SimRank: A Measure of Structural-Context Similarity

Glen Jeh and Jennifer Widom

KDD 2002

CS 519 Class Presentation
Presenter: Anh Pham

Outline of the talk

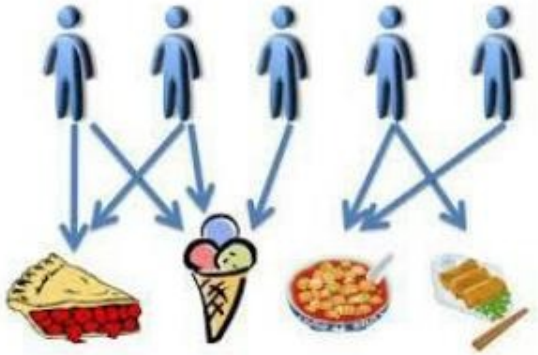
- Introduction to Structural Context Similarity
- SimRank
- Computing SimRank
  - Naïve method
  - Pruning
- Example
- Limited information problem
- Random surfer pair model
- Experimental results
- Strong and weak points
- Quiz

Finding similarity objects problem

- There are a lot of applications

1. Find similar documents:

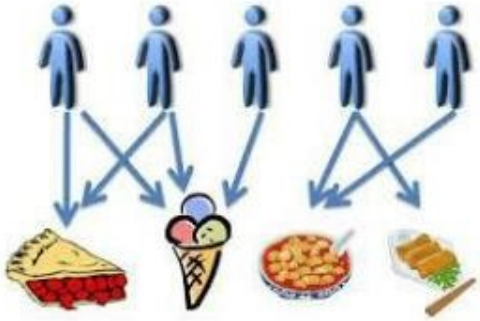
2. Collaborative filtering:
   - Find similar users
   - Find similar items



Aspects of objects for similarity

- Many aspects making similarity
  - Documents: common words, sentence…

  - Users: common preferences

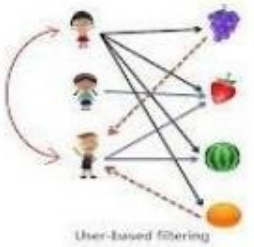


Structure similarity

- This paper proposes a general approach which can be applied when the data can be represented as graph

1. Web page cases:

2. Users preferences:

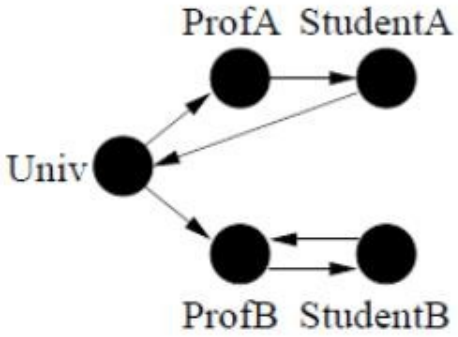


3. Scientific network:

Example of structure similarity

- Intuition: similar objects are related to similar objects

- Example:



Prof. A has student A & Prof. B has student B

1. Prof. A and Prof. B are similar, since they from the same univ.
2. Recursively, student A and student B are similar.
3. →If we know the similarity of Prof. A and B, we may estimate the similarity btw student A and B

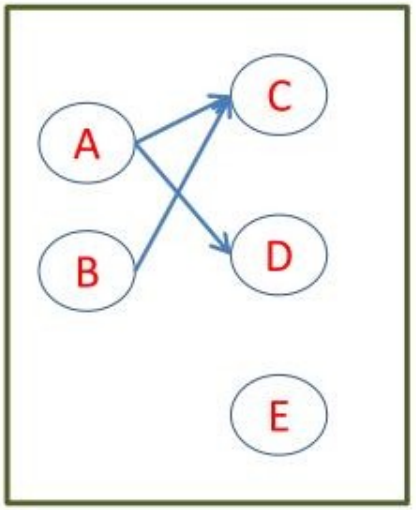
Some basic notations in graph models

- Graph G=(V,E) where V represent the nodes, and E represent the edges.

- If nodes p and q, then <p,q> denotes the edge from p to q.

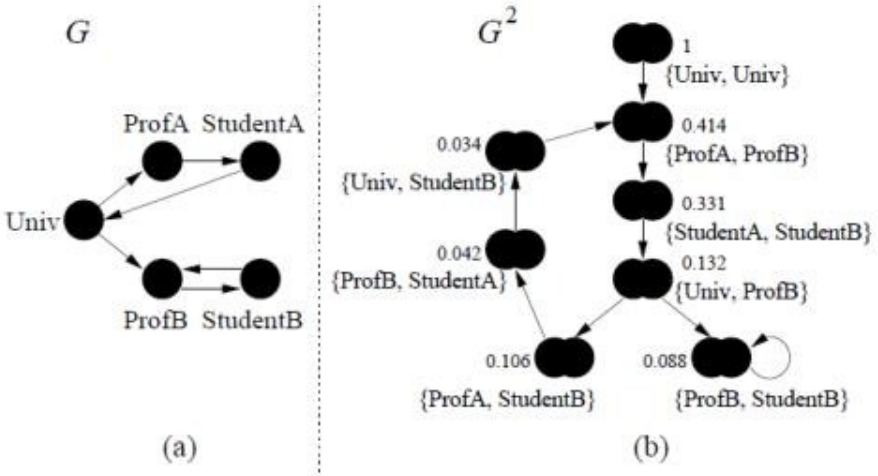
- I(v) denotes the in-neighbors of v
- O(v) denotes the out-neighbors of v

→I(C)={A,B} and O(A)={C,D}



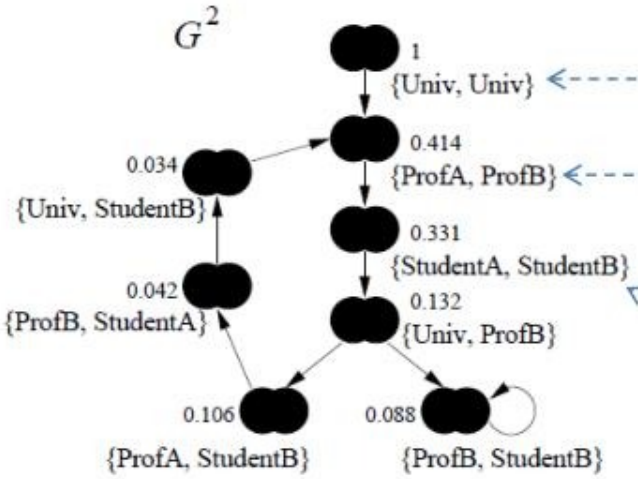
Node pair graph

- Creating a node pair graph G^2 from G
- <(p,q),(a,b)> is in G^2 if <p,a> and <q,b> are in G
- Example:



Simrank motivation

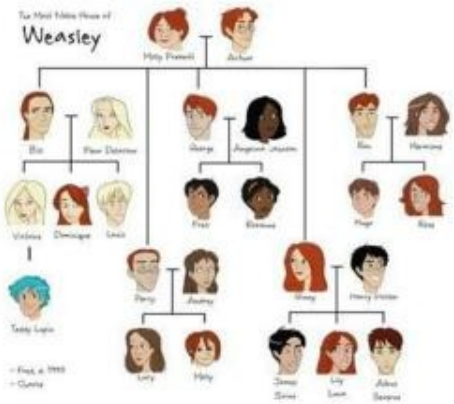
- Intuition: similar objects are related to similar objects

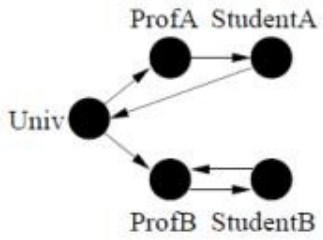


Univ=Univ → Sim(Univ, Univ)=1

Prof. A related to Univ
Prof. B related to Univ
→ Sim(Univ, Univ)=.414 <1

Student A related to Prof. A
Student B related to Prof. B
→ Sim(S_A, S_B)=.331 <1





Simrank equation

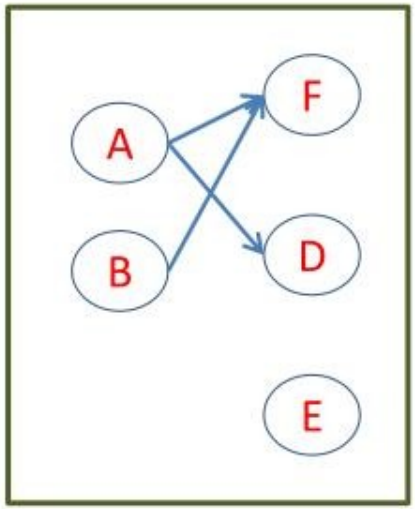
- Similarity btw a and b:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

- Example:
  - *Assume C=1*

$$S(F,D) = \frac{1}{|2|*|1|} * [S(A,A)+S(B,A)]$$

=1/2*(1+0.5)=0.75



Simrank equation (1)

- Similarity btw a and b:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

- s(a,b) is symmetric
- s(a,a)=1
- s(a,x)=0 if x has no neighbor

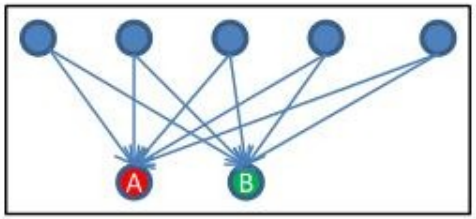
Simrank equation (2)

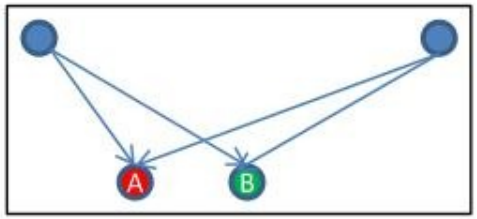
- Similarity btw a and b:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

- s(a,b) is normalized into (0,1)

- Proof: By induction
  - C<1
  - $s(I_i(a),I_j(b))<1$





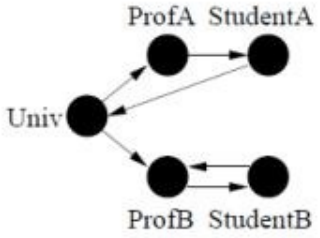
Simrank equation (2)

- Similarity btw a and b:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

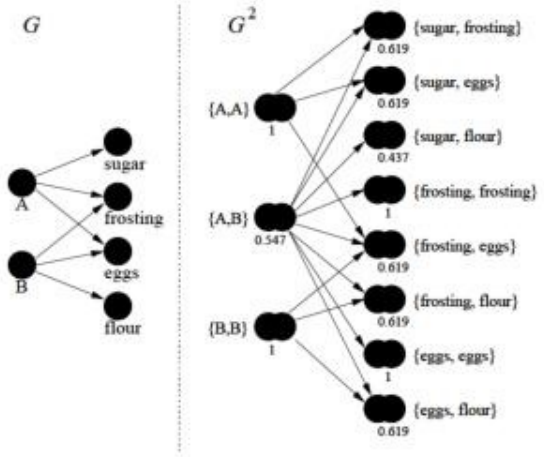
- Factor C should be <1

- C represent the confidence level, propagated from the parent nodes



Bipartite Simrank

- Consider a recommendation system:



- How we can recommend a item to a new buyer?
- A and B are similar since they both buy frosting and eggs → recommend flour for A

Bipartite Simrank (mutually-reinforcing rule)

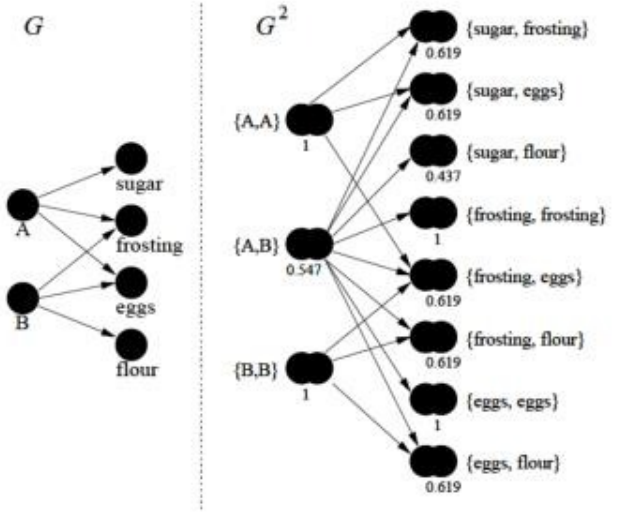
1. **Rule 1:** People are similar if they purchase similar items
2. **Rule 2:** Items are similar if they are purchased by similar people

→ Rule 1 reinforces Rule 2, and vice versa

Example:

1. If frosting and eggs are similar, then A and B also similar.
2. If A and B are similar then frosting and eggs are similar.

Observation: We can magically see the similar of **sugar** and **flour**, even though there is no common customer.



Bipartite Simrank (formula)

1. **Rule 1:** People are similar if they purchase similar items

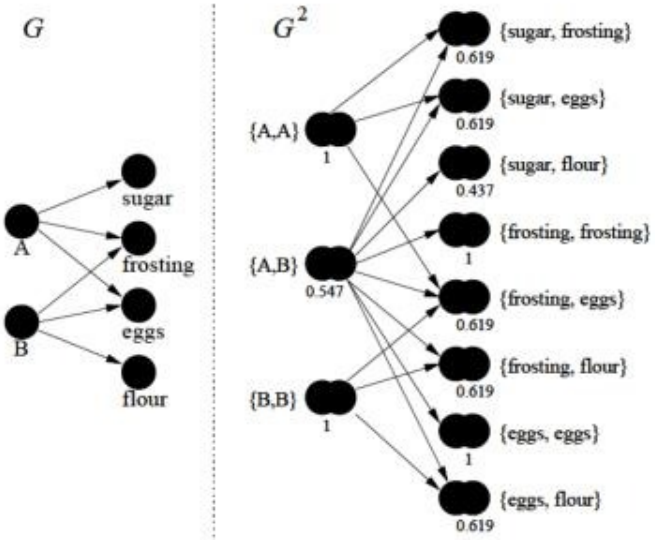
2. **Rule 2:** Items are similar if they are purchased by similar people

$$s(A,B) = \frac{C_1}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B))$$

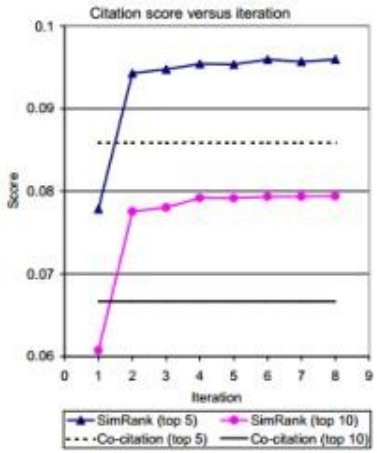
Rule 1 (in math form)

$$s(c,d) = \frac{C_2}{|I(c)||I(d)|} \sum_{i=1}^{|I(c)|} \sum_{j=1}^{|I(d)|} s(I_i(c), I_j(d))$$

Rule 2 (in math form)



Experimental results



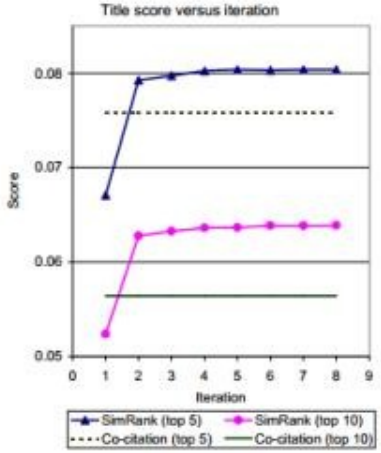


Figure 5: SimRank and co-citation on scientific papers.

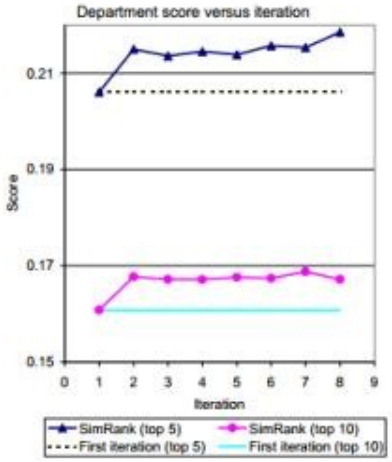
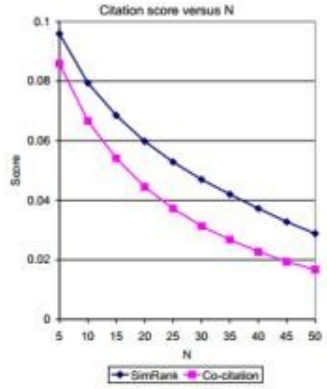


Figure 7: SimRank on courses for increasing iterations.



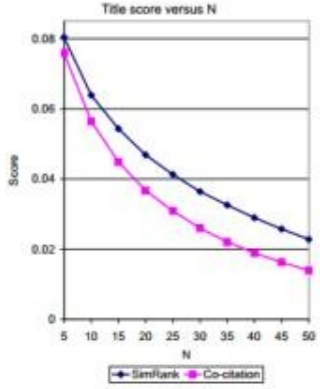


Figure 6: SimRank and co-citation on scientific papers for varying N.

Good points

- The paper proposes a novel method to compute the similarity of objects, in general, based on the structure of data
- The paper proposes a method to compute and efficient pruning technique
- The paper provides an intuition for the method
- There are good experiments results prove their idea

Weak points

- Scalability: The paper should mention about very huge size graph.

- It may incorporate distributed design. Since the algorithm is fixed point process, it should be a research problem on how to parallelize it.

Quiz

- Intuitively, in which graph, the SimRank of a and b are higher ?