Finding Experts by Semantic Matching of User Profiles

Rajesh Thiagarajan[1], Geetha Manjunath[2], and Markus Stumptner[1]

[1]Advanced Computing Research Centre
University of South Australia, Adelaide
{cisrkt,mst}@cs.unisa.edu.au

[2]Hewlett-Packard Laboratories
Bangalore, India
geetha.manjunath@hp.com

Index

- Finding experts by matching user profiles
- Spreading
  - Set-based
  - Graph-based
- Similarity Measures
  - Set intersection
  - Bipartite matching
- Evaluation
  - Finding an expert
  - Short-listing a group of experts
- Summary and Future Work

User Profiles

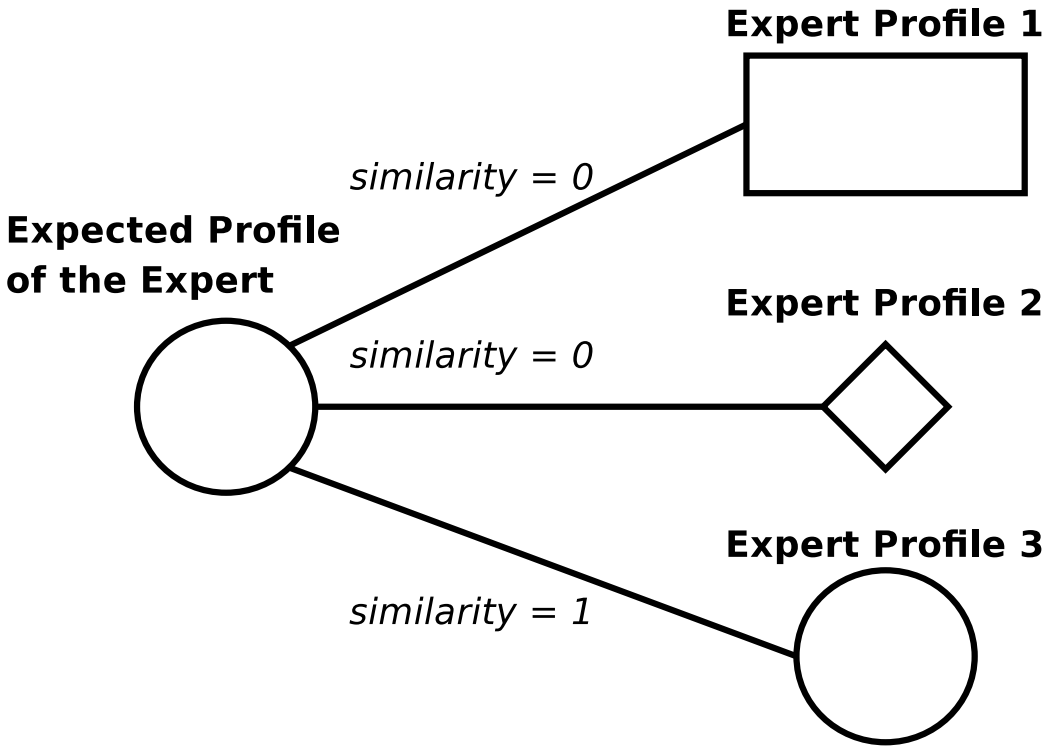
- Profiles are generally used to capture the characteristics of entities.
- **User profiles**: common model of a user in IR frameworks

Term	Weight
Web Services	3.0
RDF	2.0
AI Configuration	2.0
Semantic Web	1.0

- Number of applications: Personalized search, targeted advertisements, etc.,

User Profiles for Expert Finding

- Finding experts by matching user profiles:



- Traditional content matching techniques (Cosine etc.,) are insufficient for expert matching
  - Expected/Available expert profiles are **approximate** descriptions

User Profiles for Expert Finding: Challenges

- Existing approaches to compute similarity treat profiles as self-contained units
- Synonymy
  - Existence of synonyms in the 2 profiles. For example, ”happy” and ”glad”
- Polysemy
  - For example, ”apple” - could mean the fruit or the company
- Granularity
  - Terms may be at different granularity. For example, ”jazz” and ”music”
- Existing similarity computation techniques: Unless the profiles overlap, similarity is 0.

User Profiles for Expert Finding: Challenges

- Consider profiles p_1 and p_2

$$p_1 = \{\langle google, 1.0\rangle\}$$

$$p_2 = \{\langle yahoo, 2.0\rangle\}$$

- p_1 and p_2 don't overlap
- Intersection check: $p_1 \cap p_2 = \varnothing$ (cosine similarity is 0)
- Don't p_1 and p_2 have nothing in common?
  - They seem to be interested in Web-based IT companies or web search tools!

Contributions

- How to capture relationships that are not explicit among the profiles?
  - **Spreading**
  - Extending profiles with related terms using an ontology (Wordnet or Wikipedia)
  - Graph and Set based
- How to compute similarity between profiles post spreading?
  - New similarity measures: **Set intersection** and **Bipartite graph matching**
- How accurate are these new measures in expert matching scenarios?
  - Implementation, pilot user study, and **evaluation**.
  - 2 scenarios: finding an expert, shortlisting a group of experts

Spreading

- Extending profiles with related terms (Ontology).
- Recall *google* and *yahoo* example.

$$p_1 = \{\langle google, 1.0\rangle\}$$

$$p_2 = \{\langle yahoo, 2.0\rangle\}$$

- Intersection check: $p_1 \cap p_2 = \varnothing$ (similarity is 0)
- Spread p_1 and p_2 using the Wikipedia category graph

$$p'_1 = \{\langle google, 1.0\rangle, \langle internet\ search\ engines, 0.5\rangle\}$$

$$p'_2 = \{\langle yahoo, 2.0\rangle, \langle internet\ search\ engines, 1.0\rangle\}$$

- Intersection check: $p'_1 \cap p'_2 \neq \varnothing$ (similarity $\geqslant 0$)

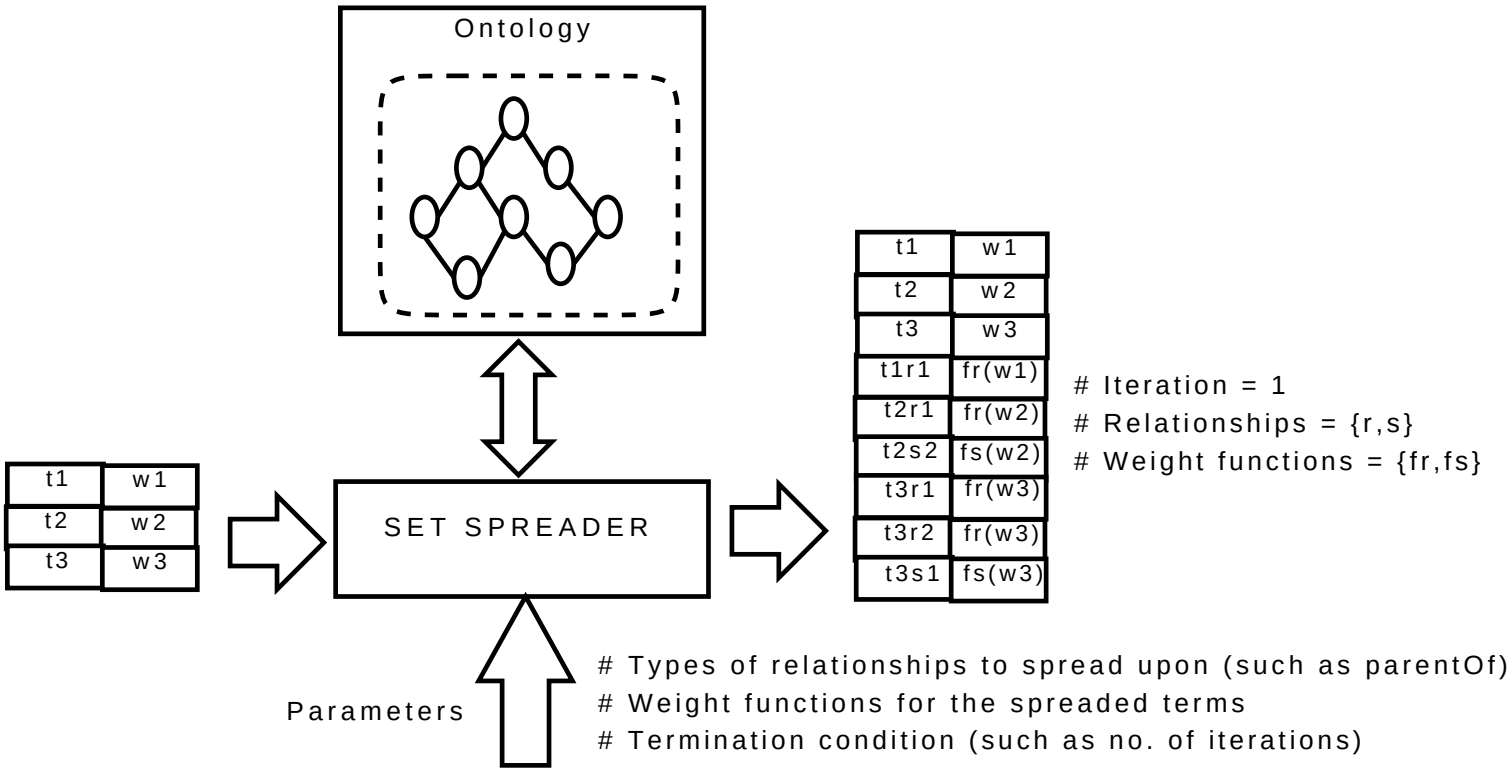
Spreading - Cont.

- Parameters
  - Relationship Types – example, parent categories only.
  - Weight Functions – example, parent category weight is 90% of the original term weight
- Types of Spreading:

Set Spreading	**Graph Spreading**
Termination conditions	
Number of iterations *Exhausted relationships*	*Number of iterations* *Exhausted relationships* *All Path Complete* *At-least Path Complete*

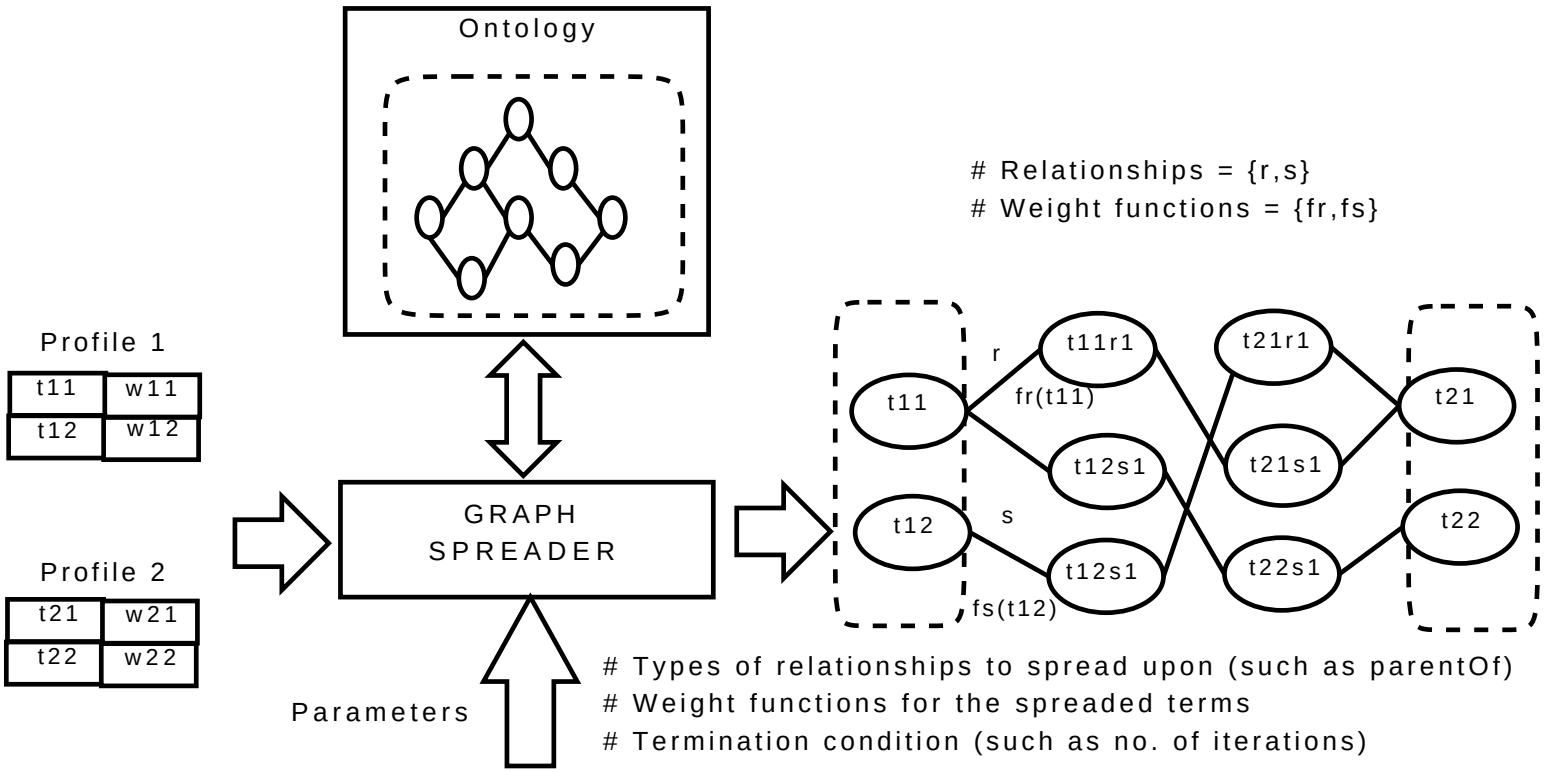
Set Spreading

- Related terms are added to the original set of terms (Ontology, parameters).
- One profile at a time.

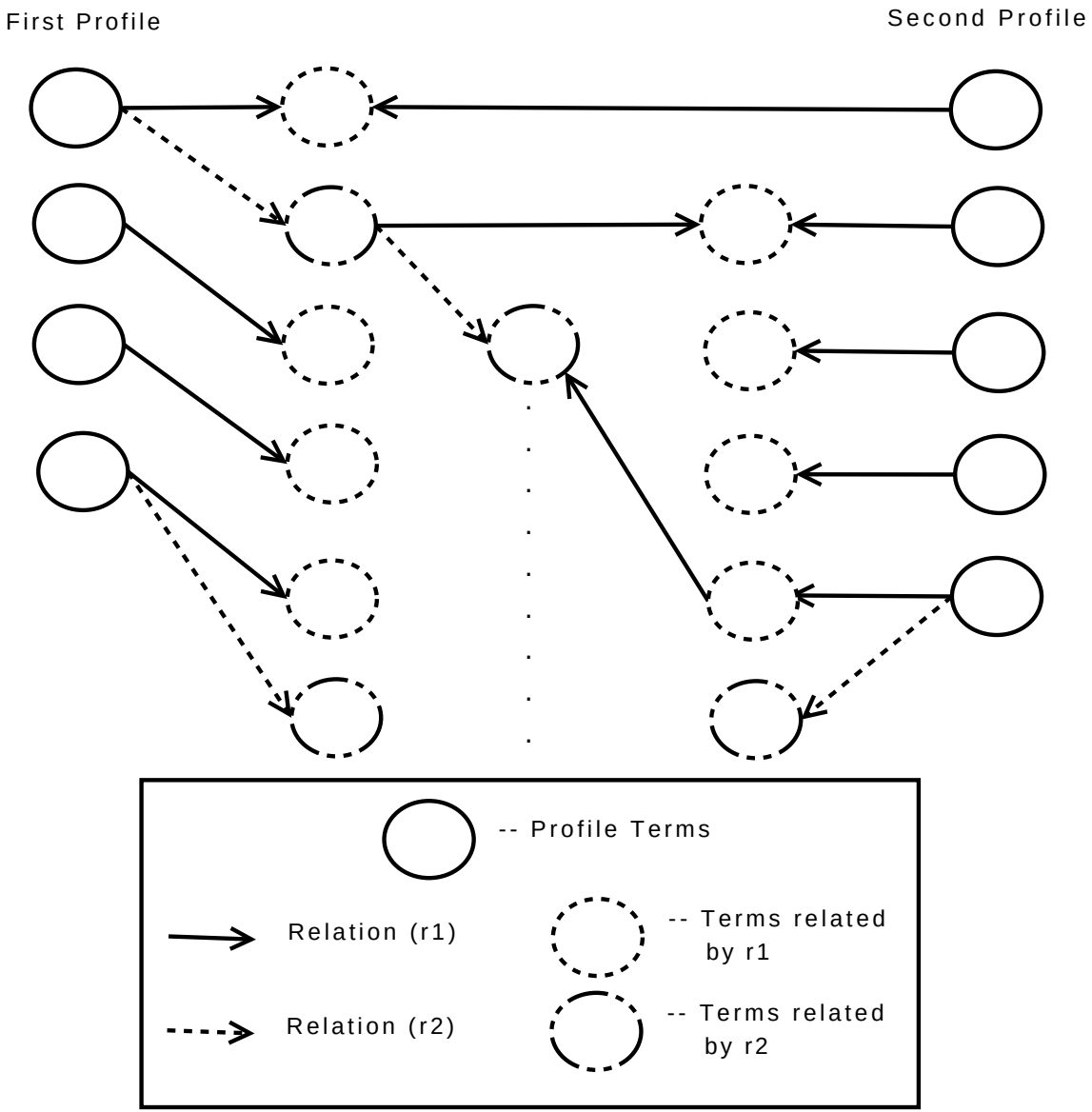


Graph Spreading

- Profile terms and the related terms are build into a graph representation (Ontology, parameters)
- 2 profiles are spread simultaneously



Graph Spreading - Cont.



Basic Concepts

- User profile

$$u = \{\langle t_1, w_1 \rangle, \ldots, \langle t_n, w_n \rangle\}$$

  - t_i are the terms that describes the user
  - w_i denotes the importance of t_i

- Cosine similarity

$$sim_{cos}(u_j, u_k) = cos(\vec{V}(u_j), \vec{V}(u_k)) = \frac{\vec{V}(u_j) \cdot \vec{V}(u_k)}{|\vec{V}(u_j)||\vec{V}(u_k)|}$$

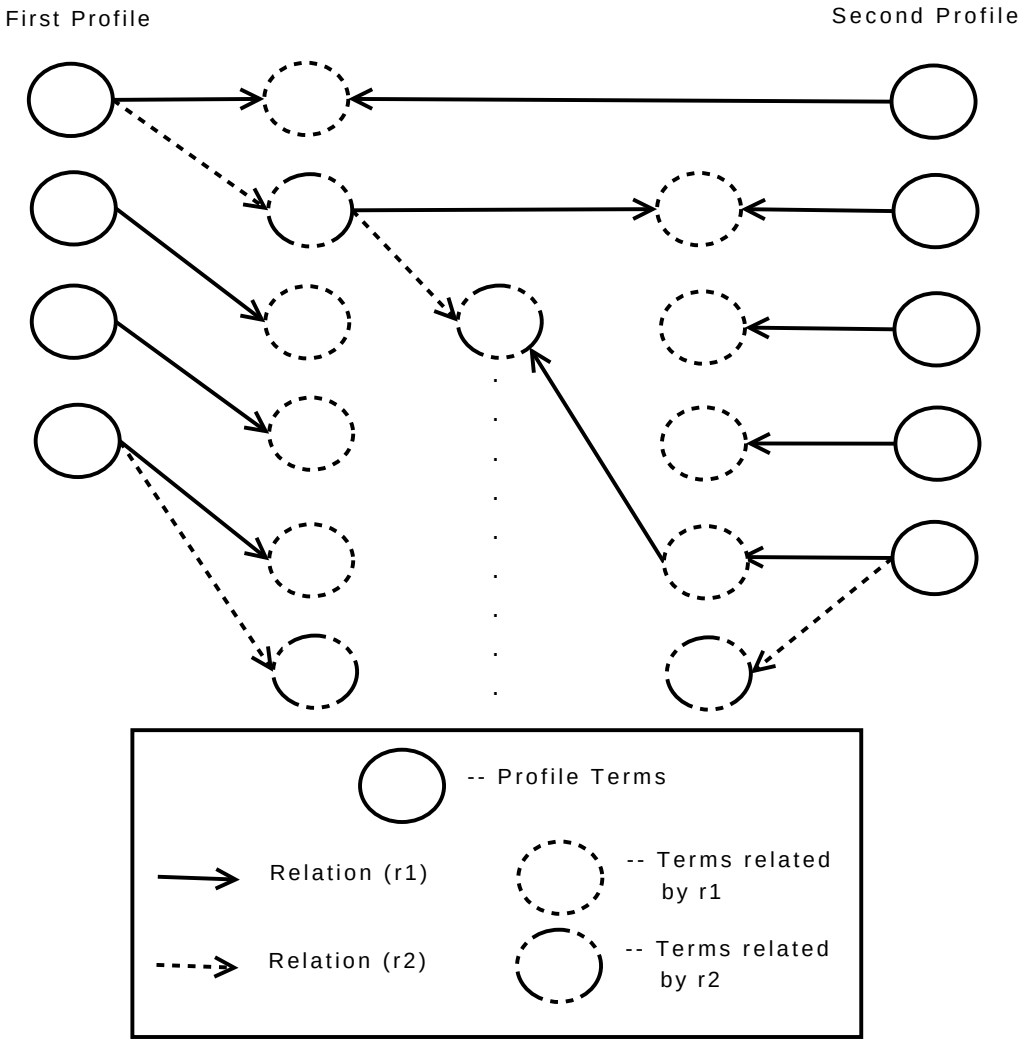
  - Euclidean length ($|\vec{V}(u_j)|$) of an entity u_j is $\sqrt{\sum_{i=1}^{n} w_i^2}$,

Similarity Computation: Set-based

1. Compute cosine similarity between original profiles.
2. Spread (set spreading) the profiles if the termination condition is not met.
3. Compute cosine similarity between the extended profiles.
4. Do step 2.
5. Average out the similarity values.

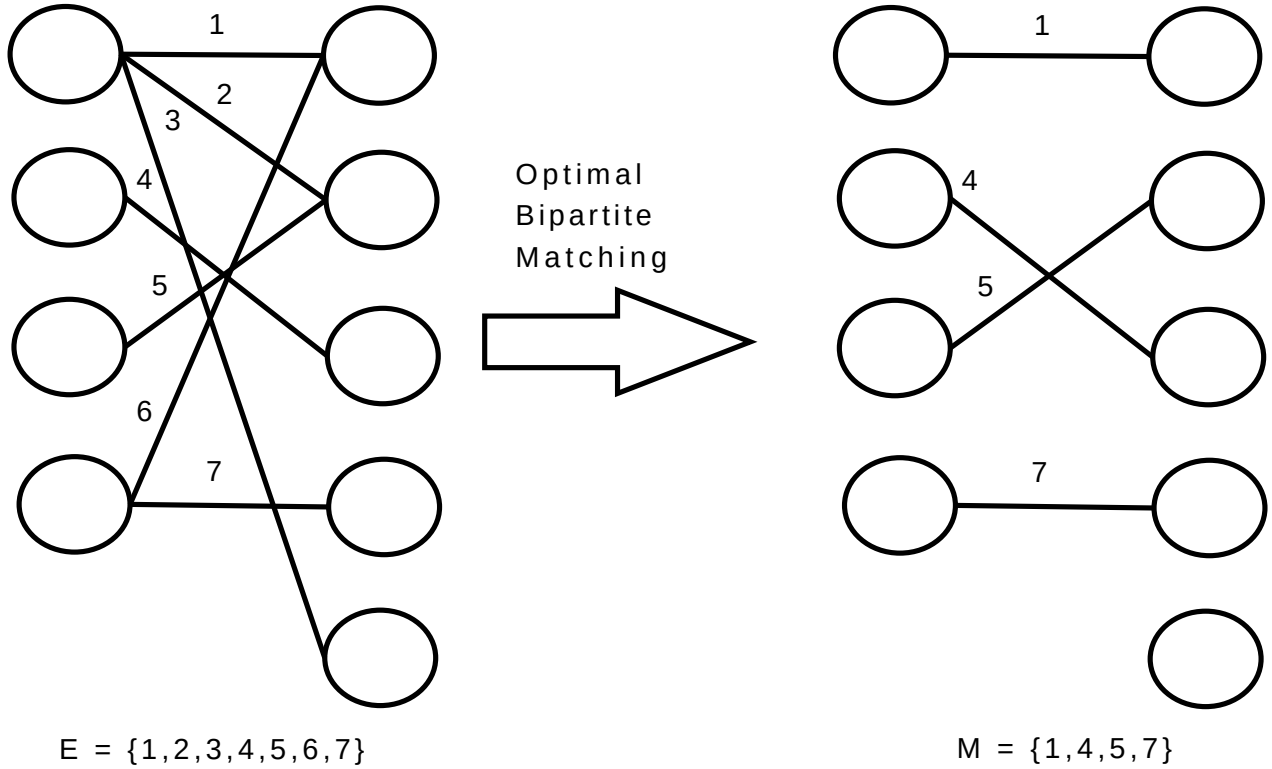
Similarity by Bipartite Matching

Build the Spreading Activation Network (SAN).



Similarity by Bipartite Matching - Reduction

Reduce it to a Bipartite graph. Using the Hungarian Algorithm find the optimal matching.



Expert finding scenario: keeping the trail of mapping between terms is particularly useful

Similarity by Bipartite Matching - Aggregate Path Distances

Rationale: The nodes that are close by and have high weights contribute more to similarity.

$$v_j^k = \langle t_j^k, w_j^k \rangle$$

v_j^k refers to the j^{th} vertex of the profile k (1 or 2)

$$path(e_{ij}) = \begin{cases} 1, & \text{if } len(v_i^1, v_j^2) \text{ is 0} \\ 0, & \text{if } len(v_i^1, v_j^2) \text{ is } \infty \\ \frac{w_i^{1\prime} \times w_j^{2\prime}}{len(v_i^1, v_j^2)}, & \text{otherwise} \end{cases}$$

$$eupath(e_{ij}) = \begin{cases} \text{as above}, & \text{if } len(v_i^1, v_j^2) \text{ is 0 or } \infty \\ \frac{w_i^{1\prime} \times w_j^{2\prime}}{e^{len(v_i^1, v_j^2)}}, & \text{otherwise} \end{cases}$$

$$euhalf(e_{ij}) = \begin{cases} \text{as above}, & \text{if } len(v_i^1, v_j^2) \text{ is 0 or } \infty \\ \frac{w_i^{1\prime} \times w_j^{2\prime}}{e^{\left(\frac{len(v_i^1, v_j^2)}{2}\right)}}, & \text{otherwise} \end{cases}$$

Similarity by Bipartite Matching - Metrics

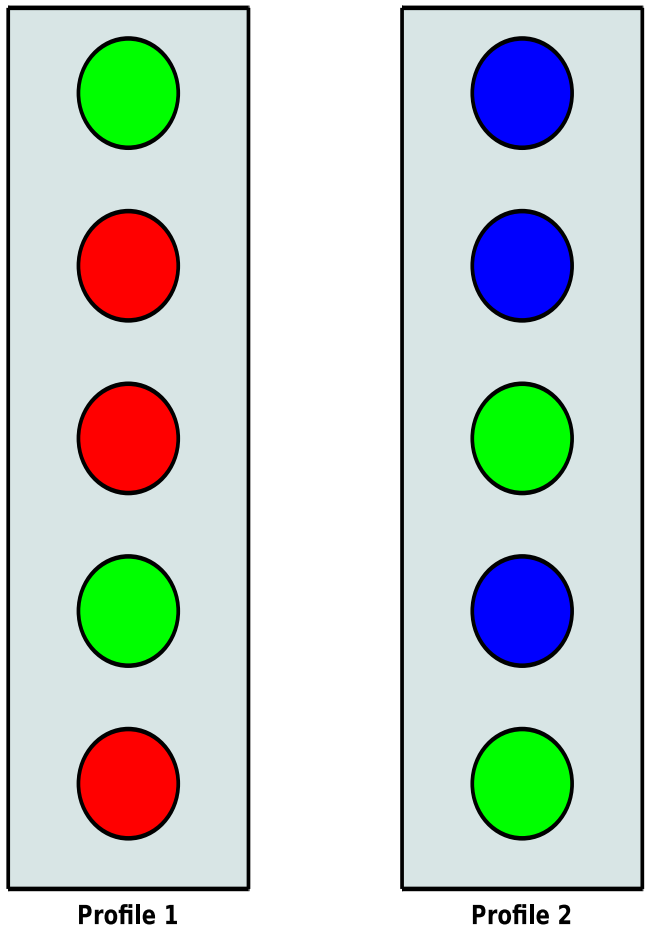
- Normalize the path distances to a value between [0,1]
- Here $max(*path(e_{ij}))$ gives the maximum path distances among all edges

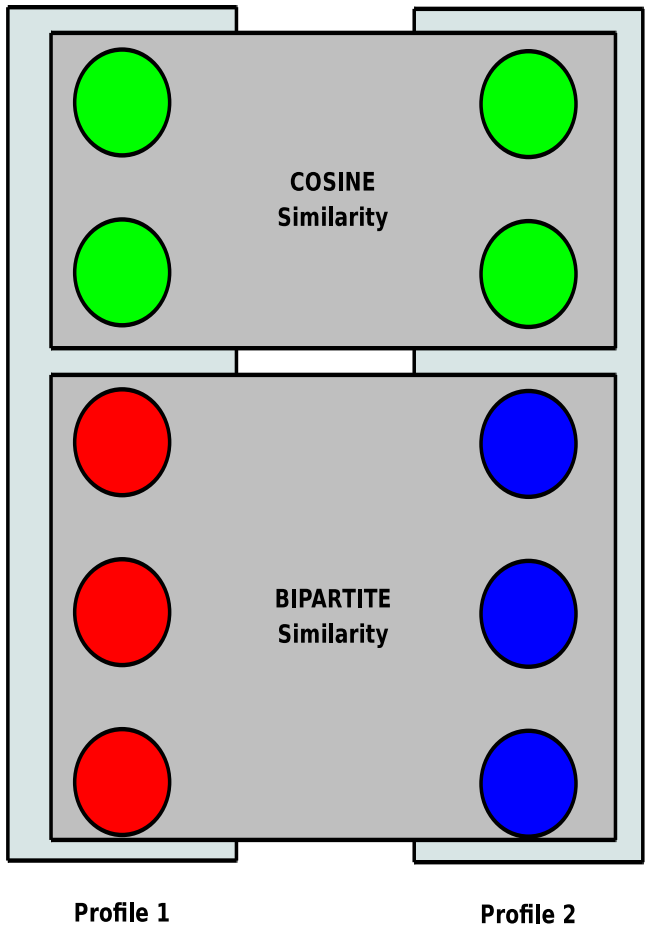
$$sim_{path}(u_1, u_2) = \frac{\sum_{\forall e_{ij} \in E'} path(e_{ij})}{min(size(terms(u_1)), size(terms(u_2))) \times max(path(e_{ij}))}$$

$$sim_{eupath}(u_1, u_2) = \frac{\sum_{\forall e_{ij} \in E'} eupath(e_{ij})}{min(size(terms(u_1)), size(terms(u_2))) \times max(eupath(e_{ij}))}$$

$$sim_{euhalf}(u_1, u_2) = \frac{\sum_{\forall e_{ij} \in E'} euhalf(e_{ij})}{min(size(terms(u_1)), size(terms(u_2))) \times max(euhalf(e_{ij}))}$$

Similarity by Bipartite Matching - Compound Measures





Evaluation - User Study

- A pilot study was conducted for evaluation
- 10 participants with expertise in different fields of CS
- Documents that best describe their research were collected (5 to 10)
- Along with the documents, the participants were asked to give 5 keywords for every document
- Similarity value judgements:
  - Each of the participant judged the similarity between their profile and other profiles
  - Each of the participant judged the similarity between every pair of profiles
- Their judgements were averaged to determine the judgements. These are used as the base/reality values

Evaluation - User Profiles

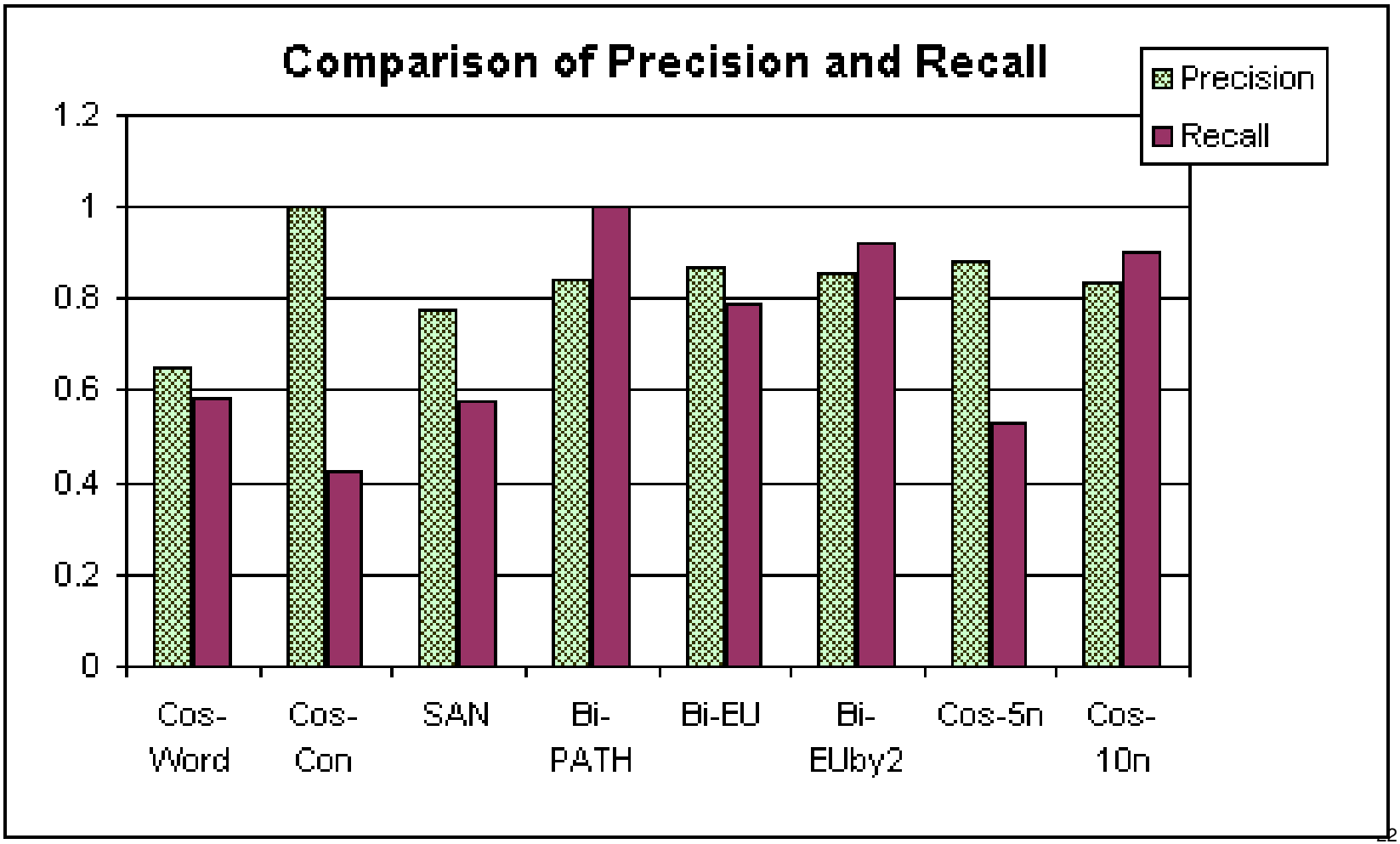
- User profiles were generated using an in-house built software based on [Banerjee et al, SIGIR 2007]
  - BOW: Word profiles (terms are Wordnet concepts)
  - BOC: Wiki profiles (terms are Wikipedia concept; article title)

Measure	Description
COS-Word	Cosine similarity measure between expert and query BOW profiles
COS-Con	Cosine similarity measure between expert and query BOC profiles
COS-5n	Mean cosine similarity between BOC profiles after 5 iterations of set spreading
COS-10n	Mean cosine similarity between BOC profiles after 10 iterations of s et spreading
Bi-PATH	Compound similarity measure after graph spreading using *path*
Bi-EU	Compound similarity measure after graph spreading using *eupath*
Bi-EUby2	Compound similarity measure after graph spreading using *euhalf*
SAN	Similarity measure after graph spreading using semantic search techniques for activating SAN

- Similarity computed between the generated profiles

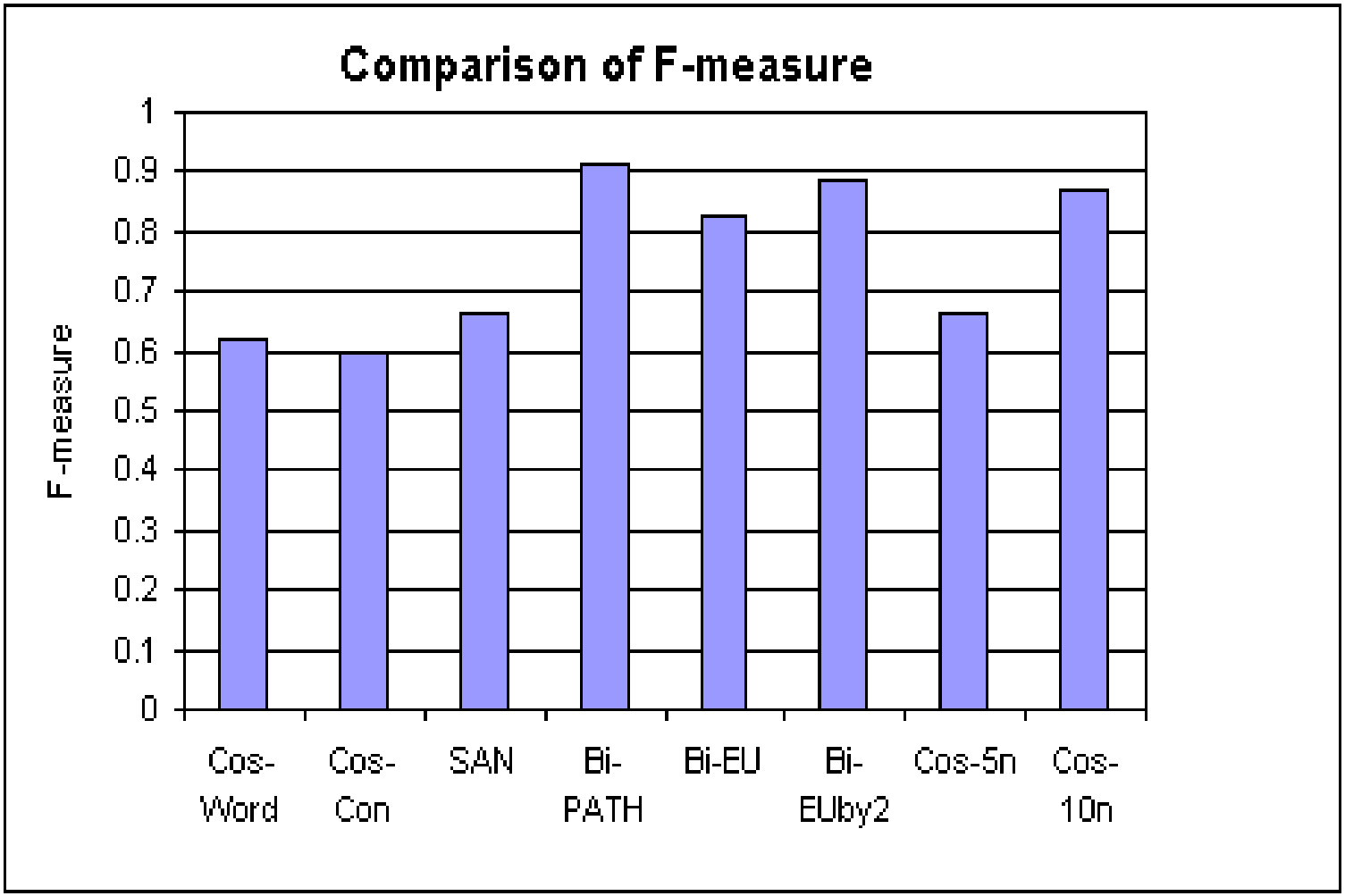
Evaluation - Short-listing a group of Experts

Bi-PATH and Bi-EUby2 perform the best followed by COS-10n



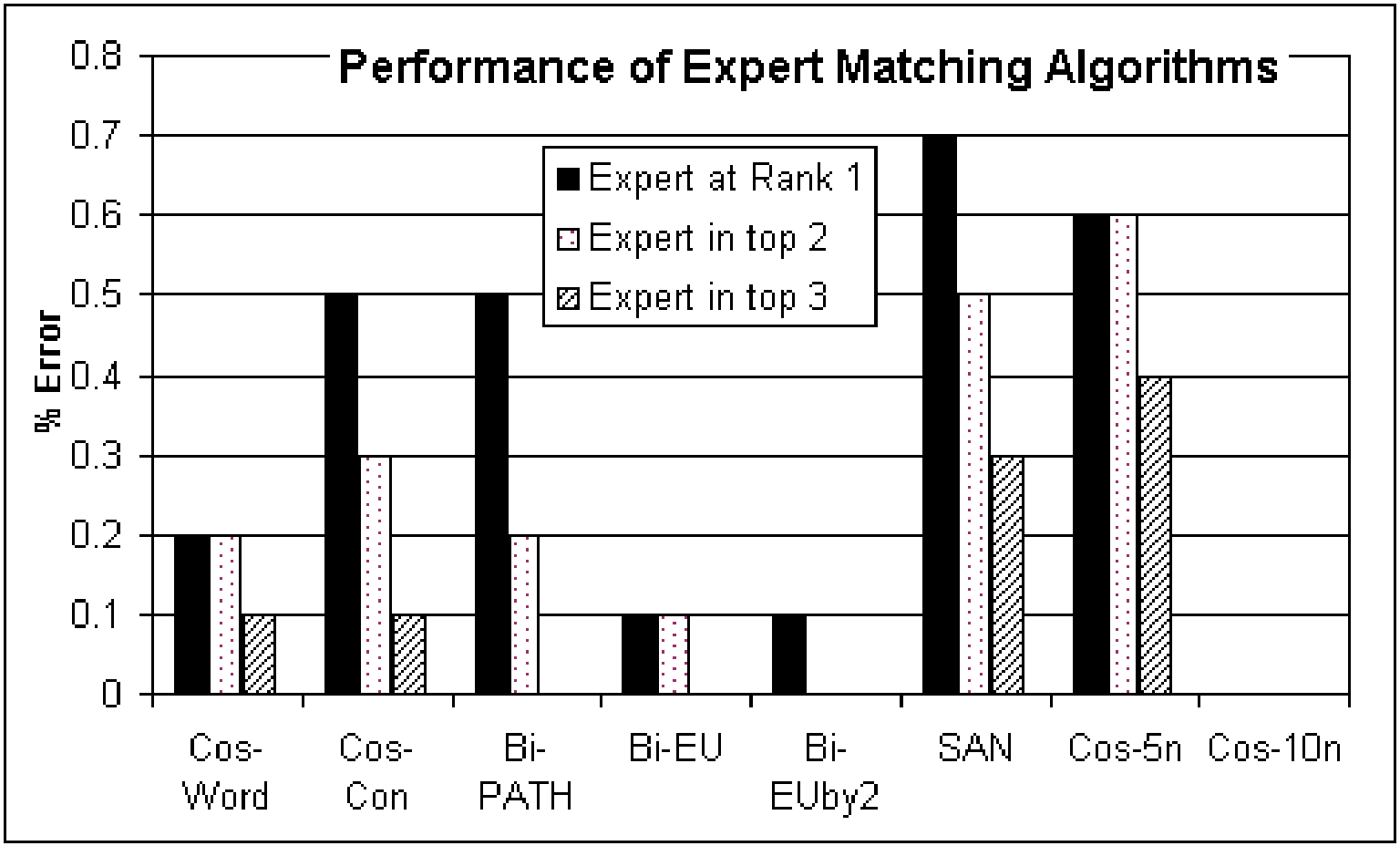
Evaluation - Short-listing a group of Experts

Bi-PATH and Bi-EUby2 perform the best followed by COS-10n



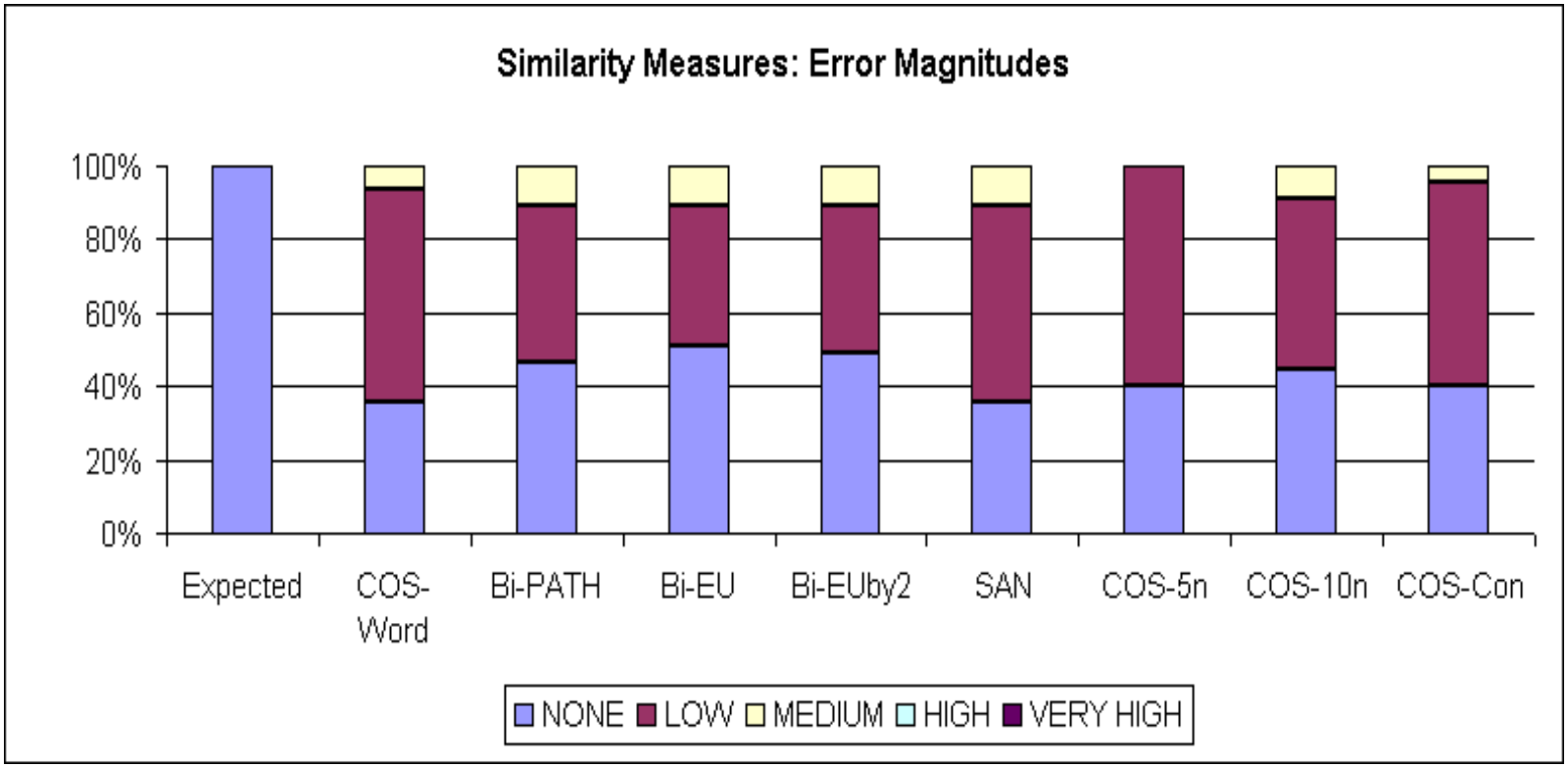
Evaluation - Finding an Expert (Top 3 matches)

COS-10n performs the best followed by Bi-*



Evaluation - Similarity measure for matching any 2 users

Bi-* have the highest percentage of *no errors*



Summary and Future Work

- A number of new similarity measures that improve accuracy of expert finding processes
- Evaluations of the approaches indicate the improvements in accuracy
- Future Work
  - To incorporate other concept-concept similarity measures instead of path measure and study its effects
  - Investigate techniques to compute similarity relative to a reference concept (black cat, red cat)
  - Extend to automatically use other domain ontologies from a repository like Swoogle

Thank you!