Using Lexical Cooccurrence Structures for Identifying the Semantic Siblings of a Set of Entities

Aditya Ramana Rachakonda
aditya.ramana@iiitb.ac.in

Open Systems Laboratory
International Institute of Information Technology, Bangalore.

December 19
COMAD

Cooccurrence → Semantics

- Human interactions are not through words but through word meanings and associated semantics.
- Systems with human interactions should recognise and address this difference.
- Information Retrieval
  - Assumption: Words are independent of each other.
  - Fact: It is not so.
  - The dependencies between words enable us to model semantics.
- Cognitive Sciences
  - Semantic memory is made of co-activations.
  - Hebbian Learning: "Cells that fire together, wire together."
  - Meaning is usage.
- Existing: Entity Relatedness, Synonyms

Semantic Siblings

- Entities with the same semantic super class in the concept hierarchy
- Example: *Roger Federer, Rafael Nadal, Andy Roddick*
- Conceptual set of entities
- Synonyms?

Semantic Siblings

- Entities with the same semantic super class in the concept hierarchy
- Example: *Roger Federer, Rafael Nadal, Andy Roddick*
- Conceptual set of entities
- Synonyms?

- Problem: Given a few such entities can an algorithm identify *similar* entities from the *same set*?

Semantic Siblings

- Entities with the same semantic super class in the concept hierarchy
- Example: *Roger Federer, Rafael Nadal, Andy Roddick*
- Conceptual set of entities
- Synonyms?

- Problem: Given a few such entities can an algorithm identify *similar* entities from the *same set*?
- **Real** Problem: No type information

Semantic Siblings

- Problem: Given a few such entities can an algorithm identify *similar* entities from the *same set*?
  - *similar*, *same set*

Semantic Siblings

- Problem: Given a few such entities can an algorithm identify *similar* entities from the *same set*?
  - *similar*, *same set*
- Example: *Federer hit three aces in the last game*
  - *Roddick hit three aces in the last game*
  - *Nadal hit three aces in the last game*
  - *Obama hit three aces in the last game*
  - *Mickey Mouse hit three aces in the last game*
  - *Hammerhead Shark hit three aces in the last game*

Semantic Siblings

- Problem: Given a few such entities can an algorithm identify *similar* entities from the *same set*?
  - *similar*, *same set*
- Example: *Federer hit three aces in the last game*
  - *Roddick hit three aces in the last game*
  - *Nadal hit three aces in the last game*
  - *Obama hit three aces in the last game*
  - *Mickey Mouse hit three aces in the last game*
  - *Hammerhead Shark hit three aces in the last game*
- Replaceability
  - Grammatical replaceability $\neq$ Semantic replaceability
  - Equivalence realtion w.r.t. a context

Problem Definition

Definition

Given a *set of entities* and an *unknown context* in which the entities are *replaceable* by one another, identify more entities which share the same property.

Problem Definition

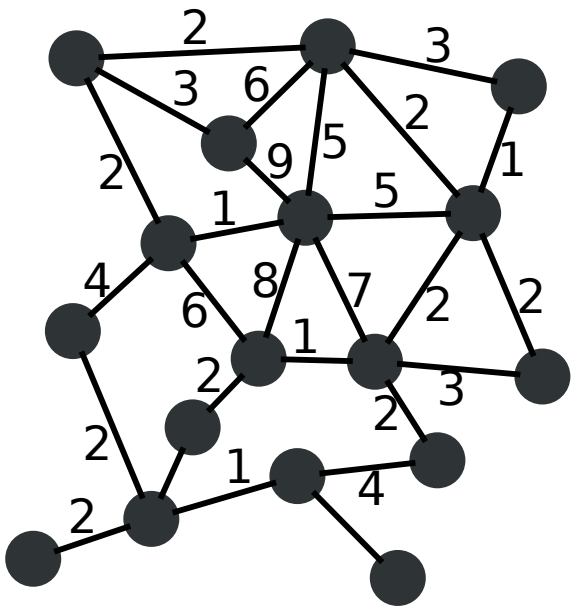
Definition

Given a *set of entities* and an *unknown context* in which the entities are *replaceable* by one another, identify more entities which share the same property.

Hypothesis

Given a set of terms $Q = \{q_1, q_2, \ldots, q_n\}$, and their respective contexts of occurrence $C(q_i)$, a semantic sibling s is a term which maximises the contextual overlap between its context $C(s)$ and each $C(q_i)$ — the context of an element in Q.

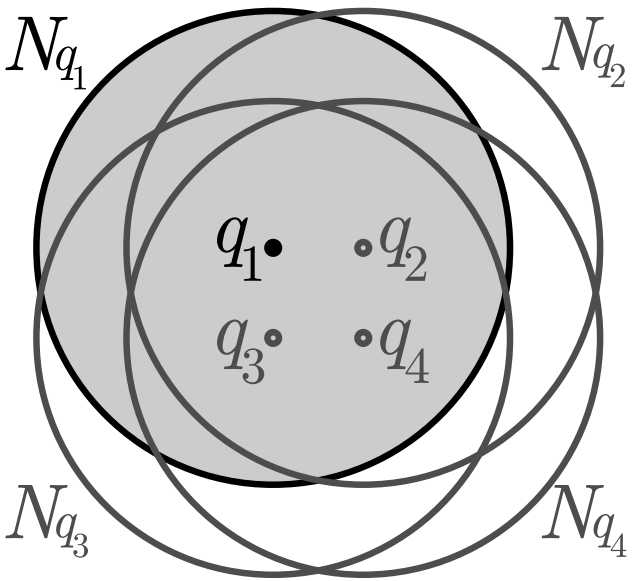
Cooccurrence Graphs



- Basic data structure for semantic siblings
- Undirected graph $G = (V, E, w)$
- V: Stemmed noun phrases
- E: Cooccurrences between the terms
- w: Edge weights indicating the number of times two terms cooccur in our universe

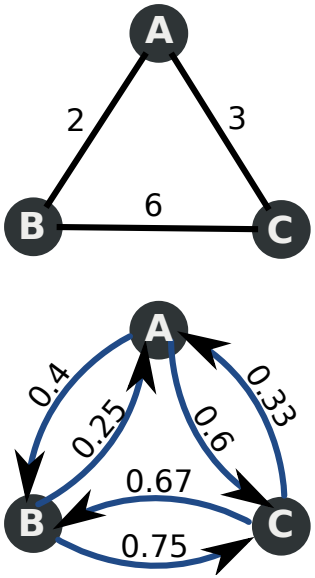
Definition (Context)

A **context** in a cooccurrence graph is any sub-graph, containing terms and cooccurrences, which can be used to describe a specific semantic universe.



- Terms (query terms) Q, are extracted from an information source
- Their neighbourhood – all terms which share an edge – is identified
- Overlap in neighbourhoods if the terms are from a coherent semantic context
- Nodes in the sub-graph $V_C = \bigcup_q N_q$, where $q \in Q$

Reachability



- Reachability: Probability that term b cooccurs with a,

$$\rho(a \rightarrow b) = \frac{e_{a,b}}{\sum_{\forall x \in N_a} e_{a,x}}$$

- $\rho(a)$ denotes the reachability distribution of a

Direct Comparison

- Roger Federer's distribution
- Rafael Nadal's distribution
- Andy Roddick's distribution
- Marat Safin's distribution

Interleaved Templates

Combined distribution of

- Roger Federer, Rafael Nadal, Andy Roddick
- **Marat Safin**, Rafael Nadal, Andy Roddick
- Roger Federer, **Marat Safin**, Andy Roddick
- Roger Federer, Rafael Nadal, **Marat Safin**

Rank

- For each term in the V_C
  - Compute a vector of scores
- Rank the terms by the magnitude of the vector.

Results I

Query: sapphire, emerald, topaz

gemstone 0.594
opal 0.593
amethyst 0.568
garnet 0.506
peridot 0.483
lapis lazuli 0.469
spinel 0.468
turquoise 0.431
beryl 0.416
onyx 0.415
pearl 0.405
gemstones 0.366
agate 0.347
corundum 0.334
tourmaline 0.320
sardonyx 0.316
crystal 0.314
moonstone 0.307
inclusion 0.294
diamond 0.279

Results II

Query: roger federer, rafael nadal, andy roddick

janko tipsarević 0.672
igor andreev 0.667
ivo karlović 0.664
potito starace 0.660
andreas seppi 0.658
arnaud clément 0.647
andrey golubev 0.646
mario ančić 0.642
jürgen melzer 0.637
mardy fish 0.635
marat safin 0.635
dmitry tursunov 0.627
marcos baghdatis 0.627
michael berrer 0.625
olivier rochus 0.625
jérémy chardy 0.623
paul-henri mathieu 0.623
josé acasuso 0.620
ernests gulbis 0.620
marcel granollers 0.620

Results III

Query: sachin tendulkar, sourav ganguly, rahul dravid

kapil dev 0.742
shahid afridi 0.740
mahendra singh dhoni 0.740
gautam gambhir 0.736
shoaib akhtar 0.733
yuvraj singh 0.733
zaheer khan 0.725
muttiah muralitharan 0.717
john wright 0.714
anil kumble 0.713
sunil gavaskar 0.709
irfan pathan 0.707
mohammad rafique 0.707
ravi shastri 0.704
sanath jayasuriya 0.702
greg chappell 0.699
mitchell johnson 0.697
sharjah 0.694
imran khan 0.693
kumar sangakkara 0.693