Context Aware Semantic Association Ranking

SWDB Workshop
Berlin, September 7, 2003

Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar, Amit Sheth
Large Scale Distributed Information Systems Lab
Computer Science Department, University of Georgia

This material is based upon work supported by the National Science Foundation under Grant No. 0219649.

From …..

Finding things

to…..

“Finding out about” [Belew00]

relationships!

Outline

- From Search to Analysis: Semantic Associations
- Using Context for Ranking
- Ranking Algorithm
- Preliminary Results / Demo
- Related Work
- Conclusion & Future Work

Changing expectations

- Not documents, not search, not even entities, but actionable information and insight
- Emergence of text/content analytics, knowledge discovery, etc. for business intelligence, national security, and other emerging markets

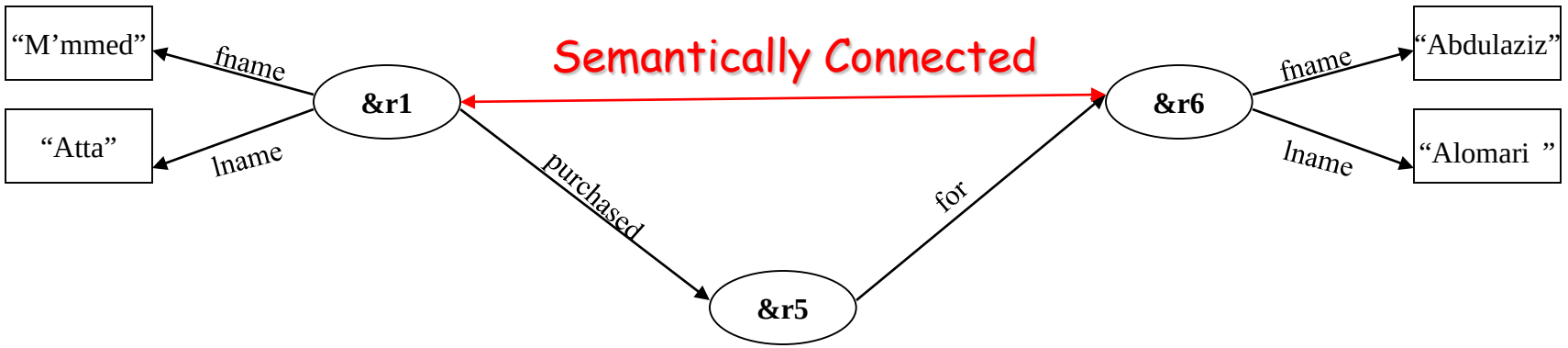
Example in 9-11 context

- What are relationships between Khalid Al-Midhar and Majed Moqed ?
  - *Connections*
    - Bought tickets using same frequent flier number
  - *Similarities*
    - Both purchased tickets originating from Washington DC paidby cash and picked up their tickets at the Baltimore-Washington Int'l Airport
    - Both have seats in Row 12
- "What relationships exist (if any) between Osama bin Laden and the 9-11 attackers"

Semantic Associations

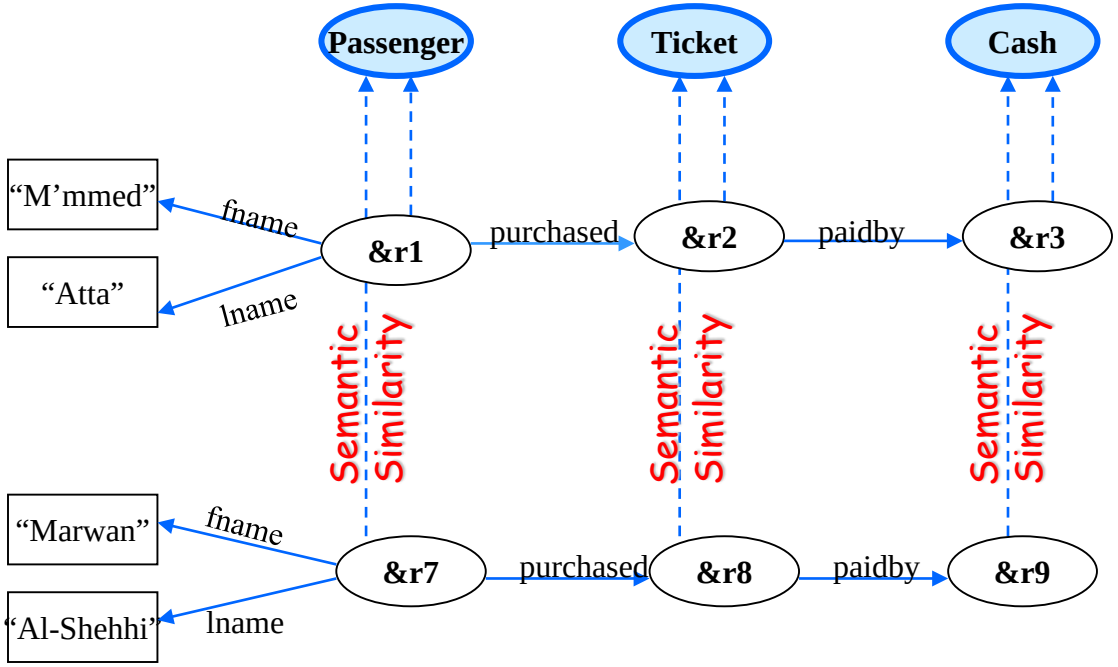
ρ - Association

- Two entities e_1 and e_n are semantically connected if there exists a sequence e_1, P_1, e_2, P_2, e_3, … e_{n-1}, P_{n-1}, e_n in an RDF graph where ei, $1 \leq i \leq n$, are entities and P_j, $1 \leq j < n$, are properties



σ - Association

- Two entities are semantically similar if both have ≥ 1 similar paths starting from the initial entities, such that for each segment of the path:
  - Property P*i* is either the same or subproperty of the corresponding property in the other path
  - Entity E*i* belongs to the same class, classes that are siblings, or a class that is a subclass of the corresponding class in the other path



The Need For Ranking

- Current test bed with > 6,000 entities and > 11,000 explicit relations
- The following semantic association query ρ(*"Nasir Ali", "AlQeada"*), results in 2,234 associations
- The results must be presented to a user in a relevant fashion…thus the need for ranking

Context Use For Ranking

Context: Why, What, How?

- Context => Relevance; Reduction in computation space
- Context captures the users' interest to provide the user with the relevant knowledge within numerous relationships between the entities
- By defining regions (or sub-graphs) of the ontology we are capturing the areas of interest of the user

Context Specification

- Topographic approach (current)
  - *Regions* ‘capture’ user’s interest, such as a *region* is a subset of classes (entities) and properties of an ontology
- View approach (future)
- Each *region* can have a relevance weight

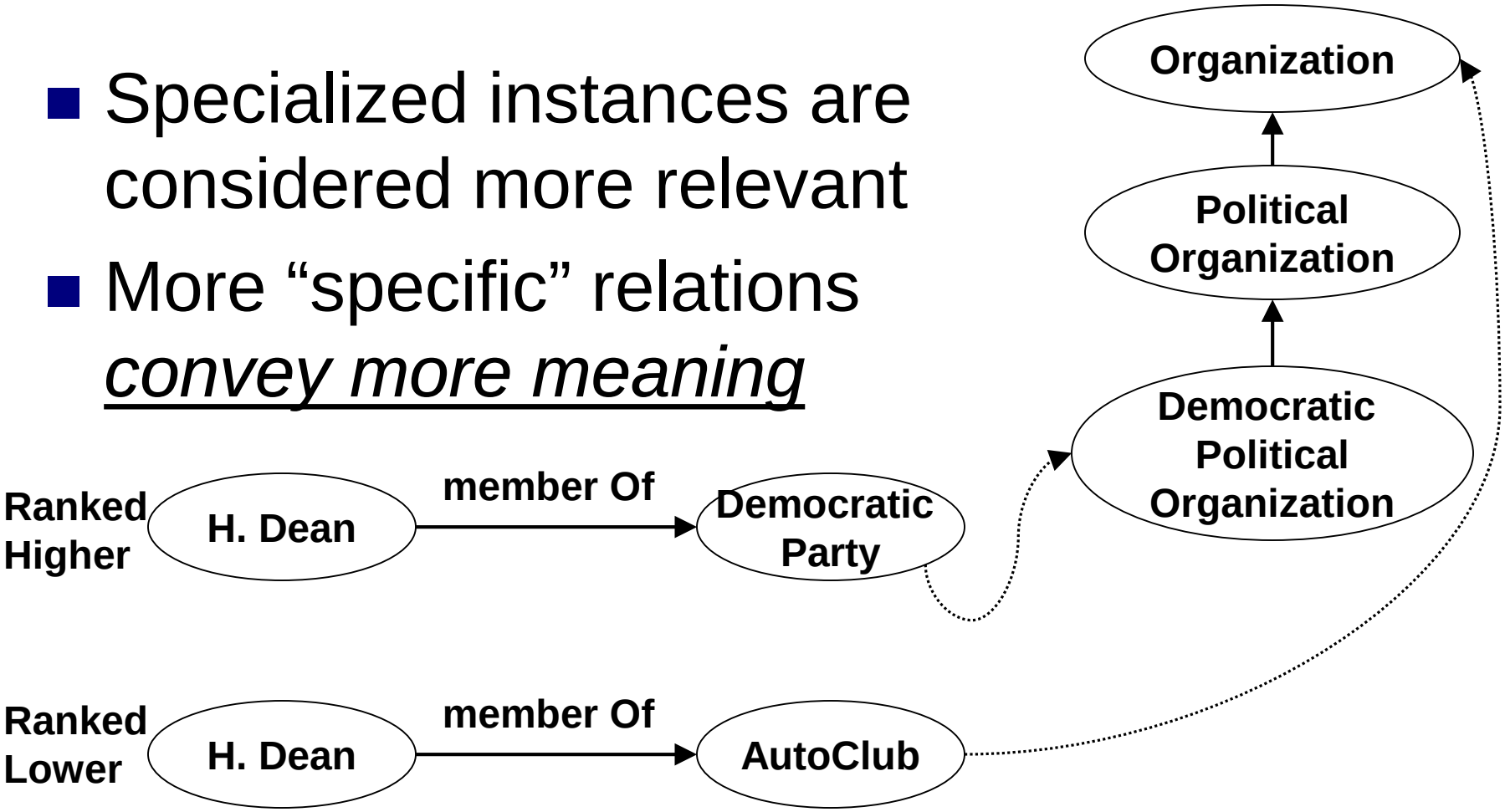
Ranking Algorithm

Ranking – Introduction

- Our ranking approach defines a path rank as a function of several ranking criteria
- Ranking criteria:
  - *Universal* – query (or context) independent
    - *Subsumption*
  - *User-Defined* - query (or context) specific
    - Path Length
    - Context
    - Trust

Subsumption Weight

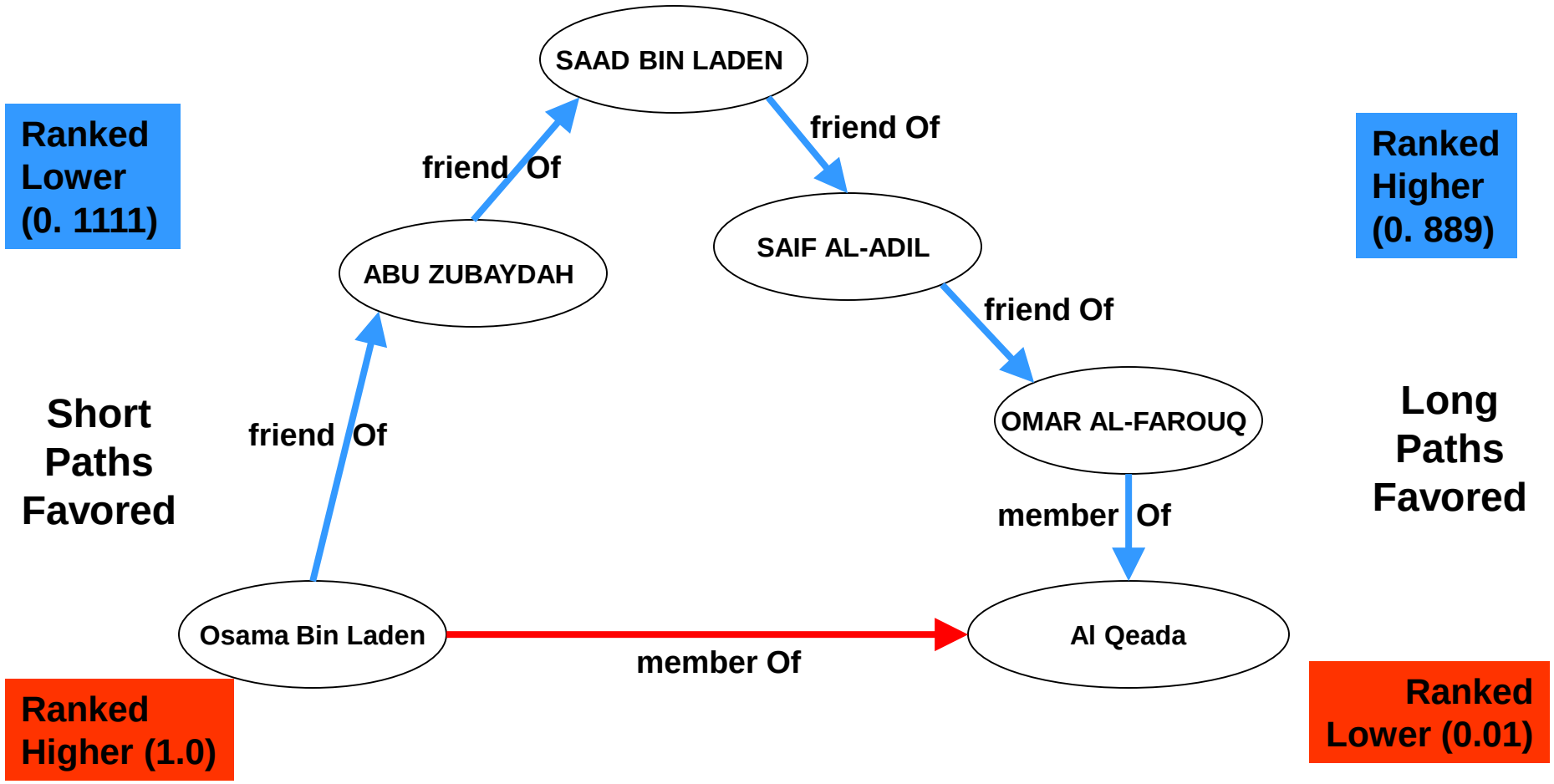
- Specialized instances are considered more relevant
- More “specific” relations *convey more meaning*



Path Length Weight

- Interest in the most direct paths (i.e., the shortest path)
  - May infer a stronger relationship between two entities
- Interest in hidden, indirect, or discrete paths (i.e., longer paths)
  - Terrorist cells are often hidden
  - Money laundering involves deliberate innocuous looking transactions

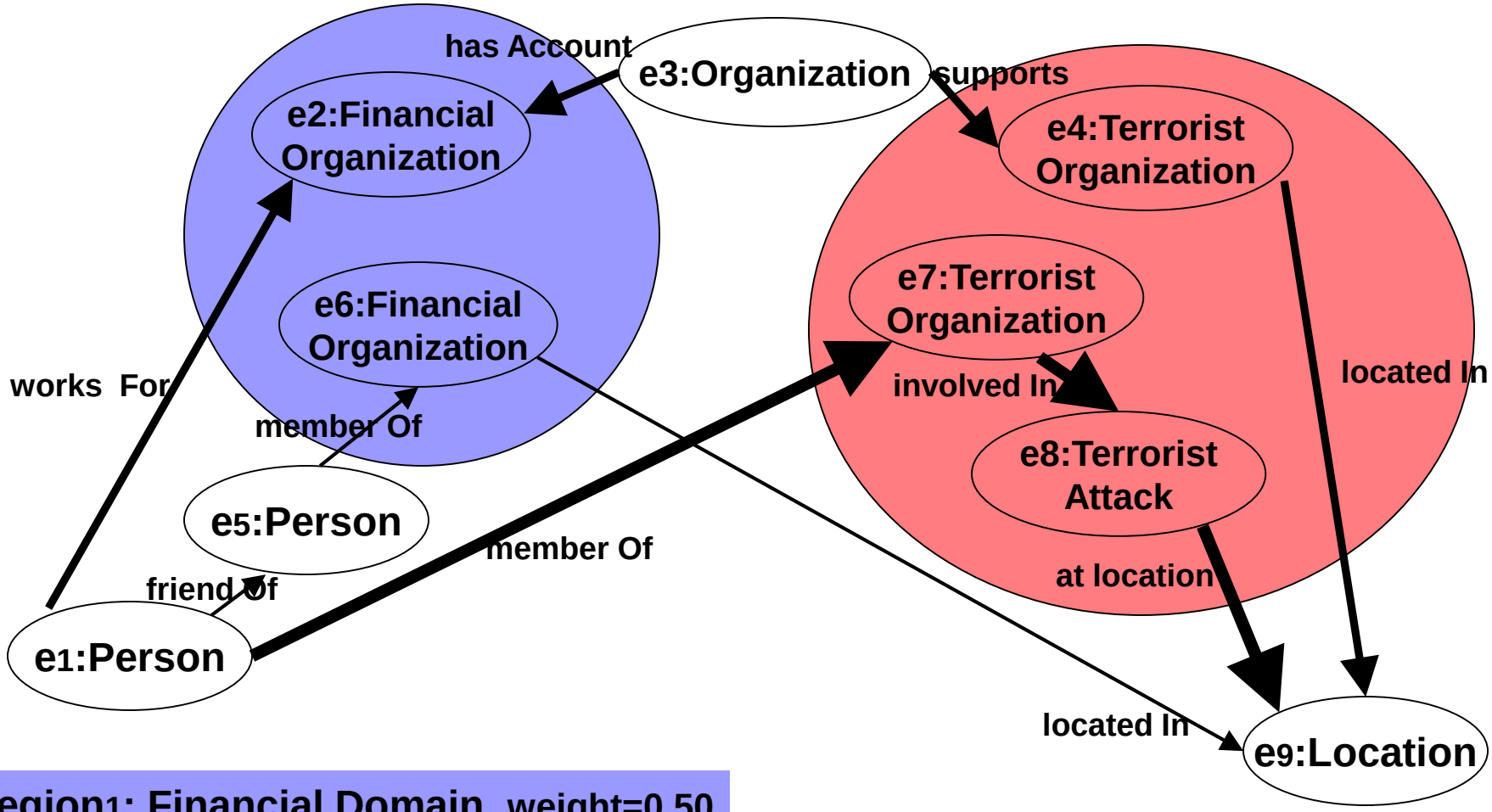
Path Length - Example



Context Weight

- Consider user's domain of interest (user-weighted *regions)*
- Issues
  - Paths can pass through numerous regions of interest
  - Large and/or small portions of paths can pass through these regions
- Paths outside context regions rank lower or are discarded

Context Weight - Example



Region1: Financial Domain, weight=0.50

Region2: Terrorist Domain, weight=0.75

Trust Weight

- Relationships (properties) originate from differently trusted sources
- Trust values need to be assigned to relationships depending on the source
- e.g., Reuters could be more trusted than some of the other news sources
- Current approach penalizes low trusted relationships (may overweight lowest trust in a relationship)

Ranking Criterion

- Overall *Path Weight* of a semantic association is a linear function

$$\begin{aligned} \textit{Ranking} \\ \textit{Score} = \end{aligned} \quad \begin{aligned} & k_1 \times \textit{Subsumption} + \\ & k_2 \times \textit{Length} + \\ & k_3 \times \textit{Context} + \\ & k_4 \times \textit{Trust} \end{aligned}$$

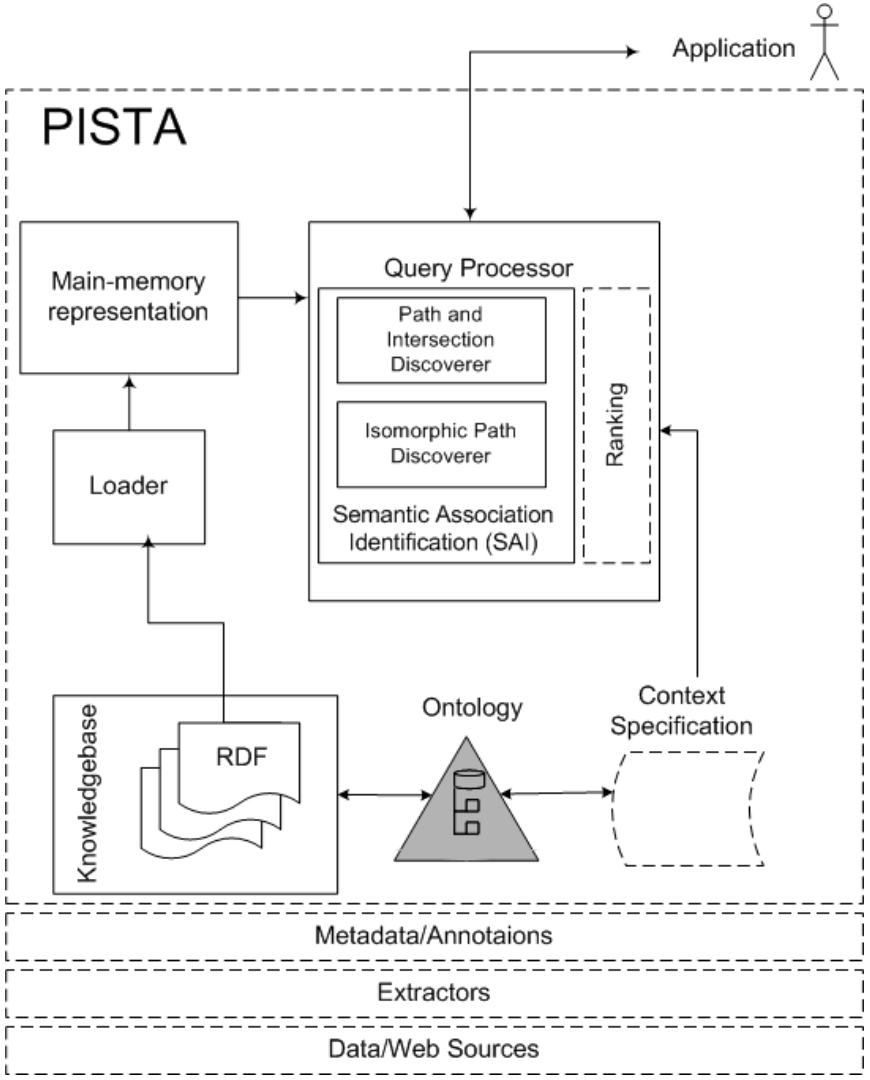
  where k_i add up to 1.0

- Allows fine-tuning of the ranking criteria

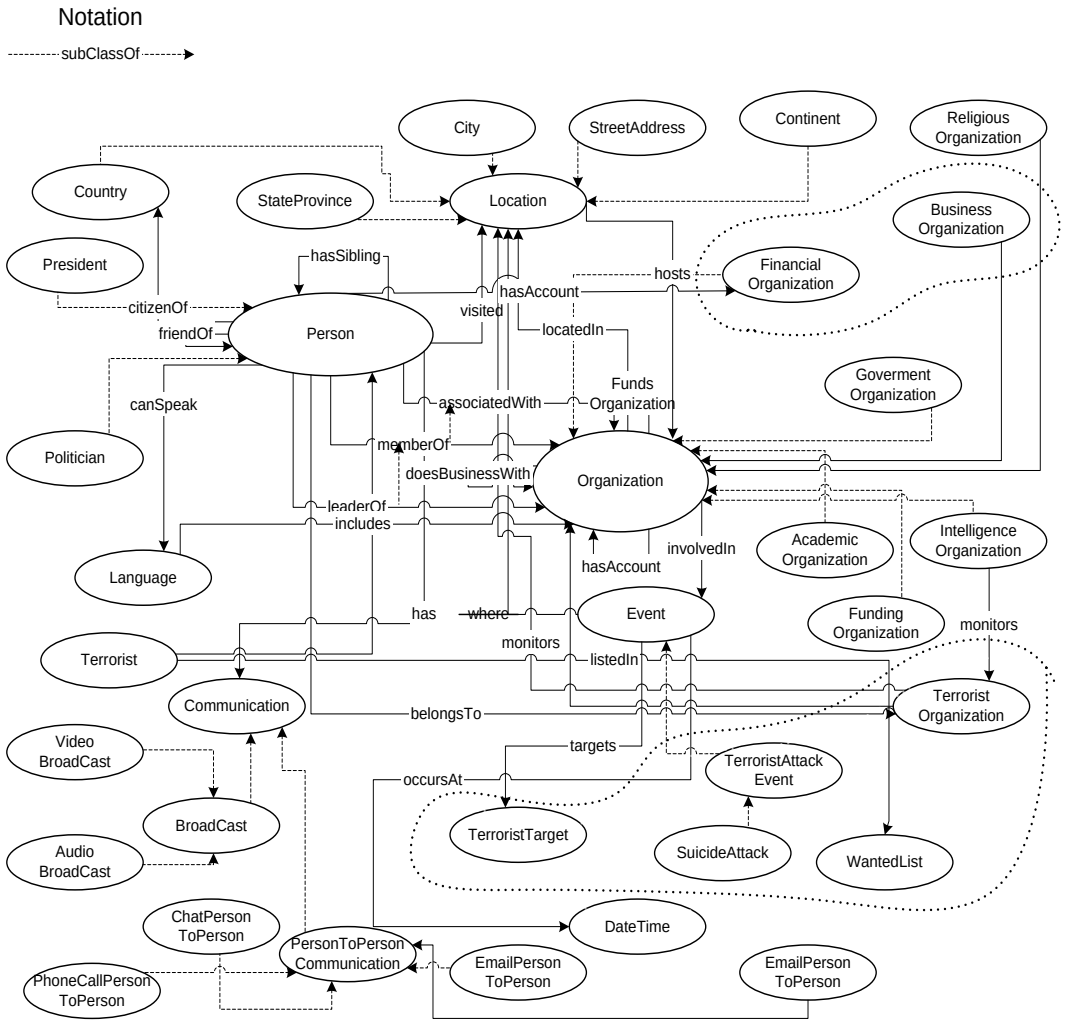
Preliminary Results & Demo

Preliminary Results

- Metadata sources cover terrorism domain
- Ontology in RDFS, metadata in RDF
- Semagix Freedom suite used for metadata extraction
- Currently > 6,000 entities and > 11,000 relations/assertions (plan to increase by 2 order of magnitude)



PISTA Ontology



Preliminary Results

- Have implemented naïve algorithms for ρ *and* σ
  - Using a depth-first graph traversal algorithm
  - Used Jena to interact with RDF graphs (i.e., metadata in main memory)

Demo

- Context
  - ‘A’ defines a region covering ‘*terrorism*’ - weight of 0.6
  - ‘B’ captures ‘*financial*’ region - weight of 0.4
- Ranking criteria (this example)
  - 0.6 to *context*
  - 0.1 to *subsumption*
  - 0.2 to *path length* (longer paths favored),
  - 0.1 to *trust weight*

Demo

- Click here to begin demo

Related Work

Related Work

- Ranking in Semantic Web Portals
  - [Maedche et al 2001]
- Our Earlier Work
  - [Anyanwu et al 2003]
- Contemporary information retrieval ranking approaches
  - [Brin et al 1998], [Teoma]
- Context Modeling
  - [Kashyap et al 1996], [Crowley et al 2002]

Conclusions & Future Work

Summary and Future Work

- This paper: ranking of ρ *path*
  - Even more important than ranking of documents in contemporary Web search
- Ongoing: ranking of σ *path*
- Future:
  - Formal query language for semantic associations is currently under development
  - Develop evaluation metrics for context-aware ranking (different than the traditional precision and recall)
  - Use of the ranking scheme for the *semantic-association* discovery algorithms (scalability in very large data sets)

Questions, Comments, . . .

- For more info:
  - http://lsdis.cs.uga.edu/proj/SAI/
    - PISTA Project, papers, presentations