Link-Based Similarity Measure for Academic Literature Data

December 11, 2013

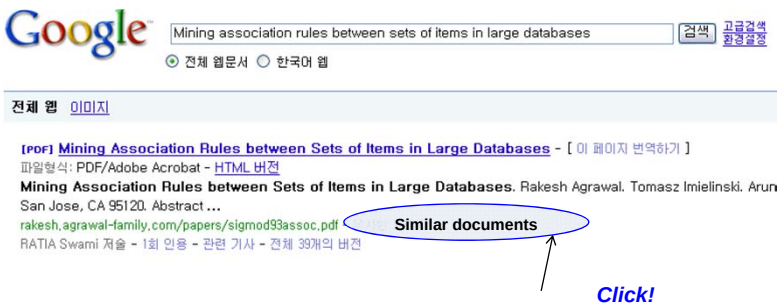
Sang-Wook Kim
Department of Computer Science and Engineering
Hanyang University

Contents

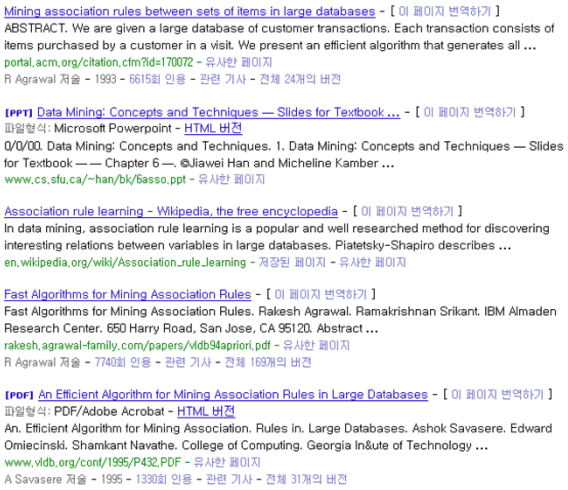
Background

- Research on Academic Literature Data
  - A lot of research papers are being published online
  - Paper search engines have been developed
    - **CiteSeer, Google Scholar, Microsoft Academic Search, and DBLP**
  - Main issues touched
    - Ranking papers according to their authority
    - ***Finding similar papers***

Motivating Example



Motivating Example



Mining association rules between sets of items in large databases - [이 페이지 번역하기]
ABSTRACT. We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all ...
portal.acm.org/citation.cfm?id=170072 - 유사한 페이지
R Agrawal 저술 - 1993 - 6615회 인용 - 관련 기사 - 전체 24개의 버전

[PPT] Data Mining: Concepts and Techniques — Slides for Textbook ... - [이 페이지 번역하기]
파일형식: Microsoft Powerpoint - HTML 버전
0/0/00. Data Mining: Concepts and Techniques. 1. Data Mining: Concepts and Techniques — Slides for Textbook — — Chapter 6 —. ©Jiawei Han and Micheline Kamber ...
www.cs.sfu.ca/~han/bk/6asso.ppt - 유사한 페이지

Association rule learning - Wikipedia, the free encyclopedia - [이 페이지 번역하기]
In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes ...
en.wikipedia.org/wiki/Association_rule_learning - 저장된 페이지 - 유사한 페이지

Fast Algorithms for Mining Association Rules - [이 페이지 번역하기]
Fast Algorithms for Mining Association Rules. Rakesh Agrawal. Ramakrishnan Srikant. IBM Almaden Research Center. 650 Harry Road, San Jose, CA 95120. Abstract ...
rakesh.agrawal-family.com/papers/vldb94apriori.pdf - 유사한 페이지
R Agrawal 저술 - 7740회 인용 - 관련 기사 - 전체 169개의 버전

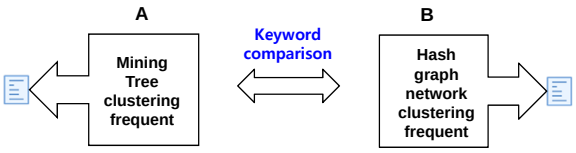
[PDF] An Efficient Algorithm for Mining Association Rules in Large Databases - [이 페이지 번역하기]
파일형식: PDF/Adobe Acrobat - HTML 버전
An. Efficient Algorithm for Mining Association. Rules in. Large Databases. Ashok Savasere. Edward Omiecinski. Shamkant Navathe. College of Computing. Georgia In&ute of Technology ...
www.vldb.org/conf/1995/P432.PDF - 유사한 페이지
A Savasere 저술 - 1995 - 1330회 인용 - 관련 기사 - 전체 31개의 버전

Computing Document Similarity

- Text-based methods
  - To compute the similarity of two documents based on keywords in each document
  - Examples: Cosine similarity, chi-sim, SVD, and LDA
- Link-based Methods
  - To compute the similarity of two documents based on in-links (or out-links) to (or from) each document
  - Examples: Bibliographic coupling, co-citation, Amsler, rvs-SimRank, SimRank, and P-Rank

Cosine Similarity: Text-Based Method

- To represent a document as a vector and to compute the similarity of two documents via the cosine measure between the two vectors
  - A dimension corresponds to a keyword
  - A value of a dimension corresponds to the frequency (or term frequency / document frequency) of the keyword
  - Equation:
    - $Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|}$

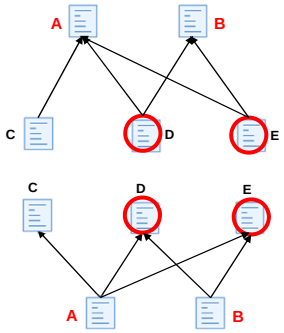


Link-Based Similarity Methods

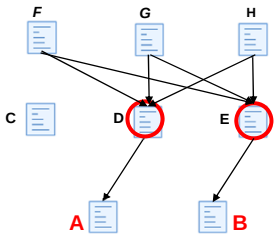
- Non-Recursive Methods
  - Co-citation
  - Bibliographic coupling (Coupling)
  - Amsler
- Recursive Methods
  - SimRank
  - rvs-SimRank
  - P-Rank

Non-Recursive Methods

- Co-citation
  - Number of common objects *pointing to* the two
  - $Sim(a,b) = |I(a) \bigcap I(b)|$
- Bibliographic coupling
  - Number of common objects *pointed by* the two
  - $Sim(a,b) = |O(a) \bigcap O(b)|$
- Amsler
  - Weighted sum of co-citation and bibliographic coupling
  - $Sim(a,b) = \lambda \times |I(a) \bigcap I(b)| + (\lambda - 1) \times |O(a) \bigcap O(b)|$

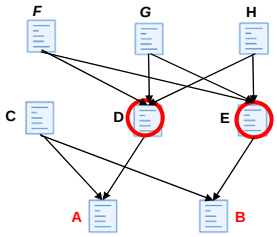


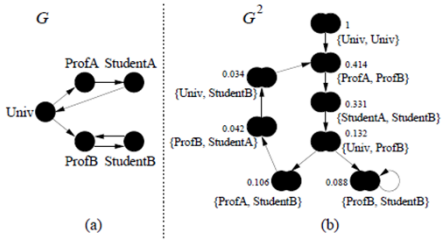
Problem with Non-Recursive Methods



Recursive Methods

- SimRank
  - Recursive version of co-citation
  - Average of similarities among *all possible pairs of objects* pointing to the two
  - $$R_0(a,b)=\begin{cases}0 \text{ (if } a\neq b)\\ 1 \text{ (if } a=b)\end{cases} \qquad R_{k+1}(a,b)=\frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}R_k(I_i(a),I_j(b))$$





Recursive Methods

- rvs-SimRank
  - Recursive version of bibliographic coupling
  - Average of similarities among *all possible pairs of objects* pointed by them
  - $$R_0(a,b)=\begin{cases}0 \text{ (if } a\neq b)\\ 1 \text{ (if } a=b)\end{cases} \qquad R_{k+1}(a,b)=\frac{C}{|O(a)||O(b)|}\sum_{i=1}^{|O(a)|}\sum_{j=1}^{|O(b)|}R_k(O_i(a),O_j(b))$$
- P-Rank
  - Recursive version of Amsler
  - Weighted sum of SimRank and rvs-SimRank

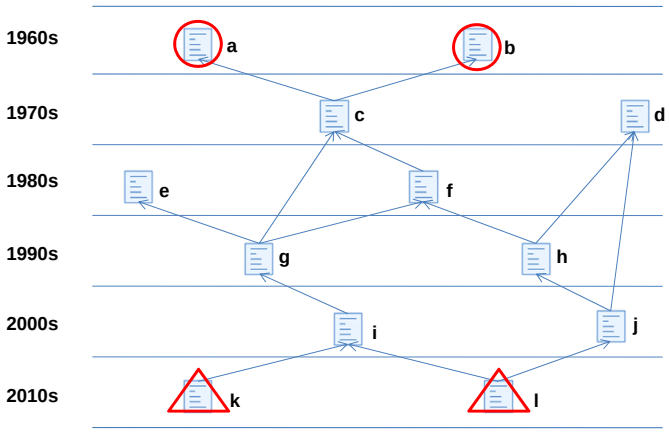
$$R_0(a,b)=\begin{cases}0 \text{ (if } a\neq b)\\ 1 \text{ (if } a=b)\end{cases}$$

$$R_{k+1}(a,b)=\lambda\times\frac{C}{|O(a)||O(b)|}\sum_{i=1}^{|O(a)|}\sum_{j=1}^{|O(b)|}R_k(O_i(a),O_j(b)) + (\lambda-1)\times\frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}R_k(I_i(a),I_j(b))$$

Characteristics of Academic Literature Data

- A paper can cite only those papers *published earlier* than it
  - Few out-links in old papers
    - An old paper does not have the papers, in the database, that it cites
  - Few in-links in young (recent) papers
    - A young (recent) paper does not have the papers, in the database, that cite it

Problems of Previous Methods

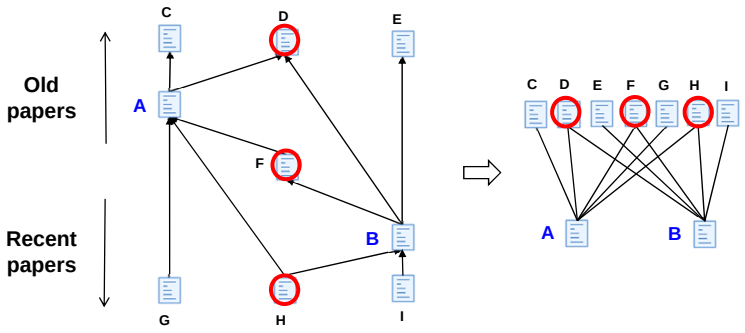


Motivation

- When are two papers A and B considered similar?
  - Case 1: A number of papers commonly cite both of A and B
    - Occurs when computing similarity between old papers – Co-citation (SimRank)
  - Case 2: A number of papers are commonly cited by both of A and B
    - Occurs when computing similarity between young papers – Coupling (rvs-SimRank)
  - Case 3: A number of papers cite A and are also cited by B, or vice versa
    - Occurs when computing similarity between one old and the other young – ***No previous methods***
- **We need a method that considers *all these three cases***

Basic Idea

- To build an undirected graph by ***ignoring the directions of citations***
- To compute the similarity of two papers A and B by considering the number of papers that are ***connected*** to both A and B



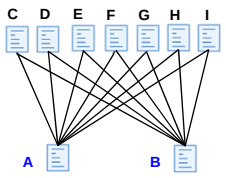
Proposed Method

- Non-recursive method
  - $S(a,b) = |L(a) \cap L(b)|$
- Recursive method
  - Pair-wise normalization
    - $R_0(a,b) = \begin{cases} 0 \,(\text{if } a \neq b) \\ 1 \,(\text{if } a = b) \end{cases}$

$$R_{k+1}(a,b) = \frac{C}{|L(a)||L(b)|} \sum_{i=1}^{|L(a)|} \sum_{j=1}^{|L(b)|} R_k(L_i(a), L_j(b))$$

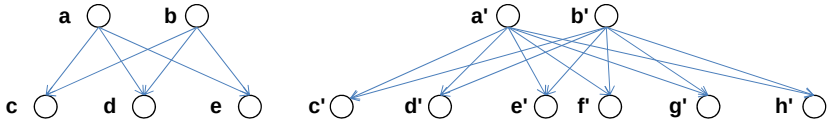
Problem with Pair-wise Normalization

- The similarity of two objects becomes smaller as the number of neighboring objects increases
  - If two objects both have common m-neighboring objects
  - Their similarity becomes 1/m (when direct neighbors are only considered)



Problem with Pair-wise Normalization

- The similarity of A and B decreases as they have more common neighbors



	rvs-SimRank
s(a,b)	**0.333 (3/9)**
s(a',b')	**0.166 (6/36)**

Evaluation

- Data
  - Papers in a database area: DBLP_DB
    - Number of papers: 55,569
    - Number of citations: 142,604

Journals and conferences related to a database area

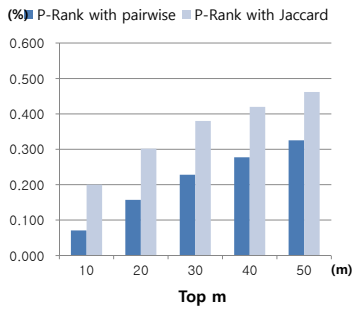
ADBIS, ADC, ARTDB, BNCOD, CDB, CIKM, CoopIS, DANTE, DASFAA, DAWAK, DB, DBPL, DBSEC, DEXA, DKD, DKE, DL, DMKD, DNIS, DOLAP, DOOD, DPD, DPDS, DS, EDBT, ER, FODO, FOIKS, FQAS, GIS, HPTS, ICDE, ICDM, ICDT, ICIS, IDA, IDEAL, IDEAS, IGIS, Inf. Process. Lett., Inf. Sci., Inf. Syst., IPM, IQIS, ISF, ISR, IW-MMDBMS, IWDM, JDM, JIIS, JMIS, K-CAP, KA, KDD, KER, KIS, KR, MDA, MFDBS, MLDM, MMDB, MSS, NLDB, OODBS, PAKDD, PKDD, PODS, RIDE, RIDS, SIGKDD Exp., SIGMOD, SIGMOD Rec., SSD, SSDBM, TKDE, TODS, TOIS, TSDM, UIDIS, VDB, VLDB, VLDB-J, WebDB, WIDM, WISE, XMLEC

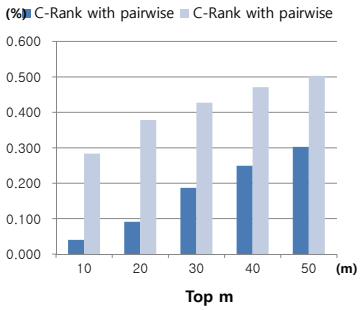
Accuracy Evaluation

- To select two data mining textbooks below and to select five chapters in the book
  - Jiawei Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann
  - P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Addison-Wesley
  - Chapters: clustering, sequential pattern mining, spatial databases, link mining, graph pattern mining
- To regard the reference papers in each chapter as ground truth
  - For each reference paper in a chapter, other papers in the same chapter are regarded as its similar papers
- Evaluation
  - To select a reference paper in a chapter as a query paper
  - To find the top-m (=10, 20, 30, 40, 50) papers from DBLP-DB, which are considered similar to a query paper by each method
  - To compare the top-m papers found and the ground truth papers for each method

Comparison of Normalization Methods

- Pair-wise normalization vs. Jaccard-based normalization
  - With P-Rank and C-Rank





Example: Top-10 Similar Papers

Rank	Paper Title
Query	BIRCH: an Efficient Data Clustering Method for Very Large ...
1	Efficient and Effective Clustering Methods ...
2	CURE: An Efficient Clustering Algorithm ...
3	A Density-Based Algorithm for Discovering Clusters ...
4	Automatic Subspace Clustering of High Dimensional ...
5	Scaling Clustering Algorithms to Large Databases
6	WaveCluster: A Multi-Resolution Clustering Approach ...
7	Fast Algorithms for Projected Clustering
8	STING: A Statistical Information Grid Approach ...
9	An Efficient Approach to Clustering in Large ...
10	OPTICS: Ordering Points To Identify the Clustering...

Example: Top-10 Similar Papers

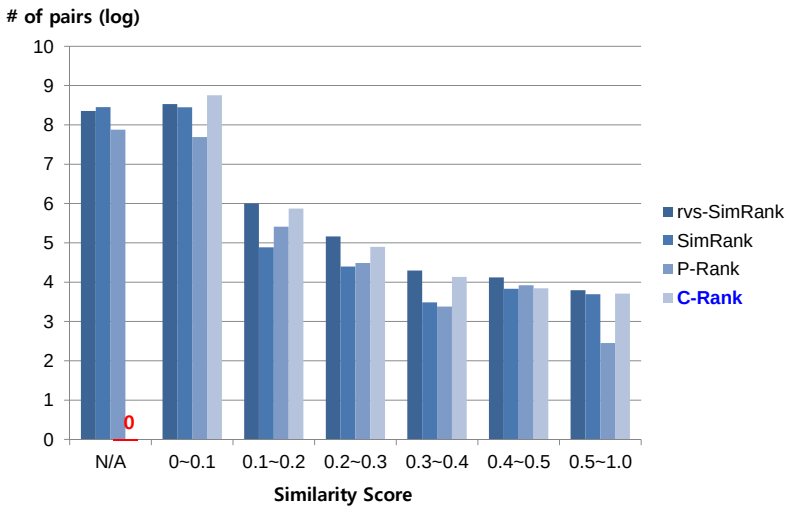
- Previous methods
  - BIRCH: An Efficient Data Clustering Method for Very Large Databases

SimRank	rvs-SimRank	P-Rank
CURE: An Efficient Clustering Algorithm for Large Databases.	A Unified Notion of Outliers: Properties and Computation.	A Unified Notion of Outliers: Properties and Computation.
WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases.	Cure: An Efficient Clustering Algorithm for Large Databases.	Cure: An Efficient Clustering Algorithm for Large Databases.
Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification.	A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.	A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.
An Efficient Approach to Clustering in Large Multimedia Databases with Noise.	Scaling Clustering Algorithms to Large Databases.	Scaling Clustering Algorithms to Large Databases.
Efficient and Effective Clustering Methods for Spatial Data Mining.	ROCK: A Robust Clustering Algorithm for Categorical Attributes.	WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases.
Scaling Clustering Algorithms to Large Databases.	WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases.	ROCK: A Robust Clustering Algorithm for Categorical Attributes.
STING: A Statistical Information Grid Approach to Spatial Data Mining.	A Linear Method for Deviation Detection in Large Databases.	Efficient Algorithms for Discovering Association Rules.
Streaming-Data Algorithms for High-Quality Clustering.	MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases.	A Linear Method for Deviation Detection in Large Databases.
A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.	Efficient Algorithms for Discovering Association Rules.	What Makes Patterns Interesting in Knowledge Discovery Systems.
A Linear Method for Deviation Detection in Large Databases.	Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes.	Mining Association Rules between Sets of Items in Large Databases.

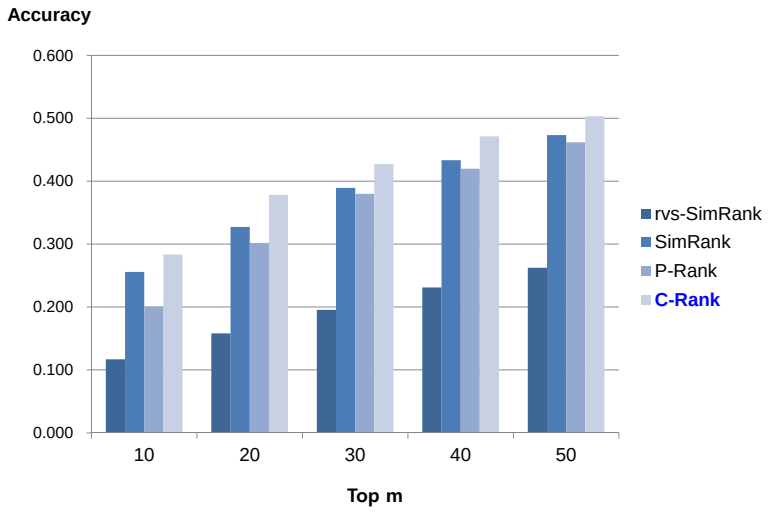
Example: Top-10 Similar Papers

Rank	Paper Title
Query	R-Trees: A Dynamic Index Structure for Spatial Searching
1	The R*-Tree: An Efficient and Robust Access Method ...
2	The R+-Tree: A Dynamic Index for Multi-Dimensional ...
3	Nearest Neighbor Queries
4	The K-D-B-Tree: A Search Structure For Large ...
5	The X-tree : An Index Structure or ...
6	On Packing R-trees
7	The Grid File: An Adaptable, Symmetric Multikey ...
8	Efficient Processing of Spatial Joins Using R-Trees
9	Hilbert R-tree: An Improved R-tree using Fractals
10	The SR-tree: An Index Structure for High-Dimensional ...

Number of Not-Applicable Pairs



Accuracy



Application 1: Paper Clustering by using Link-Based Similarity Measure

Environment

- Data
  - DBLP_DB: Academic literature data
- Similarity measure
  - C-Rank
- Network construction
  - For each paper, we made *links to its 30 most similar papers*
    - Where their similarities are used as the weights of links
- Clustering algorithm
  - Chameleon

Cluster analysis

- To select the clusters below from the clustering result
  - Clustering
  - Frequent pattern mining
  - Graph mining
  - Moving object management
  - Privacy preserving data mining
- Qualitative analysis
  - To examine the topic in each cluster by sampling papers randomly
  - To find the top-10 representative authors and keywords from each cluster

Papers Randomly Sampled in Clusters

Topics	Clustering	Frequent pattern mining	Moving Object	Privacy preserving data mining	Graph mining
#objects	441	228	191	140	135
1	Non-Redundant Data Clustering.	CT-ITL : Efficient Frequent Item Set Mining Using a Compressed Prefix Tree with Pattern Growth.	Moving Objects in Networks Databases.	Enhancing User Privacy Through Data Handling Policies.	Indexing and Mining Free Trees.
2	Effective and Efficient Distributed Model-Based Clustering.	Mining Frequent Closed Patterns in Microarray Data.	Aggregation and comparison of trajectories.	Privacy and Ownership Preserving of Outsourced Medical Data.	Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism.
3	CACTUS - Clustering Categorical Data Using Summaries.	Information-Based Classification by Aggregating Emerging Patterns.	ASPEN: an adaptive spatial peer-to-peer network.	Privacy-Preserving Top-K Queries.	Efficient Discovery of Common Substructures in Macromolecules.
4	An Incremental Hierarchical Data Clustering Algorithm Based on Gravity Theory.	Distribution-Based Synthetic Database Generation Techniques for Item-set Mining.	Modeling and Querying Moving Objects.	Improved Privacy-Preserving Bayesian Network Parameter Learning on Vertically Partitioned Data.	Mining for Tree-Query Associations in a Graph.
5	Electricity Based External Similarity of Categorical Attributes.	From frequent itemsets to semantically meaningful visual patterns.	R-trees with Update Memos.	Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking.	MARGIN: Maximal Frequent Subgraph Mining.
6	On the complexity of finding balanced one-way cuts.	Statistical Supports for Frequent Itemsets on Data Streams.	STRIPES: An Efficient Index for Predicted Trajectories.	Privacy Preserving Nearest Neighbor Search.	Razor: mining distance-constrained embedded subtrees.
7	On the Efficiency of Best-Match Cluster Searches.	Mining Top-k Covering Rule Groups for Gene Expression Data.	A data model for multi-dimensional transportation applications.	Ask a Better Question, Get a Better Answer A New Approach to Private Data Analysis.	Discovering frequent topological structures from graph datasets.
8	Hierarchical Taxonomy Preparation for Text Categorization Using Consistent Bipartite Spectral Graph Copartitioning.	Optimization of Constrained Frequent Set Queries with 2-variable Constraints.	Relaxed space bounding for moving objects: a case for the buddy tree.	Deriving Private Information from Arbitrarily Projected Data.	Clustering Document Images Using Graph Summaries.
9	Efficient Disk-Based K-Means Clustering for Relational Databases.	On compressing frequent patterns.	Querying Imprecise Data in Moving Object Environments.	On the Design and Quantification of Privacy Preserving Data Mining Algorithms.	A Quantitative Comparison of the Sub-graph Miners MoFa, gSpan, FFSM, and Gaston.
10	Iterative Projected Clustering by Subspace Mining.	Research issues in data stream association rule mining.	Indexing Animated Objects Using Spatiotemporal Access Methods.	Revealing information while preserving privacy.	Graph Indexing: A Frequent Structure-based Approach.

Representative Authors and Keywords

Topic	Clustering				
Rank	Author	RWR Score	Frequency	Keyword	Frequency
1	Jiawei Han	9.644	12	clustering	283
2	Hans-Peter Kriegel	6.083	29	data	132
3	Martin Ester	4.919	8	algorithm	46
4	Xiaowei Xu	4.769	6	large	41
5	Jorg Sander	3.227	12	hierarchical	37
6	Inderjit S. Dhillon	3.212	9	mining	35
7	Philip S. Yu	2.740	15	clusters	34
8	Charu C. Aggarwal	2.262	8	high	32
9	Wei Wang	1.822	9	cluster	31
10	Jiong Yang	1.737	6	dimensional	30

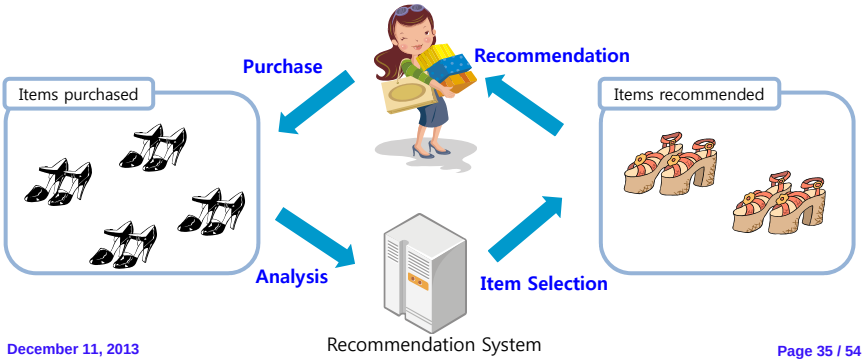
Representative Authors and Keywords

Topic	Frequent pattern mining				
Rank	Author	RWR Score	Frequency	Keyword	Frequency
1	Jiawei Han	4.367	25	frequent	125
2	Jian Pei	3.115	13	mining	113
3	Heikki Mannila	1.414	8	patterns	55
4	Mohammed Javeed Zaki	1.043	9	itemsets	44
5	Laks V. S. Lakshmanan	1.014	7	pattern	39
6	Toon Calders	0.891	6	data	37
7	Jianyong Wang	0.499	6	closed	24
8	Osmar R. Zaiane	0.486	8	efficient	24
9	Hong Cheng	0.483	7	itemset	23
10	Anthony K. H. Tung	0.471	6	algorithm	17

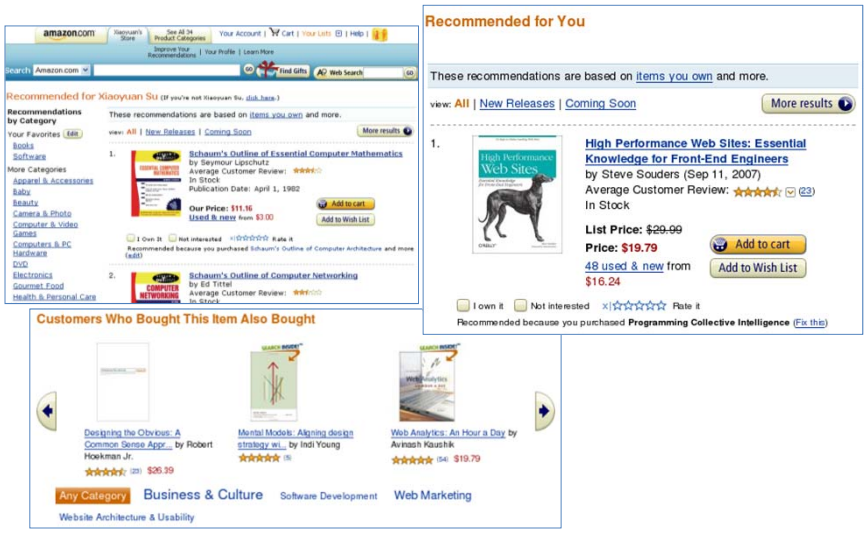
Application 2: Recommendation Using Link-Based Similarity Measure

Background

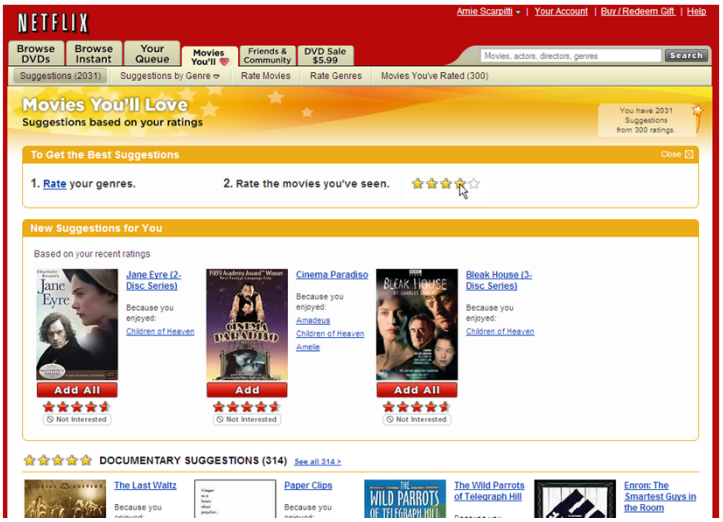
- Recommendation systems
  - To predict the degree of preferences on items that a target customer did not purchase yet
  - To recommend the top-k items to the customer



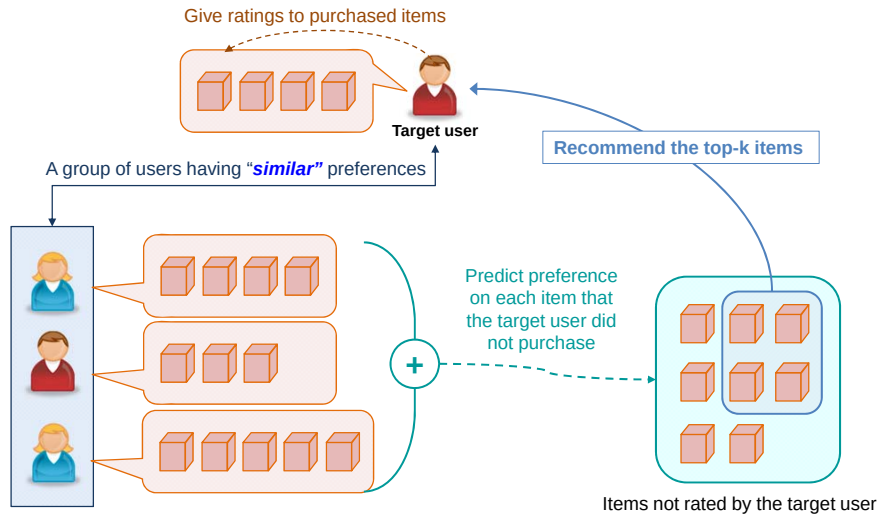
Example – Amazon.com



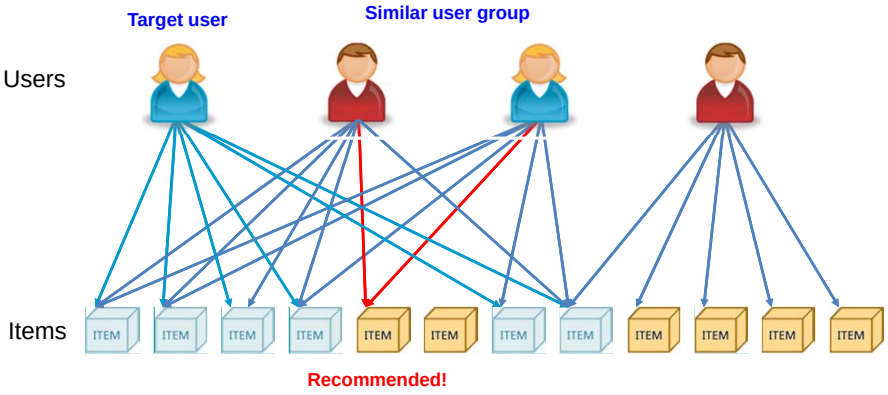
Example – Netflix



Collaborative Filtering



Collaborative Filtering



Data Sparsity

- Density in price-comparison shopping data
  - 9,997 users
  - 310,841 items
  - 349,167 user-item pairs
  - **Density: 0.01%**
- Collaborative filtering suffers from the data sparsity problem in this case

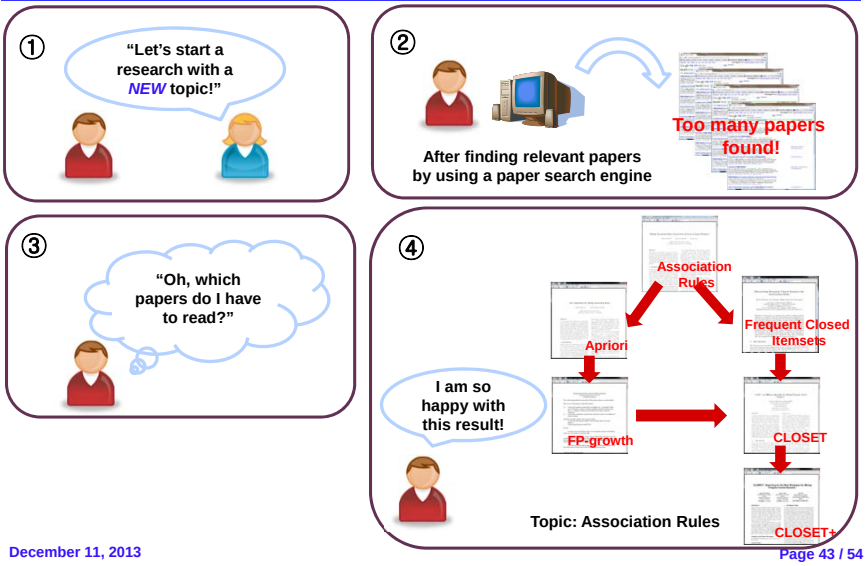
Evaluation: Accuracy

- Accuracy (%)

	Recall@10	Recall@20	Precision@10	Precision@20
Our Approach	**23.83**	**27.67**	**3.70**	**2.15**
Graph-Based RS	13.17	18.17	1.90	1.40
User-Based CF	16.16	21.87	3.15	2.19
Item-Based CF	12.00	19.06	2.25	1.68

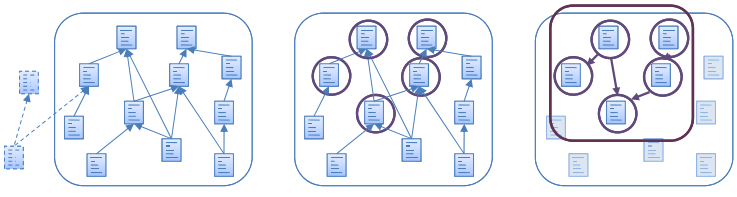
Application 3: Seminal Paper Genealogy by using Link-Based Similarity Measure

Motivation



Building Seminal Paper Genealogy

- Problem formulation
  - Given
    - Papers belonging to the same topic
    - A citation graph
    - *k*, the number of seminal papers
  - Find
    - *k* representative seminal papers and their genealogy
- Overview



(a) Extracting papers belonging to the same topic. (b) Finding seminal papers. (c) Constructing seminal paper genealogy.

Extracting Papers of the Same Topic

- Unsupervised clustering using text or link-based similarity
  - Spectral clustering, modularity-based clustering, and Chameleon
- Finding *n* papers most similar to a few key papers in a topic
  - k-nearest neighbor searching

Finding Seminal Papers

- Our wish list
  1. Find such papers *cited by many papers and/or highly cited papers*
  2. Find such papers *cited by relevant papers in the same topic*
  3. Find such papers *cited by the papers that are published much later*
  4. Include young seminal papers *even though they do not get a lot of citations*
- We achieved this by **ArtRank**, our own ranking algorithm

- **Seminal paper list on the clustering topic**

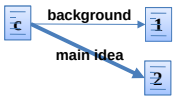
Title	Author	Publisher	Year
Efficient and Effective Clustering Methods for Spatial Data Mining	Raymond T. Ng, Jiawei Han	VLDB	1994
A Database Interface for Clustering in Large Spatial Databases	Martin Ester, Hans-Peter Kriegel, Xiaowei Xu	KDD	1995
A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise	Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu	KDD	1996
BIRCH: An Efficient Data Clustering Method for Very Large Databases	Tian Zhang, Raghu Ramakrishnan, Miron Livny	SIGMOD	1996
STING: A Statistical Information Grid Approach to Spatial Data Mining	Wei Wang, Jiong Yang, Richard Muntz	VLDB	1997
Scaling Clustering Algorithms to Large Databases	Paul S. Bradley, Usama M. Fayyad, Cory Reina	KDD	1998
Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications	Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan	SIGMOD	1998
CURE: An Efficient Clustering Algorithm for Large Databases	Sudipto Guha, Rajeev Rastogi, Kyuseok Shim	SIGMOD	1998
Algorithms for Mining Distance-Based Outliers in Large Datasets	Edwin M. Knorr, Raymond T. Ng	VLDB	1998
Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values	Zhexue Huang	DMKD	1998

- **Seminal paper list on the clustering topic**

Title	Author	Publisher	Year
OPTICS: Ordering Points To Identify the Clustering Structure	Mihael Ankerst, MarkusM. Breunig, Hans-Peter Kriegel, Jorg Sander	SIGMOD	1999
Fast Algorithms for Projected Clustering	Charu C. Aggarwal, Cecilia Magdalena Procopiuc, Joel L. Wolf, Philip S. Yu, Jong Soo Park	SIGMOD	1999
Chameleon: Hierarchical Clustering using Dynamic Modeling	George Karypis, Eui-Hong Han, Vipin Kumar	IEEE Computer	1999
Clustering Data Streams: Theory and Practice	Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, Liadan O'Callaghan	IEEE TKDE	2000
Biclustering of Expression Data	Y. Cheng, G.M. Church	ISMD	2000
LOF: Identifying Density-Based Local Outliers	Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jorg Sander	SIGMOD	2000
NiagaraCQ: A Scalable Continuous Query System for Internet Databases	Jianjun Chen, David J. DeWitt, Feng Tian, Yuan Wang	SIGMOD	2000
ROCK: A Robust Clustering Algorithm for Categorical Attributes	Sudipto Guha, Rajeev Rastogi, Kyuseok Shim	Inf. Syst.	2000
Models and Issues in Data Stream Systems	Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, Jennifer Widom	PODS	2002
A Framework for Clustering Evolving Data Streams	Charu C. Aggarwal, Jiawei Han, Jianyong Wang, Philip S. Yu	VLDB	2003

Constructing Genealogy

- A new paper is influenced by the contribution of previously published papers



  - Finding significant influence relationships is a key of constructing genealogy

- Procedure
  1. For paper *c*, measure the degree of influence from every paper *p* cited by paper *c*
  2. Select the top *m* papers having the largest influence scores (parent papers)
  3. Draw directed links from the parent papers to child paper *c*

Influence Metrics

- Requirements
  1. Should reflect the strength of the influence correctly
     - If a cited paper has a great influence on a citing paper, the influence score should be high
  2. Should consider the temporal distance between citing and cited papers
     - If the difference of publication years is large, the influence score should be small
     - We do not want to lose the true ***influence chains***
  3. Should be able to compute the influence between all the pairs of papers having citation relationships
     - No not-applicable cases

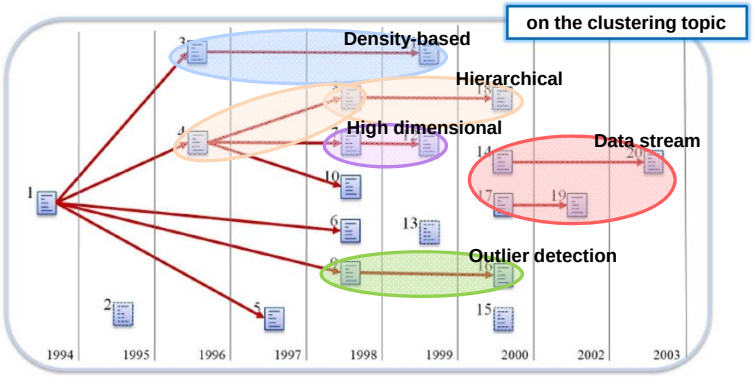
Influence Metrics

- Summary

	R1	R2	R3	
Cosine similarity	O	X	O	Text similarity
Bibliographic coupling	O	O	X	Link-based similarity
Co-citation	O	O	X	
Amsler	O	O	X	
SimRank	O	O	X	
rvs-SimRank	O	O	X	
P-Rank	O	O	X	
C-Rank	**O**	**O**	**O**	
Keyword-Extension	O	△	O	Combined similarity

Results

- Constructing Paper Genealogy using C-Rank



  - C-Rank produces a nice genealogy
    - It separates a whole topic into sub-topics appropriately

Summary

- Link-Based Similarity Measures in Academic Literature Data
  - Previous Methods
  - C-Rank: Proposed Method
- Applications of Link-Based Similarity Measures
  - Paper Clustering
  - Recommendation
  - Paper Genealogy Construction
- On going work
  - To combine the links and contents together in a sophisticated way for similarity computations

Thank You !