A Survey of Network Mining

Xinyu Ma
Institute of Computing Technology
Chinese Academy of Science
Beijing, P.R.C
xinyuma2016@gmail.com

1. Problem Definition

1.1 What is Network? (Not Computer Network)

Most real systems usually consist of a large number of interacting, multi-typed components, such as human social activities, communication and computer systems, and biological networks. In such systems, the interacting components constitute interconnected networks, which can be called **information networks** without loss of generality.

DEFINITION 1: **Information Network**. An information network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\varphi : V \rightarrow A$ and a link type mapping function $\psi : E \rightarrow R$. Each object $v \in V$ belongs to one particular object type in the object type set $A: \varphi(v) \in A$, and each link $e \in E$ belongs to a particular relation type in the relation type set $R: \psi(e) \in R$. If two links belong to the same relation type, the two links share the same starting object type as well as the ending object type.

DEFINITION 2: **Heterogeneous/homogeneous information Network**. The information network is called heterogeneous information network if the types of objects $|A| > 1$ or the types of relations $|R| > 1$; otherwise, it is a **homogeneous information network**.

However, with the boom of social network analysis, all kinds of networked data have emerged, and numbers of concepts to model networked data have been proposed. Here we compare heterogeneous network concept with these related concepts.

Heterogeneous network vs multi-relational network. Different from heterogeneous network, multi-relational network has only one type of objects, but more than one kind of relationship between objects. So multi-relational network can be seen as a special case of heterogeneous network.

Heterogeneous network vs multi-dimensional/mode network. Tang et al. proposed the multi-dimensional/mode network concept, which has the same meaning with multi-relational network. That is, the network has only one type of objects and more than one kind of relationship between objects. So multi-dimensional/mode network is also a special case of heterogeneous network.

Heterogeneous network vs composite network. Qiang Yang et al. proposed the composite network concept, where users in networks have various relationships, exhibit different behaviors in each individual network or subnetwork, and share some common latent interests across networks at the same time. So composite network is in fact a multi-relational network, a special case of heterogeneous network.

Heterogeneous network vs complex network. A complex network is a network with nontrivial topological features and patterns of connection between its elements that are neither purely regular nor purely random. Such non-trivial topological features include a heavy tail in the degree distribution, a high clustering coefficient, community structure, and hierarchical structure. The studies of complex networks have brought together researchers from many areas including mathematics, physics, biology,

computer science, sociology, and others. The studies show that many real networks are complex networks, such as social networks, information networks, technological networks, biological networks, and so on. So we can say that many real heterogeneous networks are complex networks.

So, we can see that heterogeneous network include most of the other network, except for complex network. However, the studies on complex networks usually focus on the structures, functions, and features of networks.

1.2 What to Mine?

Recently, the information network analysis has become a hot research topic in data mining and information retrieval fields in the past decades. The basic paradigm is to mine hidden patterns through mining link relations from networked data. The analysis of information network is related to the works in link mining and analysis, social network analysis, hypertext and web mining, network science, and graph mining.

Particularly in heterogeneous network, Yizhou Sun and her team have analyzed more than 100 papers in this field, and divided them into **7 categories** according to their data mining tasks.

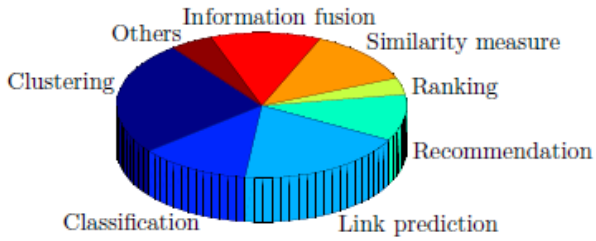

Fig 1. Paper distribution of heterogeneous information network analysis on different data mining tasks.

A. Similarity Measure

Similarity measure is to evaluate the similarity of objects. It is the basis of many data mining tasks, such as web search, clustering, and product recommendation.

B. Clustering

Clustering analysis is the process of partitioning a set of data objects (or observations) into a set of clusters, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters.

C. Classification

Classification is a data analysis task where a model or classifier is constructed to predict class (categorical) labels. Traditional machine learning has focused on the classification of identically-structured objects satisfying independent identically distribution (IID). However, links exist among objects in many real-world datasets, which makes objects not satisfy IID. So link based object classification has received considerable attention, where a data graph is composed of a set of objects connected to each other via a set of links.

D. Link Prediction

Link prediction is a fundamental problem in link mining that attempts to estimate the likelihood of the existence of a link between two nodes, based on observed links and the attributes of nodes.

Link prediction is often viewed as a simple binary classification problem: for any two potentially linked objects, predict whether the link exists (1) or not (0).

E. Ranking

Ranking is an important data mining task in network analysis, which evaluates object importance or popularity based on some ranking functions. Ranking in heterogeneous information networks is an important and meaningful task, but faces several challenges. First, there are different types of objects and relations in HIN, and treating all objects equally will mix different types of objects together. Second, different types of objects and relations in HIN carry different semantic meanings, which may lead to different ranking results.

F. Recommendation

Recommender systems help consumers to make product recommendations that are likely to be of interest to the user such as books, movies, and restaurants. It uses a broad range of techniques from information retrieval, statistics, and machine learning to search for similarities among items and customer preferences.

G. Information Fusion

Information fusion denotes the process of merging information from heterogeneous sources with differing conceptual, contextual and typographical representations. Due to the availability of various data sources, fusing these scattered distributed information sources has become an important research problem. In the past decades, dozens of papers have been published on this topic in many traditional data mining areas, e.g., data schemas integration in data warehouse, protein-protein interaction (PPI) networks and gene regulatory networks matching in bioinformatics, and ontology mapping in web semantics. Nowadays, with the surge of HIN, information fusion across multiple HINs has become a novel yet important research problem. By fusing information from different HINs, we can obtain a more comprehensive and consistent knowledge about the common information entities shared in different HINs, including their structures, properties, and activities.

H. Other Tasks

Besides the tasks discussed above, there are many other applications in heterogeneous networks, such as **influence propagation**, **privacy risk problem and sematic mining**.

1.3 Where to use?

Recommender systems. Public opinion analysis. Social network. Web search and etc.

2. Evaluation Metric

A. Similarity measure metric

   Euclidean distance, Manhattan distance, Cosine, Jaccard Similarity Coefficient, information entropy and etc.

B. Clustering metric

   1. Not given label

Compactness, Separation, Davies-Bouldin Index, Dunn Validity Index, Modularity

2. Given label

Cluster Accuracy, Rand index, Normalized Mutual Information

Different problems need to be treated differently!

C. Classification metric

Precision, Recall, F-score, Accuracy, ROC, AUC, etc.

D. Link Prediction

AUC, Precision, Ranking Score

E. Influence Propagation

Running Time, Memory Usage, The Accuracy of Spread Prediction

3. Related Work

A. Similarity Measure

1.Traditional algorithm

Feature based: cosine similarity, Jaccard coefficient, and Euclidean distance

Link based: Personalized PageRank, SimRank

2.State-of-art

Path based: PathSim, PCRW, HeteSim, and etc

B. Clustering(community detection) NP-hard

1.Traditional algorithm

k-means, normalized cuts, modularity(spectral method, greedy method and sampling technique)

3.Integrate HIN

TCSC(attribute), LSA-PTM(text), SemiRPClus(guide), RankClus, NetClus and etc.

C. Classifccation

1.Traditional algorithm

GNetMine, HetPathMine, IMBHN

2.Integrate HIN

RankClass, F-RankClass

D. Link Prediction

1.Traditional algorithm(on one HIN)

PathPredict(meta-path), MRIP and TFGM (Probabilistic models),JMF(Matrix factorization)

2.across multiple aligned HIN

SCAN-PS, TRAIL, COSNET

E. Influence Propagation

M&M

4. Benchmark

4.1 Datasets

1. DBLP: This dataset provides bibliographical information about computer science journals and

proceedings. It includes 50,000 objects.

2. Yelp Dataset: The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes.
3. SNAP：Stanford Large Network Dataset Collection
   http://snap.stanford.edu/data/index.html
4. UCI Machine Learning Repository: This is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.
   http://archive.ics.uci.edu/ml/datasets.html
5. Social network dataset in China：http://www.socialysis.org/data/project/project
6. Wikipedia：http://download.wikipedia.org
7. DBpedia: http://wiki.dbpedia.org/
8. YAGOhttps://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/#c10444

4.2 Experiments

A. Similarity

Table 1. TOP 10 RELATED AUTHORS TO “CHRISTOS FALOUTSOS” BASED ON “APVCPVA” PATH ON ACM DATASET

	HeteSim		PathSim		PCRW		SimRank	
Rank	Author	Score	Author	Score	Author	Score	Author	Score
1	Christos Faloutsos	1	Christos Faloutsos	1	Charu C. Aggarwal	0.0063	Christos Faloutsos	1
2	Srinivasan Parthasarathy	0.9937	Philip Yu	0.9376	Jiawei Han	0.0061	Edoardo Airoldi	0.0789
3	Xifeng Yan	0.9877	Jiawei Han	0.9346	Christos Faloutsos	0.0058	Leejay Wu	0.0767
4	Jian Pei	0.9857	Jian Pei	0.8956	Philip Yu	0.0056	Kensuke Onuma	0.0758
5	Jiong Yang	0.9810	Charu C. Aggarwal	0.7102	Alia I. Abdelmoty	0.0053	Christopher R. Palmer	0.0699
6	Ruoming Jin	0.9758	Jieping Ye	0.6930	Chris B. Jones	0.0053	Anthony Brockwell	0.0668
7	Wei Fan	0.9743	Heikki Mannila	0.6928	Jian Pei	0.0034	Hanghang Tong	0.0658
8	Evimaria Terzi	0.9695	Eamonn Keogh	0.6704	Heikki Mannila	0.0032	Evan Hoke	0.0651
9	Charu C. Aggarwal	0.9668	Ravi Kumar	0.6378	Eamonn Keogh	0.0031	Jia-Yu Pan	0.0650
10	Mohammed J. Zaki	0.9645	Vipin Kumar	0.6362	Mohammed J. Zaki	0.0027	Roberto Santos Filho	0.0648

B. Clustering

Table 2. Object clustering performance of different methods on (a) DBLP and (b) NSF-Awards datasets.

(a) DBLP

Object	Paper (%)		Author (%)		Venue (%)		Average (%)	
Metric	AC	NMI	AC	NMI	AC	NMI	AC	NMI
NMF	44.55	22.92	-	-	-	-	44.55	22.92
PLSA	59.45	32.75	*65.0*	*37.97*	*80.0*	*74.74*	68.15	48.49
LapPLSI	61.35	33.93	-	-	-	-	60.70	33.37
LDA	47.00	20.48	-	-	-	-	47.00	20.48
ATM	77.00	52.21	74.13	40.67	-	-	75.57	46.44
NetClus	65.00	40.96	70.82	47.43	79.75	76.69	71.86	55.03
TMBP-RW	73.10	53.13	82.59	67.76	81.75	**77.53**	79.15	66.14
TMBP-Regu	**79.15**	**59.16**	**89.81**	**74.25**	**82.75**	76.56	**83.90**	**69.99**

(b) NSF-Awards

Object	Doc (%)	
Metric	AC	NMI
NMF	45.97	40.92
PLSA	63.00	64.48
LapPLSI	63.65	64.58
LDA	65.06	63.36
ATM	65.69	69.58
NetClus	63.51	66.11
TMBP-RW	64.84	68.74
TMBP-Regu	65.15	69.83

C. Classification-RankClus

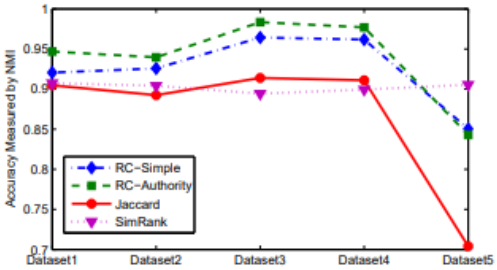

Fig 3. Accuracy of Clustering

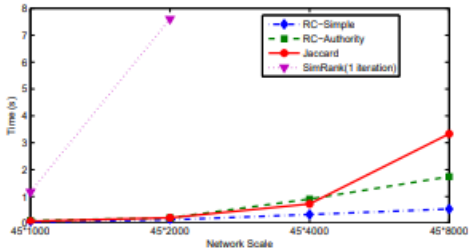

Fig 4. Efficiency Analysis

D. Link Prediction

Table 2. Performance comparison of different methods for inferring social and location links for Foursquare of different remaining information rates. The anchor link sample rate ρ is set as 1.0.

link	measure	methods	remaining information rates σ 0.1	0.2	0.3	0.4	0.5	0.6	0.7
social	AUC	TRAIL	**0.810±0.012**	**0.824±0.009**	**0.837±0.008**	**0.844±0.009**	**0.832±0.003**	**0.852±0.009**	**0.847±0.009**
		TRAILT	0.691±0.040	0.684±0.039	0.704±0.033	0.729±0.006	0.718±0.020	0.732±0.005	0.730±0.008
		TRAILS	0.572±0.007	0.578±0.007	0.580±0.004	0.575±0.012	0.580±0.011	0.583±0.009	0.578±0.009
		SCAN	0.772±0.050	0.788±0.004	0.811±0.009	0.830±0.005	0.809±0.004	0.825±0.008	0.824±0.012
		SCANT	0.524±0.023	0.559±0.008	0.559±0.017	0.554±0.044	0.630±0.008	0.599±0.007	0.627±0.004
		SCANS	0.583±0.005	0.579±0.003	0.583±0.010	0.562±0.005	0.579±0.004	0.585±0.003	0.584±0.003
		CN	0.494±0.002	0.500±0.015	0.504±0.006	0.496±0.012	0.495±0.018	0.491±0.015	0.489±0.018
		JC	0.497±0.003	0.503±0.004	0.501±0.002	0.502±0.010	0.496±0.008	0.496±0.019	0.492±0.008
		AA	0.494±0.002	0.499±0.014	0.501±0.006	0.494±0.012	0.492±0.018	0.489±0.015	0.493±0.022
	Accuracy	TRAIL	**0.855±0.002**	**0.849±0.004**	**0.850±0.008**	**0.854±0.005**	**0.850±0.003**	**0.851±0.001**	**0.852±0.004**
		TRAILT	0.622±0.046	0.627±0.036	0.655±0.022	0.676±0.009	0.674±0.019	0.677±0.004	0.679±0.008
		TRAILS	0.548±0.004	0.551±0.006	0.552±0.004	0.549±0.000	0.551±0.002	0.553±0.003	0.544±0.001
		SCAN	0.747±0.003	0.752±0.007	0.748±0.000	0.754±0.008	0.746±0.005	0.745±0.007	0.747±0.003
		SCANT	0.512±0.009	0.522±0.002	0.520±0.001	0.537±0.006	0.554±0.008	0.542±0.003	0.567±0.007
		SCANS	0.557±0.002	0.547±0.006	0.553±0.002	0.545±0.006	0.552±0.007	0.551±0.002	0.551±0.004
		NAIVE	0.525±0.014	0.526±0.006	0.525±0.008	0.526±0.007	0.525±0.013	0.525±0.009	0.525±0.013
location	AUC	TRAIL	**0.848±0.005**	**0.856±0.010**	**0.870±0.010**	**0.878±0.007**	**0.899±0.007**	**0.886±0.022**	**0.887±0.009**
		TRAILT	0.839±0.006	0.850±0.003	0.857±0.009	0.866±0.008	0.862±0.005	0.871±0.005	0.869±0.003
		TRAILS	0.631±0.003	0.632±0.002	0.631±0.001	0.634±0.001	0.634±0.002	0.634±0.002	0.635±0.001
		SCAN	0.712±0.010	0.757±0.002	0.758±0.009	0.770±0.005	0.775±0.005	0.784±0.004	0.792±0.003
		SCANT	0.676±0.009	0.711±0.005	0.730±0.005	0.749±0.003	0.756±0.001	0.763±0.005	0.769±0.003
		SCANS	0.633±0.003	0.633±0.003	0.633±0.001	0.636±0.001	0.637±0.000	0.633±0.001	0.634±0.001
		FCF	0.598±0.008	0.638±0.015	0.638±0.005	0.654±0.012	0.664±0.007	0.661±0.007	0.664±0.010
	Accuracy	TRAIL	**0.719±0.004**	**0.736±0.001**	**0.749±0.006**	**0.754±0.003**	**0.753±0.002**	**0.760±0.002**	**0.761±0.002**
		TRAILT	0.674±0.009	0.697±0.004	0.706±0.005	0.709±0.001	0.717±0.006	0.716±0.007	0.717±0.002
		TRAILS	0.536±0.003	0.527±0.001	0.537±0.005	0.553±0.003	0.560±0.002	0.565±0.000	0.566±0.001
		SCAN	0.658±0.000	0.670±0.002	0.682±0.001	0.697±0.003	0.699±0.003	0.723±0.003	0.723±0.007
		SCANT	0.610±0.001	0.623±0.001	0.631±0.001	0.647±0.001	0.653±0.002	0.671±0.003	0.676±0.002
		SCANS	0.536±0.025	0.531±0.008	0.535±0.002	0.547±0.004	0.557±0.004	0.565±0.001	0.566±0.001
		NAIVE	0.536±0.014	0.536±0.002	0.536±0.001	0.537±0.008	0.536±0.012	0.536±0.009	0.537±0.019

5. Conclusion

There is a surge on heterogeneous information network analysis in recent years because of rich structural and semantic information in this kind of networks. There is still no structural research system because so many fields need to be mined and so many unknowns. I see challenge, I see opportunity.

6. References

[1] Chuan Shi, “A Survey of Heterogeneous Information Network Analysis” in TKDE, 2015

[2] Y. Sun and J. Han, “Mining heterogeneous information networks: a structural analysis approach,” SIGKDD Explorations, vol. 14, no. 2, pp. 20–28, 2012.

[3] L. Getoor and C. P. Diehl, “Link mining: a survey,” SIGKDD Explorations, vol. 7, no. 2, pp. 3–12,

2005.
[4] D. Jensen and H. Goldberg, AAAI Fall Symposium on AI and Link Analysis. AAAI Press, 1998.
[5] R. Feldman, “Link analysis: Current state of the art,” Tutorial at the KDD, vol. 2, 2002.
[6] S. Wasserman, Social network analysis: Methods and applications. Cambridge university press, 1994.
[7] E. Otte and R. Rousseau, “Social network analysis: a powerful strategy, also for the information sciences,” Journal of information Science, vol. 28, no. 6, pp. 441–453, 2002.
[8] S. Chakrabarti et al., Mining the Web: Analysis of hypertext and semi structured data. Morgan Kaufmann, 2002.
[9] T. G. Lewis, Network science: Theory and applications. John Wiley & Sons, 2011.
[10] D. J. Cook and L. B. Holder, “Graph-based data mining,” IEEE Intelligent Systems, vol. 15, no. 2, pp. 32–41, 2000.
[11] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, “New perspectives and methods in link prediction,” in KDD, 2010, pp. 243–252.
[12] V. Leroy, B. B. Cambazoglu, and F. Bonchi, “Cold start link prediction,” in KDD, 2010, pp. 393–402.
[13] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu, “Relevance search in heterogeneous networks,” in EDBT, 2012, pp. 180–191.
[14] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu, “Pathsim: Meta path-based top-k similarity search in heterogeneous information
networks,” in VLDB, 2011, pp. 992–1003.
[15] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, “Integrating meta-path selection with user-guided object clustering in heterogeneous information networks,” in KDD, 2012, pp. 1348–1356.
[16] X. Kong, P. S. Yu, Y. Ding, and D. J. Wild, “Meta path-based collective classification in heterogeneous information networks,” in CIKM, 2012, pp. 1567–1571.
[17] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, “RankClus: integrating clustering with ranking for heterogeneous information network analysis,” in EDBT, 2009, pp. 565–576.
[18] Y. Sun and J. Han, Mining Heterogeneous Information Networks: Principles and Methodologies. Morgan & Claypool Publishers, 2012.
[19] Y. Sun and J. Han, “Meta-path-based search and mining in heterogeneous information networks,” Tsinghua Science and
Technology, vol. 18, no. 4, pp. 329–338, 2013.
[20] Y. Sun, Y. Yu, and J. Han, “Ranking-based clustering of heterogeneous information networks with star network schema,” in KDD, 2009, pp. 797–806.
[21] Y. Li, C. Shi, S. Y. Philip, and Q. Chen, “Hrank: A path based ranking method in heterogeneous information network,” in WAIM, 2014, pp. 553–565.
[22] J. Tang, H. Gao, X. Hu, and H. Liu, “Exploiting homophily effect for trust prediction,” in WSDM, 2013, pp. 53–62.
[23] I. Konstas, V. Stathopoulo, and J. M. Jose, “On social networks and collaborative recommendation,” in SIGIR, 2009, pp. 195–202.
[24] D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks,” in CIKM, 2003, pp. 556–559.
[25]Similarity metric: https://blog.csdn.net/txwh0820/article/details/51791739
[26]Cluster metric: https://blog.csdn.net/u012102306/article/details/52423074