British influence ou the directly controlled Empire,[4] though by the end of the nineteenth century, distinct regional codes were already developing: Gaelic Football, for example, deliberately incorporat rules of local traditional football games in order to maintain their heritage.[5] In 1888, The Football League was founded in England, becoming the first of many professional football competitions. During the twentieth century, the various codes of football became amongst the most popular team sports in the world.[6]

Categories: Ball games | Football | Sports originating in England | 19th-century introductions

- Preferential Path Constraint

$$P_{n_{ij}} = {}^{1}/_{priority_{ji}}$$

Where $priority_{ji}$ is the priority of category i for subcategory j

Many to Many – Preferential Path

Article | Talk | Read | View sou

Football

From Wikipedia, the free encyclopedia

For other uses, see Football (disambiguation).

Football refers to a number of sports that involve, to varying degrees, kicking a ball with the foot to score a goal. The most popular of these sports worldwide is association footba commonly known as just "football" or "soccer". Unqualified, the word *football* applies to whichever form of football is the most popular in the regional context in which the word app including association football, as well as American football, Australian rules football, Canadian football, Gaelic football, rugby league, rugby union,[1] and other related games. The variations of football are known as football codes.

Various forms of football can be identified in history, often as popular peasant games. Contemporary codes of football can be traced back to the codification of these games at Eng public schools in the eighteenth and nineteenth century.[2][3] The influence and power of the British Empire allowed these rules of football to spread to areas of British influence ou the directly controlled Empire,[4] though by the end of the nineteenth century, distinct regional codes were already developing: Gaelic Football, for example, deliberately incorporat rules of local traditional football games in order to maintain their heritage.[5] In 1888, The Football League was founded in England, becoming the first of many professional football competitions. During the twentieth century, the various codes of football became amongst the most popular team sports in the world.[6]

Categories: Ball games | Football | Sports originating in England | 19th-century introductions

- Preferential Path Constraint

$$P_{n_{ij}} = {}^{1}/_{priority_{ji}}$$

Where $priority_{ji}$ is the priority of category i for subcategory j

Many to Many – Preferential Path

Article | Talk | Read | View sou

Football

From Wikipedia, the free encyclopedia

For other uses, see Football (disambiguation).

Football refers to a number of sports that involve, to varying degrees, kicking a ball with the foot to score a goal. The most popular of these sports worldwide is association footba commonly known as just "football" or "soccer". Unqualified, the word *football* applies to whichever form of football is the most popular in the regional context in which the word app including association football, as well as American football, Australian rules football, Canadian football, Gaelic football, rugby league, rugby union,[1] and other related games. The variations of football are known as football codes.

Various forms of football can be identified in history, often as popular peasant games. Contemporary codes of football can be traced back to the codification of these games at En public schools in the eighteenth and nineteenth century.[2][3] The influence and power of the British Empire allowed these rules of football to spread to areas of British influence ou the directly controlled Empire,[4] though by the end of the nineteenth century, distinct regional codes were already developing: Gaelic Football, for example, deliberately incorporat rules of local tr nal footb mes in order t ntain their heritage.[5] In 18 The Football League was founded in England, becoming the first of many professional football competitions. D the twe century, the v s codes of football became gst the most popular team sports in the world.[6]

1 2 3 4

Categories: Ball games | Football | Sports originating in England | 19th-century introductions

- Preferential Path Constraint

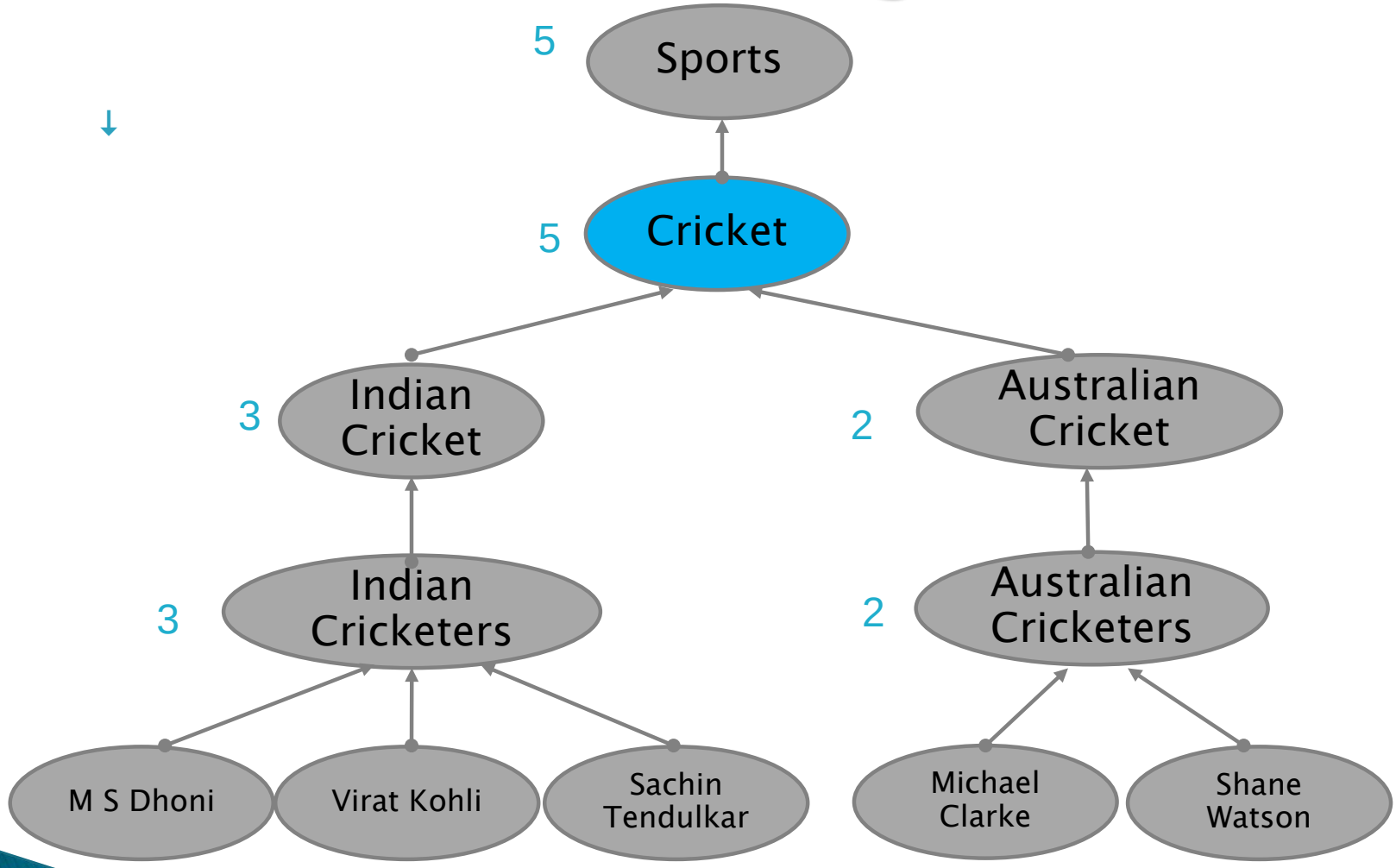
$$P_{n_{ij}} = {}^{1}/_{priority_{ji}}$$

Where $priority_{ji}$ is the priority of category i for subcategory j

Boost Intersecting Nodes

- Nodes that intersect domains/subcategories activated by diverse entities

Boost Intersecting Nodes



Boost Intersecting Nodes

- Nodes that intersect domains/subcategories activated by diverse entities
  - Intersect Booster

$$B_i = {}^{N_{e_i}}/_{N_{e_{icmax}}}$$

N_{e_i} is the number of entities activating e_i

e_{icmax} is the subcategory of i which is activated by maximum number of entities

- Example

$$B_{cricket} = {}^{5}/_{3}$$

Activation Functions

- Bell

$$A_j = \sum_{i=0}^{n} A_i \times F_j$$

- Bell Log

$$A_j = \sum_{i=0}^{n} A_i \times FL_j$$

- Priority Intersect

$$A_j = \sum_{i=0}^{n} A_i \times FL_j \times P_{ji} \times B_j$$

Agenda

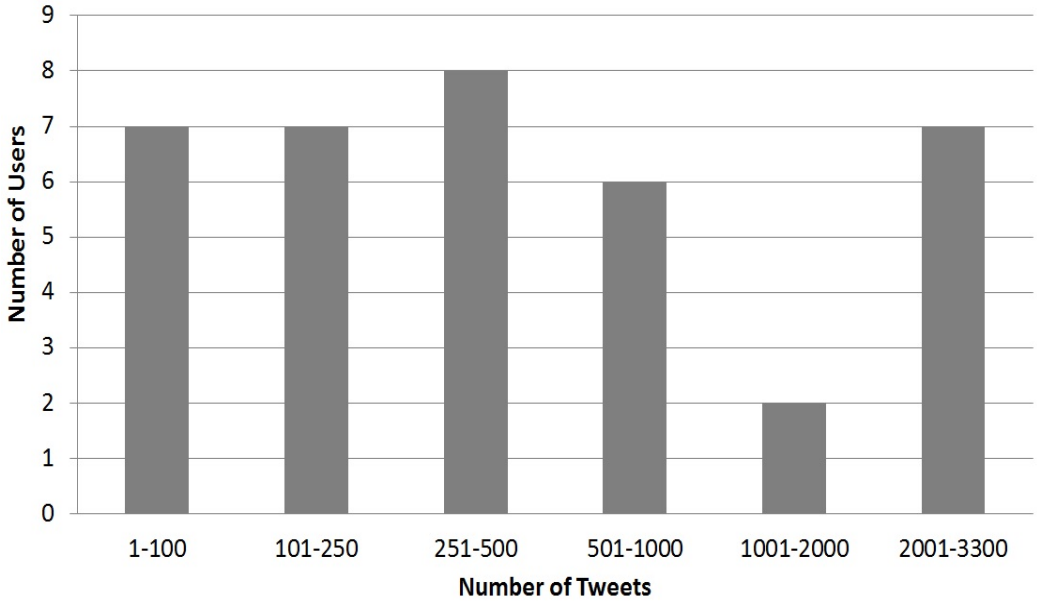
- Motivation
- Approach
- **Evaluation**
- Conclusion & Future Work

User Study

- User Study Data
  - 37 Users
  - 31927 Tweets

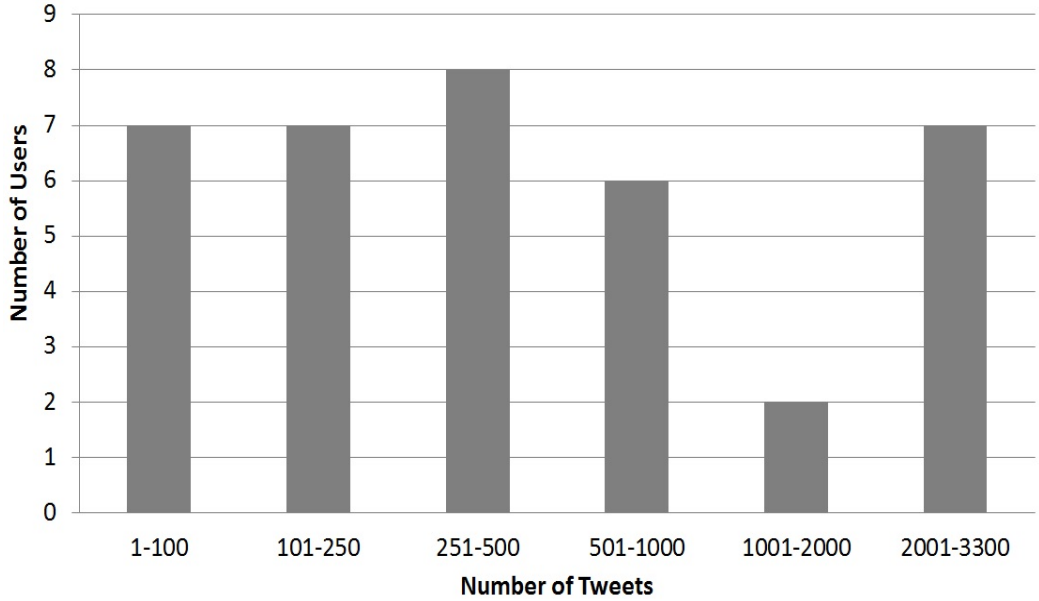
User Study

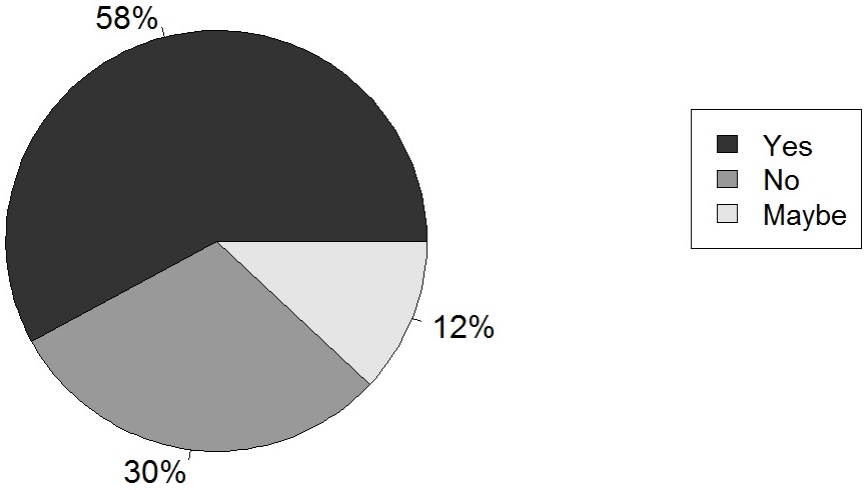
- User Study Data
  - 37 Users
  - 31927 Tweets



User Study

- User Study Data
  - 37 Users
  - 31927 Tweets



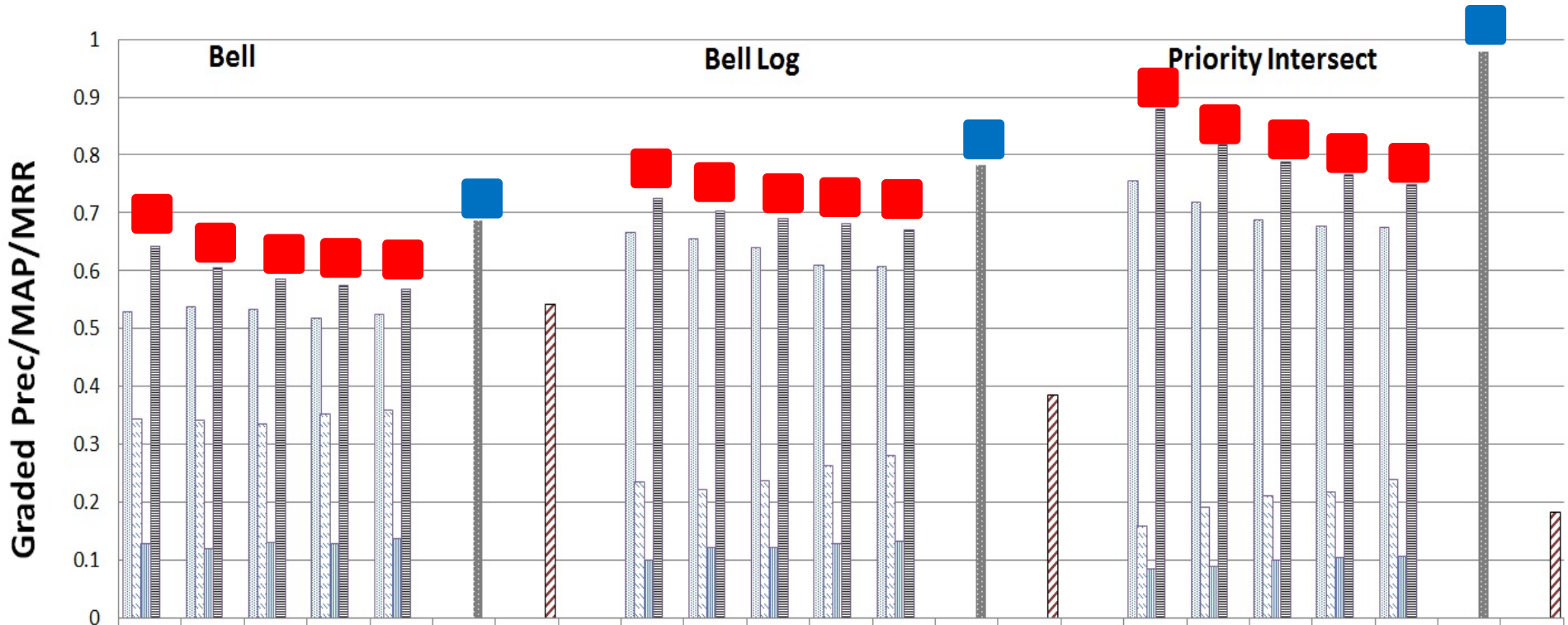


- Hierarchical Interest Graph
  - 111,535 Category Interests.
  - 3000 Categories/user
  - Ranking Evaluation -- Top-50 Categories.

Evaluation Metrics

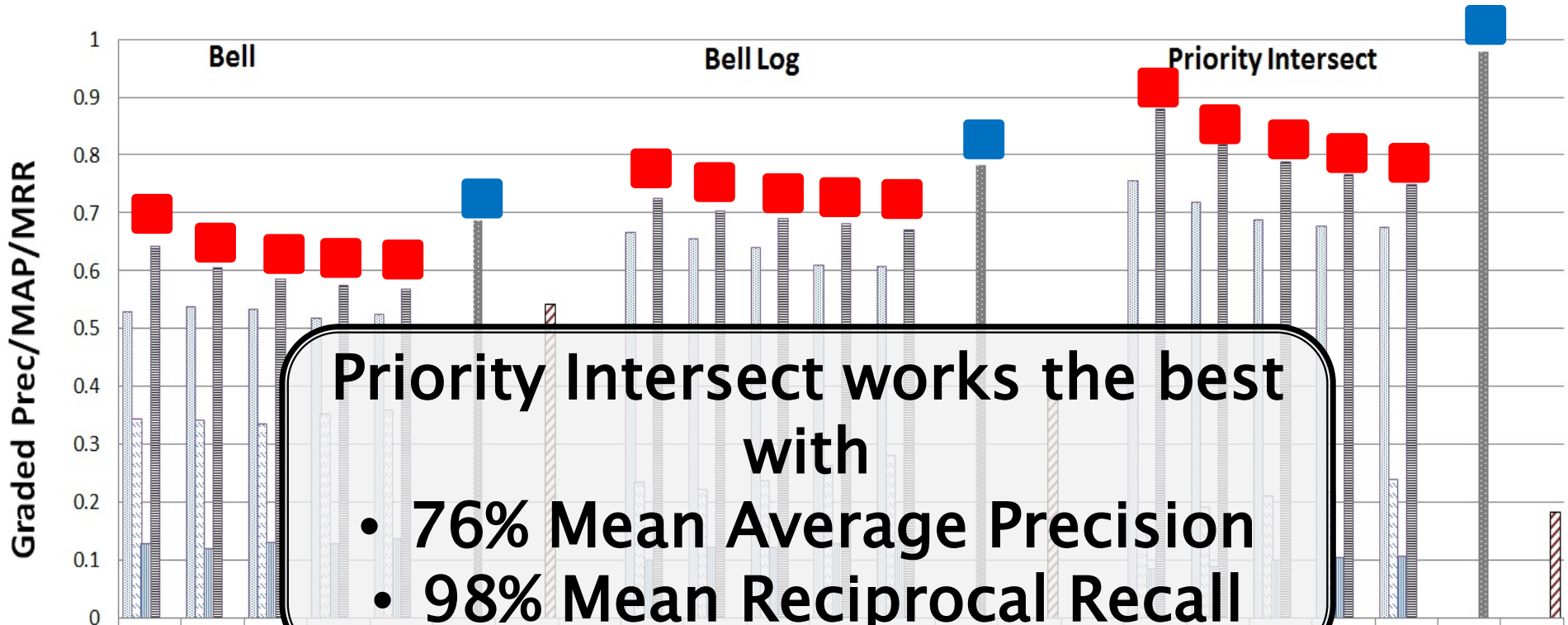
- How many relevant/irrelevant Hierarchical Interests are retrieved at top-k ranks?
  - Graded Precision
- How well are the retrieved relevant Hierarchical Interests ranked at top-k?
  - Mean Average Precision
- How early in the ranked Hierarchical Interests can we find a relevant result?
  - Mean Reciprocal Recall

Evaluation Results



	@10	@20	@30	@40	@50	MRR Y	MRR N		@10	@20	@30	@40	@50	MRR Y	MRR N		@10	@20	@30	@40	@50	MRR Y	MRR N
yes	0.53	0.54	0.53	0.52	0.52				0.67	0.66	0.64	0.61	0.61				0.76	0.72	0.69	0.68	0.67		
no	0.34	0.34	0.34	0.35	0.36				0.23	0.22	0.24	0.26	0.28				0.16	0.19	0.21	0.22	0.24		
maybe	0.13	0.12	0.13	0.13	0.14				0.1	0.12	0.12	0.13	0.13				0.08	0.09	0.1	0.11	0.11		
MAP	0.64	0.61	0.59	0.58	0.57				0.72	0.7	0.69	0.68	0.67				0.88	0.82	0.79	0.77	0.75		
MRRY						0.69								0.78								0.98	
MRRN							0.54								0.39								0.18

Evaluation Results



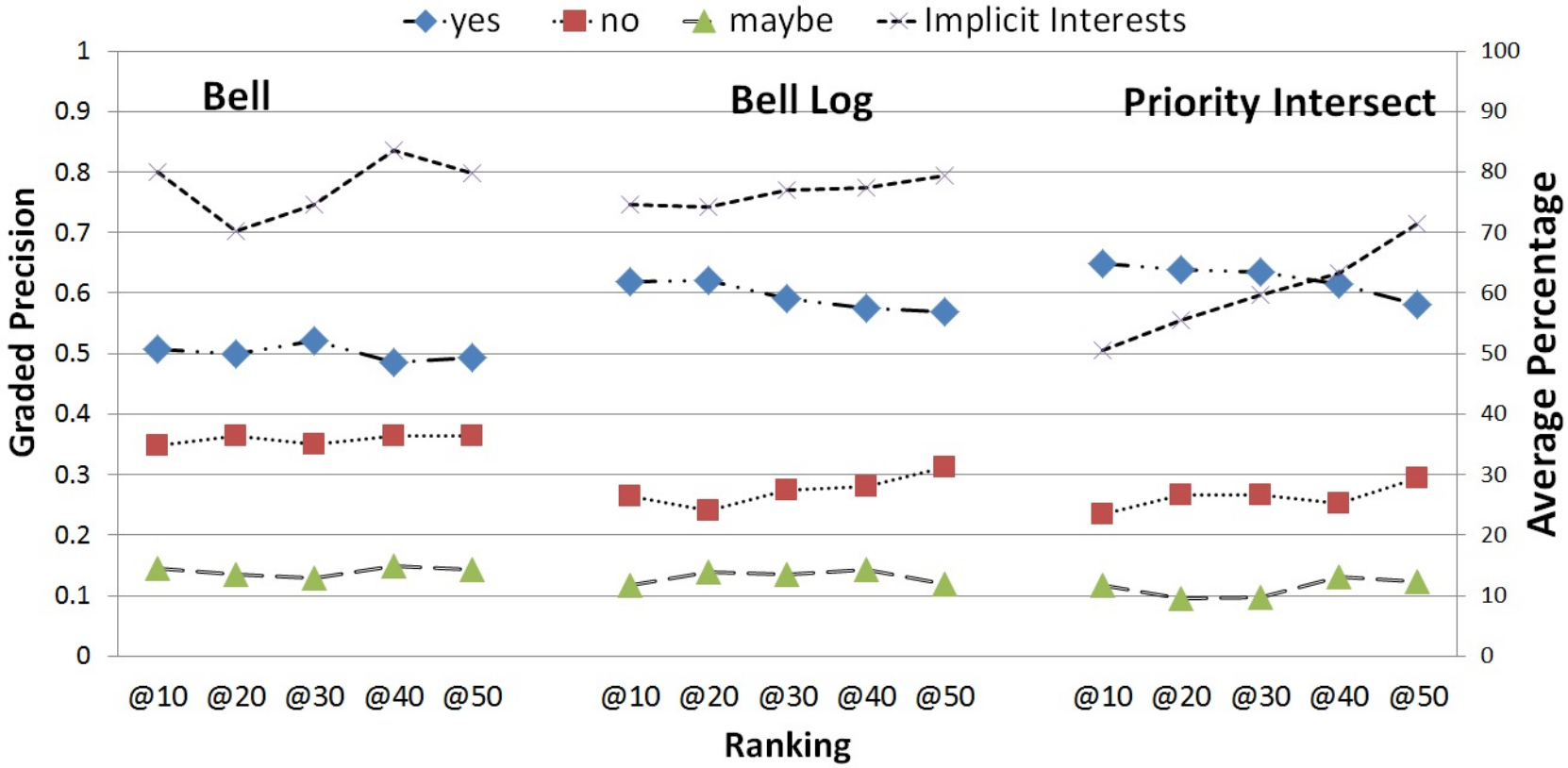
Priority Intersect works the best with

- **76% Mean Average Precision**
- **98% Mean Reciprocal Recall**

	@10	@20	@30	@40	@50	MRR Y	MRR N		@10	@20	@30	@40	@50	MRR Y	MRR N		@10	@20	@30	@40	@50	MRR Y	MRR N
yes	0.53	0.54	0.53	0.52	0.52				0.67	0.66	0.64	0.61	0.61				0.76	0.72	0.69	0.68	0.67		
no	0.34	0.34	0.34	0.35	0.36				0.23	0.22	0.24	0.26	0.28				0.16	0.19	0.21	0.22	0.24		
maybe	0.13	0.12	0.13	0.13	0.14				0.1	0.12	0.12	0.13	0.13				0.08	0.09	0.1	0.11	0.11		
MAP	0.64	0.61	0.59	0.58	0.57				0.72	0.7	0.69	0.68	0.67				0.88	0.82	0.79	0.77	0.75		
MRRY						0.69								0.78								0.98	
MRRN							0.54								0.39								0.18

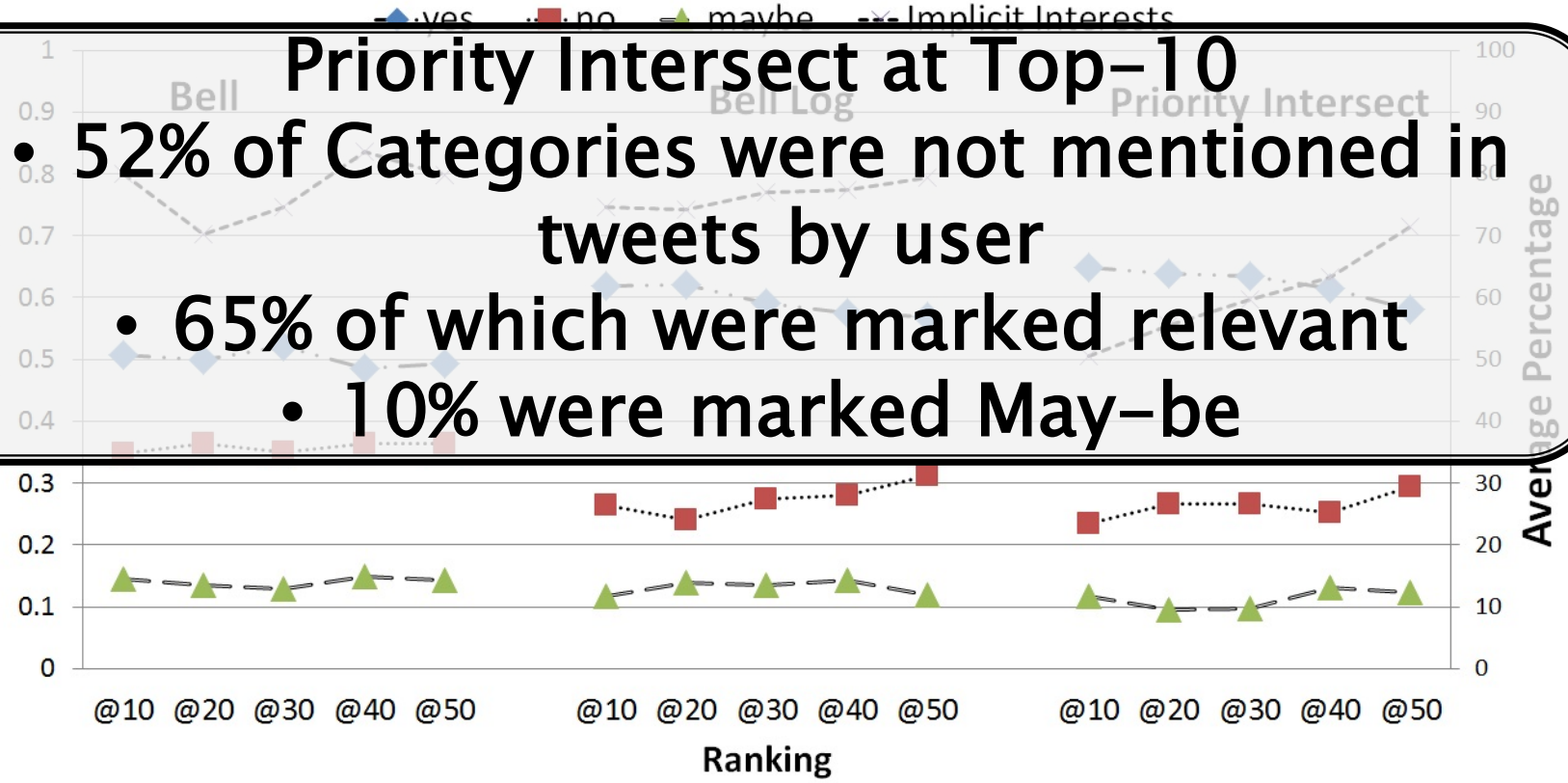
Implicit Interests – Syntactic

- How many of the categories inferred by the system were not explicitly mentioned by the user in tweets? (Semantic Web and Category:Semantic Web)



Implicit Interests – Syntactic

- How many of the categories inferred by the system were not explicitly mentioned by the user in tweets? (Semantic Web and Category:Semantic Web)



Priority Intersect at Top-10

- **52% of Categories were not mentioned in tweets by user**
- **65% of which were marked relevant**
- **10% were marked May-be**

Hierarchy Evaluation

- Mapped (String match) categories of Wikipedia to Dmoz.
  - ~141K categories mapped
- Compared all the category and sub-category relationships of the mapped categories in the hierarchy to manually created Dmoz.
  - 87% precise (in hierarchy were also found in Dmoz)

Agenda

- Motivation
- Approach
- Evaluation
- **Conclusion & Future Work**

Conclusion

- Hierarchical Interest Graph (Hierarchy representation of user interests)
  - With hierarchical levels of each interest to have flexibility for personalizing and recommending based on its abstractness.
- We semantically enhanced user profiles of interests from Twitter using Knowledge bases.
  - Inferred abstract/hierarchical interests of Twitter users using Wikipedia
  - This can help reducing the data sparcity problem by inferring relevant interests.
- The top-1 hierarchical-interest generated by the system was correct for 36 out of 37 user-study participants.
  - Mean Average Precision at Top-10 is 0.76

Future Work

- Measuring impact of Hierarchical Interest Graphs for recommendation of Movies/Music
  - Datasets
    - Movielens
    - Lastfm
- Tuning the system to utilize the hierarchical levels of interests for personalization and recommendation
  - Sports (most abstract interest)
  - Baseball (specific interest)

Thanks

More info: Knoesis Wiki – Hierarchical Interest Graph

Paper at: http://j.mp/user-ig

Contact: Pavan Kapanipathi
Twitter:@pavankaps
Email: pavan@knoesis.org