International Journal of Engineering & Technology, 7 (2.6) (2018) 293-300

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A Survey of Data Mining Techniques on Information Networks

Sadhana Kodali[1]*, Madhavi Dabbiru[2], B Thirumala Rao[3]

[1]PhD scholar,Department of CSE,Koneru Lakshmaiah Education foundation,Vaddeswaram,Guntur Dt,Andhra Pradesh,India.
[2] Professor,Dept of CSE,Dr L Bullayya College of Engineering for Women,Visakhapatnam,Andhra Pradesh,India.drlbcse@gmail.com
[3] Professor,Department of CSE,Koneru Lakshmaiah Education foundation,Vaddeswaram,Guntur Dt,Andhra Pradesh,India.
**Corresponding author E-mail:sadhanalendicse@gmail.com*

Abstract

An Information Network is the network formed by the interconnectivity of the objects formed due to the interaction between them. In our day-to-day life we can find these information networks like the social media network, the network formed by the interaction of web objects etc. This paper presents a survey of various Data Mining techniques that can be applicable to information networks. The Data Mining techniques of both homogeneous and heterogeneous information networks are discussed in detail and a comparative study on each problem category is showcased.

Keywords: *InformationNetworks,DataMining Techniques, Homogeneous Information Networks,HeterogeneousInformation Networks*

1. Introduction

When different objects interact and communicate with each other it forms an information network. The examples include the social networks where different people interact with each other, an author-paper network where the communication is between the authors and conferences related via a paper, the DNA structure in a biological network etc. The object relationships in the database form an information network that may be categorized as homogeneous (formed between objects of same type) or heterogeneous (formed between objects of different types which may also be distributed at various locations).An information network is the set of objects that are related to each other and these objects can communicate and interact with each other. There may exist a set of objects of different types and a set of relationships. Therefore an information network is an object from the set of objects that can communicate with another object in a different set in the form of a directed graph. The edge is the relationship that exists between these objects. An information network is categorized as homogeneous if the objects belong to the same set and the relationship belongs to the same type of relationship. Otherwise the information network is said to be heterogeneous which means that there may be more than one type of objects or the objects may communicate with more than one type of relationship. A homogeneous network can be derived from a heterogeneous network by excluding the diversified object features. The link mining techniques are based on homogeneous network analysis and can also be extended to heterogeneous networks. If a network has only multiple interactions among the same type of objects it is called a multi relational network and can be treated as a special case of heterogeneous network. The multi-relational network is equivalent to a multi dimensional network. A network can be composite if the relationships exhibited between the objects depend on the context in a sub-network. And the example of a social media network and genetically structured biological networks are treated as complex networks. The heterogeneous networks can be formed from objects that can be derived from and are not restricted to only structured, semi-structured data but also to unstructured data. A combined path from one object to another with interlinking relationships is called as the meta-path between the objects. A meta-path from one object to another is symmetric if the path is equal to its inverse. The information network adheres to a template called the network schema. The meta-path is a semantic and a significant feature of the heterogeneous information network. The network schema is a mapping from the source object to the target object using a directed link.

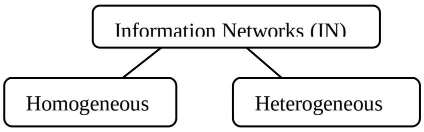

Figure1: Types of Information Networks.

The Figure1 shows the classification of the Information Networks and the Figure 2 depicts the type of Data-Mining techniques like similarity, clustering, classification and it can also be extended for recommender systems.

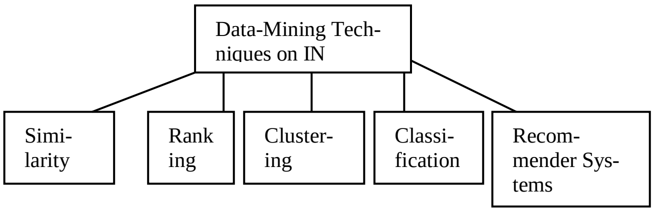

Figure2: Knowledge extraction techniques on Information Networks.

The various algorithms corresponding to each technique are discussed and compared in the preceding sections. The rest of the paper is organized as: Section 2 discusses the Similarity measures Section 3 focuses on Ranking, Section 4 gives an insight to clustering ,Section 5 a briefing about classification, Section 6 about Recommender Systems , followed by conclusions.

2.Similarity Measure

Similarity measure estimates the similarity or the likeliness between the objects. Similarity measure has its applications in web-search, clustering and product recommendation [1].Similarity measures can be of two types: Feature based similarity and web based similarity. In the feature based similarity the objects similarity are measured with cosine similarity, Euclidean distance, Jaccard coefficient. If the similarity is measured with the help of link structures in the graph it is called web based similarity. Similarity measure plays a vital role in product search and web search**.**

The SimRank algorithm [2] states that two objects are similar if they are related to similar objects. The SimRank works on a graph-theoretic model. The basic directed graph theory is considered where the nodes represent the objects and the directed edge refers to the web page link or reference. If the object is similar to itself the similarity score to it is given as one. In the graph, there may be nodes with singleton similarity or zero similarity which can be excluded. Similarity can be extended to any two types of objects. SimRank equations are applicable to homogeneous objects as well as heterogeneous objects. The SimRank algorithm considers an ordered pair of a graph G called the node pair graph G^2.The pairs in G^2 represent similarity of nodes and if they refer to the same object they are said to be similar. If the example of Web pages is considered, the graph G is given as (V, E) where V is the set of nodes in the domain and E is the relationship between the nodes. Here each node is an object which is homogeneous in nature. The user-item example is considered as a bipartite graph where the nodes in V represent the users-items and the directed edge indicates the purchase between them. The in-degree and out-degree for every node in the set v are considered and the computed similarity is between 0 and 1.To compute the similarity a naive method is used to compute the SimRank score between *a* and *b* for every iteration *k*. The score is 0 if $a \neq b$, else if *a* and *b* are equal it is 1.When the neighbourhood of G^2 node pairs is observed the nodes with nearby neighbourhood have more similarity than the nodes which are far away and have little overlap. Within a given radius *r* the node pairs are considered and the remaining are pruned. To have a clear perception about the computation of similarity scores a Random surfers model [3] is observed where it starts from two directions namely node *a* and node *b*. A random walk is applied backwards towards a meeting point and the number of steps that lead toward the meeting point is called the Expected Meeting Distance (EMD). In a directed graph which forms a cycle the EMD between the random surfers is always ∞ and this is called a lock step. The EMD is said to be 1 if the random surfers meet. Another new technique to compute the structural similarity is proposed by the authors in [4] called the P-Rank algorithm where P stands for *Penetrating*. The P-Rank algorithm states that two objects in an information network are similar if they exhibit relationship between similar entities. The similarity score computed is similar to SimRank except that it is based on a damping factor C and it depends on the factor for balancing the in-link and the out-link which is denoted as λ. If the P-Rank score computed between two nodes *a* and *b* is the same as that from *b* to *a* in a given iteration then it exhibits similarity. If the similarity score computed for an iteration k and its next iteration k+1 lies between 0 and 1 it exhibits monotonicity. P-Rank similarity score has the property of existence in which the similarity score unite at a fixed point. The P-Rank score displays the property of uniqueness if the damping factor $C \neq 1$.The other forms of P-Rank are *Co-citation, Coupling and Amsler.* In the Co-citation measure [5] on iteration 1, the damping factor C=1 and λ =1.In the Coupling [6] technique which is an another form of P-Rank on iteration 1, damping factor is 1 and λ =0. Amsler [7] is a one step form of P-Rank which works with iteration k=1, C=1, λ =1/2.As the number of iterations increase and becomes ∞ using λ =1 P-Rank exhibits the property of SimRank. As the number of iterations increase and becomes ∞ using λ =1, the P-Rank works as reverse-SimRank also known as rvs-SimRank [4] which is a more practical measure than the SimRank. The computation of P-Rank is preceded by computing the similarity score for two nodes at each iteration and it is stored in a data structure like sparse matrix and hash tables. The time complexity O (n^2) and space complexity O (n^4) of P-Rank and SimRank are the same. The time and space complexities can be reduced in homogeneous networks using radius based pruning and in heterogeneous networks category based pruning.

To make the similarity search faster and to identify the similarity between the peer objects in an information network the authors in [8] proposed an algorithm called as PathSim. PathSim measures the similarity between the objects based on semantic similarity of the objects. The other similarity measures are applicable to densely visible objects. PathSim proposes a framework based on meta-path based similarity. To compute the similarity measures we need to be compute Path count, random walk, and pair-wise random walk. Path count is the number of occurrences between node *x* and node *y*. Random walk is a probability of traversing a path from node x to node y. Pair-wise Random Walk is the probability of conjunction between two path occurrences. A commuting matrix needs to be formed for a network for which a network schema is defined. Figure 4 shows a meta-path symptoms-disease-patient and Figure 3a indicates the adjacency matrix for Symptom-Disease and Figure 3b depicts an adjacency matrix for Disease-Treatment. The commuting matrix is the product of the adjacency matrices that are formed for the meta-path. This matrix is the adjacency matrix formed between two different types of object sets.

With the PathSim similarity measure a top-k similarity measure is proposed where the objects in a given set are of the same type and are sorted. The similarity measure is computed based on symmetric meta-path based calculation. The efficiency of this measure is increased by applying co-clustering pruning technique. To apply this pruning the set of query dependent objects called target clusters and the set of objects for which the features to compute the similarity measure is called as the feature cluster.

The target cluster is partitioned into co-clusters (objects with similar features) by partitioning and these are observed to have any likely objects, if the co-clusters don’t have any likely objects they are pruned which is shown in the Figure 5. The PathSim uses a pair wise random walk and is called as a two random walker model. Another approach SimFusion [9] was proposed to integrate different relationships of objects in a heterogeneous network. SimFusion is based on the Unified Relationship Matrix which is an adjacency matrix formed between two different set of objects which are said to have inter-object relationship. The essence of SimFusion is similar to that of SimRank but with less time complexity. The assumption considered in the SimFusion algorithm is called similarity reinforcement which means the similarity between two data objects is supported by the same type of data objects which are related and also by the different types of relationships between the objects.

	Dis-	Dis-	Dis-		Dis-
Symptom1				....	1
Symptom2	1	1		...	
Symptom3		1	1	....	
....					
Symp-tomN					

Figure 3a: adjacency matrix symptom-disease

	Treatment1	Treatment2	TreatmentN
Disease 1	1	1	
....	1		1
Disease N			

Figure 3b: adjacency matrix for disease-treatment.

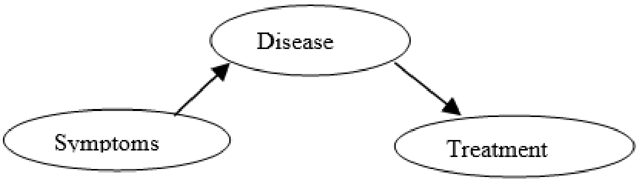

Figure 4a: meta-pathsymptom-disease-treatment

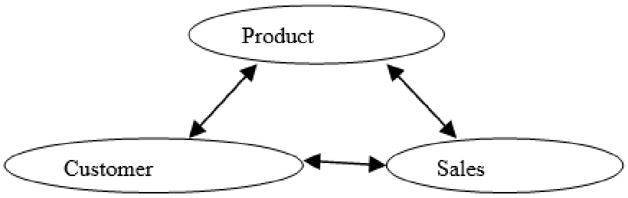

Figure 4b: Meta-path Product-sales-customer

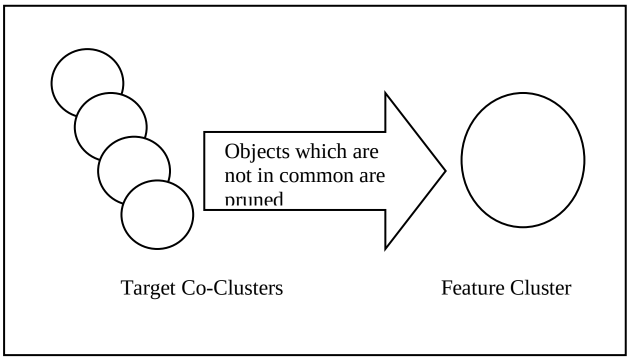

Figure 5: extracting feature cluster from target co-clusters.

The target cluster is partitioned into co-clusters (objects with similar features) by partitioning and these are observed to have any likely objects, if the co-clusters don't have any likely objects they are pruned which is shown in the Figure 5. The PathSim uses a pair wise random walk and is called as a two random walker model. Another approach SimFusion [9] was proposed to integrate different relationships of objects in a heterogeneous network. SimFusion is based on the Unified Relationship Matrix which is an adjacency matrix formed between two different set of objects which are said to have inter-object relationship. The essence of SimFusion is similar to that of SimRank but with less time complexity. The assumption considered in the SimFusion algorithm is called similarity reinforcement which means the similarity between two data objects is supported by the same type of data objects which are related and also by the different types of relationships between the objects.

The novel framework HeteSim is proposed in [10] to identify the object relevance using three metrics called the uniform measure, path constrained measure and a semi-metric measure. HeteSim algorithm is a framework to identify the similarity measure in heterogeneous objects. HeteSim is a path based relevance measure. HeteSim has the following properties: A relation can be decomposed to unique relationships. The significance of relationship between objects in two different sets is based on a relevance path. HeteSim generally exhibits symmetric property not only for paths which are symmetric but also for asymmetric paths. HeteSim exhibits the path maximum property for two objects whose value is 1 if they have similar structure based on the given path. HeteSim is related to SimRank. The only difference between SimRank and HeteSim is that SimRank adds the meeting probability of the objects after all possible steps but HeteSim just computes the meeting probability in the relevance path. In [11] the paper authors proposed an approach for mining meta-paths in heterogeneous information networks using the HeteSim algorithm for the given data to identify the Meta paths then the MapReduce [12] is applied to form the clusters and to get the summarized result. The algorithm proceeds by constructing a constraint based matrix to which HeteSim algorithm can be applied and the MapReduce is applied to the resultant clusters.

HN-similarity is proposed in [13] which is cost-effective and considers the heterogeneous neighbourhood of objects which is a network formed by the structural similarity between nodes. The similarity between objects is computed by an influence dependent function named as HN-Sim. The similarity between two nodes v1 and v2 is a function of weighted factor λ and a product of homogeneous similarity and heterogeneous similarity computed for v1, v2.The similarity function acts in three modes when λ=0 the similarity function computes only heterogeneous neighbourhood. If λ=1 the HN-Sim works just like a homogeneous similarity function and if λ value lies between 1 and 0 it exhibits both homogeneous similarity and heterogeneous similarity.
The Table 1 shows the similarity measures and the models or approaches used by the similarity measures. Also the usage for each of the measures is discussed in the table.

Table 1: Similarity measures and their usage

Name	Model	Time and Space Complexity	Usage
SimRank	Uses random surfers model which is a graph theoretic model based on expected meeting distance. Considers only In-links in the graph	$O(n^3)$-Time $O(n^2)$-Space	1.Used in web document ranking especially in the search engine. 2.Used for document corpora clustering
Prank	Graph based model which uses iterative computing and converges to a fixed point. Considers both the in-link and out-link of the node in a graph.	$O(k(d1^2+d2^2)n^2)$ Average Time $O(n^4)$-worst case Time $O(n^2)$-Space	It is a unified framework for structured similarity computation.
PathSim	A commuting Matrix is used.Meta-path is based on sequence of relations of different objects.	O(n*d)-Time O(n)-Space	For online concatenation and combination of a path to give a top k result for a query.
HeteSim	Uses a transition probability matrix.	O(l*d*n*2)-time O(n2)-Space	1.Automatic object profiling. 2.Expert finding

3. Ranking

Ranking is used to compute the importance of an object. Ranking techniques can be classified into two types: homogeneous ranking techniques and heterogeneous ranking techniques. The popularity of an object is computed based on the ranking functions. HITS-Hyperlink Induced Topic Search [14] identifies the pages that are authoritative for a given query which contain information relevant to the given search query. The HITS classifies second category of pages called hubs contains link towards the authoritative pages. The pages are assigned two kinds of weights authority weight ai and hub weight hi. It is a ranking algorithm that analyses the link among the web pages and rates them. This algorithm is a kind of link prediction by observing the sub-graph which is formed from a base set of web pages that are linked from it and to the pages that link to it. It has structural link among mostly homogeneous kind of entities. It is a ranking algorithm for web pages. Another ranking algorithm to rank the websites was proposed in [15] known as PageRank and is used in Google Search. PageRank is also a link analysis algorithm among the entities which are hyperlinked set of documents. The PageRank is also an algorithm for homogeneous objects. Page Rank assigns global ranking to the web pages. Similarly one more link structure algorithm was proposed by R Lempel and S Moran Stochastic approach for Link Structure Analysis, SALSA [16], which combines the features of PageRank and HITS. In SALSA a result set R is considered from which a neighbourhood graph is considered. For each vertex in the neighbourhood graph an authority and a hub score is computed. A markov chain model is used to perform a random walk in which the sub-graph will visit the authorities which have high probability. To improve the context based searching the topic-sensitive PageRank was proposed in [17].The topic-sensitive PageRank algorithm is a combination of both the page rank and HITS algorithms. This algorithm has two steps: PageRank score vector is captured for each predefined topic statically. The probability that a query of every topic is resolved dynamically. The final ranking is the weighted blend of the rankings for every topic.

Personalized PageRank [18] aims at computing biased PageRank score to a customized question vector q, which is referred to as preference vector. The preference vector is unique in relation to topic sensitive page rank where the query vectors have predefined topics which are fixed. The query vectors are subjective in Personalized PageRank. The Personalized PageRank starts a random walk from source s to target t and a teleport probability of α these steps are carried out iteratively and either the process stops with a probability α else will continue and it is given as πs(t) whose value is calculated and named as P.P is nothing but the walk from source s to target t until it reaches a threshold estimate δ. TrustRank [19] is a kind of Link analysis technique which separates the useful web pages from spam. Many web spam pages are made just with the goal of deceiving web indexes. These pages, mainly made for business reasons, utilize different procedures to accomplish fake rankings on the web search tools' outcome pages. While human specialists can without much of a stretch distinguish spam, a manual survey of the Internet is not feasible. One famous technique for enhancing rankings is to build the apparent significance of an archive through complex connecting plans. Google's PageRank and other pursuit positioning calculations have been subjected to such control.
TrustRank looks to battle spam by filtering the web based on unwavering quality. The strategy calls for choosing a little arrangement of seed pages to be assessed by a specialist. Once the legitimate seed pages are physically identified, a slighter augmenting outward from the seed set searches out comparably solid and dependable pages.

Here we discuss about the heterogeneous ranking algorithms. Sorting the objects from a given order from a database is called as Object Ranking [20]. Object Ranking is very much useful for surveys, retrieval of the information and decision making. The Object Ranking method is Cohen's method which is a greedy algorithm that sequentially chooses the most-preceding object. Other Ranking methods are for Object ranking are the Rank Boost algorithm [21], Support Vector Machine based algorithms like Order SVM [22] and Herbrich's method [23]. After one of the application of above methods to the objects a regression method called Expected Rank Regression is applied to obtain the expected ranks of the objects. Pop-Rank [24] is an extension to the Page Rank model with the addition of a Popularity Propagation Factor for each link that points to an object which uses different types of propagation factors to each structural link with different types of relationships. Pop Rank model is used in a paper search engine called as Libra. A random object finder model which keeps clicking on successive Web page links, Web page to object links, and object relationship links at random. The search strategy Pop Rank is based on Simulated Annealing for Factor Assignment, the algorithm adapts the simulated annealing algorithm to automatically assign popularity propagation factors. The neighbours are observed to get a best combination of Popularity Propagation factor. To make the search strategy better a sub-graph with k-diameter is chosen which forms k concentric circles and all these circles contain the objects that are less than k-links away from the main object inclusive of all the links from these objects. The distance between the ranking results and the ranking given by domain experts is the final computed cost. PopRank gives a global score for every object. The Co-ranking method [25] is a ranking framework on heterogeneous networks by using two random walks on GA a Social Network of Authors, GD a network containing connecting documents. It is a method based on Page Rank and the mutual reinforcement principle. Co Rank has three drawbacks .Firstly it does not consider the venue information. Secondly it is good for an older paper with good citation. Thirdly the Self Citation is given equal priority to the other citations. The Co Rank considers the un-weighted undirected graph which is used to indicate the set of heterogeneous objects. The bidirectional edges between them are the relationships between the objects. For example in an author document graph If the given graph is subdivided into three sub-graphs like one sub-graph ssconsists of the other sub-graph the nodes that represent authors and the edges that represent the social ties between them. That contains nodes which represent the set of documents and the links between them .The last sub-graph consists of the nodes which contain the set of authors and documents and the edges contain the relationship between these nodes. On the three sub-graphs identified, the random walk is applied for each path. The random walk on the graph is the Markov chain which is represented by a stochastic matrix which contains non-negative entities and the sum of each row is 1. The co-ranking is done by coupling two intra-class random walks. However the Co-rank algorithm has some drawbacks such as lack of venue information, the citations of previous papers are more and they are not given more weightage ,the priority for self citation should be less when considered with other papers. To prevail over the above said drawbacks a heterogeneous ranking framework has been proposed by the Tri-Rank Algorithm [26] three bipartite graphs are considered. Here they have considered a heterogeneous connectivity of objects which are all papers, authors and venues. The intra-relationships within the graph are denoted using a bidirectional edge the inter-relationships between the bipartite graphs is called uni-directional relationships. Tri-Rank can be observed as the extension of co-rank and it follows certain rules. The first rule specifies that the score of the current paper is mostly affected by the score of the papers cited. The second one emphasizes on the rank of the author and the venue. The third depend on the score of the venue and on the rank of the author. The score of the author is also based upon the rank of the co-authors and also on those who cite the author. The fourth rule states that score is also influenced by the scores of the paper published and also depends on the venues attended by the author. The fifth rule is inclined by the scores of the other venues that cite this venue. Lastly the score of the paper depends on the average score of the papers at the venues and the average of the score of authors who attended it. The Tri-Rank framework proceeds in four steps .On the first iteration the PageRank is applied on the graph of papers. Now as the scores of the papers are restructured, using these scores of the authors and venues are reorganized and the scores of the author network are normalized by applying one iteration PageRank. As a next step the venue's score is updated and normalized with one iteration PageRank. The scores of the papers are updated and normalized and this procedure iterates. Table 2 shows the ranking algorithms their time complexities and applications.

Table 2: Ranking techniques their time complexities and applications.

Algorithm Name	Model Used	Time and Space Complexity	Usage
SALSA	Stochastic approach with less computing between hubs and authorities by isolating Tightly Knit community	Time-O(k*E) Space – O(k*E)	Query search
Personalized Page Rank	Make use of Scalable Hyper-link score .The graph is stored in a social store. All jumps are made to the seed node.	Θ(n)	Personalized recommendations
Trust Rank	Seed set is computed using inverted page Rank. Trust Rank scores are calculated by trust damping and trust splitting.	O(p)	For the separation of spam pages from useful pages.
Object Rank	Object Rank is computed using Index based table. The authority transfer data graph is DAG on which object rank is computed.	O(n^2)- Cohens Approach O(n log n) – Rank-Boost,Order SVM	To rank Objects
TriRank	A tripartite graph is used where each edge carries a weight and a global minimum function is applied	O(L_u)	Used for recommender purpose

4. Clustering

Clustering forms groups of objects which are similar according to a closeness rule that objects within the same cluster are more similar to each other and objects of two different clusters are less similar to each other. Clustering can be done on objects which are homogeneous in nature and also can be extended to objects which are heterogeneous in nature. The basic data-mining techniques like the K-Means, K-medians are good examples for homogeneous clustering. Clustering can also be done on heterogeneous objects. Clustering on heterogeneous objects is quite different and complex when compared with homogeneous techniques. Clustering on heterogeneous web objects is performed using compression technique by considering a clustering structure C {c1, c2,.., cm} of a data set 'x' and the defined partition set for the data P as {p1,p2,…,ps}. A pair of points are defined on these data sets using the points from SS which is a cluster which belong to the same cluster and the same partition which is indicated as 'a'. A pair of points are said to be from a set of points named as SD if the cluster belong to the equivalent structure. The dissimilar groups of partitions are indicated as 'b', the pair of points from the set DS if the points belong to different cluster C but of the same group given as 'c', a pair of points are said to be from DD if they are from different cluster and belong to diverse group and named as'd'. The sum of the four points is said to be M. Using the

four points a, b, c, d and the value M indices are defined as follows: The rank index is defined as (a+d)/M this value should lie between 1 and 0. If the value is close to 1 then C and P have high similitude. The Jaccard coefficient is calculated whose value lies between 0 and 1 which is used to estimate the similitude between C and P. Similarly the Fowles –Mallow measure is calculated whose value close to 1 indicates the similitude between C and P. In order to classify the network as heterogeneous in nature the objects must be at least of two different categories which are said to be bi-typed. The RankClus is proposed in [27] which perform clustering and then ranking on a given set of data objects so that the objects after clustering can be ranked within a cluster and this could give a better analysis about that cluster. The RankClus algorithm runs on a bi typed information network. The ranking functions considered to rank the objects in an information network are simple ranking and authority ranking. The example heterogeneous information network considered here is author-conference network in which authors who publish more papers are ranked high in-spite of the value of the conferences. The next approach is the authority ranking which proceeds with certain rules that the set of ranks that are highly ranked will publish in highly ranked conference. The second rule is that the conferences which are very popular draw the papers from authors which are highly ranked. The next rule tells that if an author may be co-authored with highly ranked authors then his popularity increases. The algorithm proceeds with assigning a conditional rank to each object depends on that cluster. The conditional ranks within the clusters are not similar to each other. A conditional rank can be used as an important attribute to observe the cluster. The mixture model components are calculated using the EM algorithm [28] and the cluster is adjusted. The drawbacks of the RankClus algorithm is that it has less number of object types and if we observe the bi-typed networks the clusters formed have only homogeneous objects. To overcome these drawbacks the NetClus [29] algorithm is proposed by Sun et al with a star schema. In the initial step of the algorithm net clusters are formed. The probabilistic generative model is used where the probability of an object is calculated based on the probability of visiting that object in the entire heterogeneous networks and the probability of that object corresponding to all other objects. The probability of two different objects may not be the same and is independent of each other. If the target object belongs to a different cluster the posterior probabilities are calculated. Once the clusters are formed the ranking techniques adopted to rank the cluster are simple ranking and authority ranking.

Table 3: Clustering techniques their time complexities and applications

Algorithm Name	Model Used	Time and Space Complexity	Usage
RankClus	A mixture model is used and objects are adjusted to the clusters. Conditional rank for each cluster is calculated.	$O(t(t_1\|E\|+t_2(K\|E\|+K+mK)+mK^2))$	Ranking databases
NetClus	Ranking on a star schema network using authority ranking .	$O(k.\eta_p.\xi_p)$	Trajectory aware services, Facility Location queries
MedRank[30]	Ranks Heterogeneous Objects in medical information. Authority ranking formula is used.	$O(l\|E'\|)$	For Ranking heterogeneous objects in medical databases.

5. Classification

Classification is a way of specifying categorical groups and giving them labels. The basic classification techniques like the decision tree induction can be quoted a good example for classification of data objects which are homogeneous in nature. Classification can also be extended to data objects which are heterogeneous which are discussed in this section. ComClus [31] is defined on hybrid networks which include the homogeneous as well as the heterogeneous objects. This approach is a graph based classification and is applicable for heterogeneous networks. The nodes are of distinct types and are heterogeneous in nature they belong to various classification labels. The NetClus is proposed with a star schema and the ComClus is a star schema with a self loop. ComClus uses a probability model to represent a generative probability. The generative probability model and the expert’s model are used to combine the relationships from heterogeneous and homogeneous relations. The algorithm proceeds by randomly partitioning the given network G. For each subnet in the graph the homogeneous probability, conditional probability, the heterogeneous probability and the mixed probability are calculated on centre node. For every centre node the posterior probabilities are calculated and these are used for ranking the nodes. The Graffiti [32] works as follows: Firstly the contents of the nodes which indicate the description is identified. If there is an edge among the same type of node it is called as the S-edge. If there exits an edge among different type of nodes it is called the X-edge. A random walk procedure is applied on the graph to provide a convergence solution and also it gives a unique solution. The algorithm proceeds with taking initial graph nodes with labels as input, then it identifies the nodes which have similar neighbours by identifying the reciprocal influence of each node and it must be higher to form the S-edge. The RankClass [33] algorithm gives the ranking of an object among a class of objects. The RankClass proceeds with the following steps: 1) A class of labelled data is considered and it initializes the ranking distribution function. The type of network structure considered is also identified and is initialized. 2) In each class the network structures are adjusted. This resembles the similarity of boosting technique and the ranking distribution is done based on the graph based ranking model.3) Steps 1 and 2 are iteratively repeated and the posterior probability is calculated until the object is assigned a class label. It is a transductive classification [34] strategy. This is a new graph based classification approach in identifying the link structures in heterogeneous networks where the information is processed from labelled data to unlabelled data. The confidence of two objects of an estimated class label k should be same.

6. Recommender Systems

Recommender systems utilize the knowledge discovery techniques and statistical methods to predict recommendations and have its prominence in e-commerce applications. The recommendation algorithms are designed based on collaborative filtering model or cluster model. The other models are search based models which identifies similar items. The working of the collaborative filtering uses a database where the customer is matched with the entities or other customers who have similar interests in purchasing items and the other products which are purchased by the neighbours are recommended to the customer. The two main challenges of the recommender systems are 1) scalability which is a great factor that has influence on its neighbours for high dimensional data when trying to apply recommender algorithms for huge, dynamic and realistic datasets2) the veracity derived from the recommender systems should not be false. This factor also affects the quality of the recommender systems. The most prominent data mining technique to apply for the recommender systems is the association rule mining technique which applies the support and the confidence thresholds to suggest the entity recommendations in a given transaction. Collaborative filtering is the most effective way for suggesting the recommendations among the neighbours. The recommendation for any customer depends on the opinions of the group of other customers. In [35] collaborative filtering is designed in three subtasks 1)representation 2)formation of the neighbourhood 3)recommendation generation. In the representation step a cross product matrix is used with the number of row a is based on products and the number of columns b represent the transactions. The corresponding value in the matrix is one if the user purchases that product otherwise zero. The second step which includes forming of similar neighbours using the measure of similarity. The similarity is measured using the correlation technique or the cosine similarity. The final step is to generate recommendation using most-frequent item based recommendation where the count of most frequent items purchased are calculated ,sorted and the most frequent products are given for recommendation or the next approach for generating recommendation is Association rule based recommendation where only few neighbours are considered for generating association rules. As a limited set is considered for recommendations there may be a drawback of inadequate products for recommendation and the reduction in dimensions may not be considered as a best approach for recommendations. The most primitive application of the collaborative filtering is used in the Tapestry which is an experimental mail system.

The basic algorithms used in collaborative filtering are K-NearestNeighbor or the K-means algorithm which may not be useful for dense data and may not be apt for recommendations. The other approach for recommender systems is cluster model which suggests the recommendations treating it as a classification problem. The method implemented to form clusters is called as *repeated clustering*. One method is to cluster a set of entities based on the relationships with the set of second category of entities. Later in the second step the set of second category of entities are clustered using K-Means using the influence of the first category of entities. The clustered formed are re-clustered based on the number of first category of entities influencing the second category and vice-versa. A *“soft clustering”* method is also used which is based on K-Means clustering method in which a set of objects are assigned to a class based on the degree which is equivalent to the similarity measure calculated.

The other category of recommendation system algorithms include search based models which identify the solution for proposing a recommendation as a search for interconnected items. But this search based model is not so effective for recommendation systems. The most popularly used recommendation system for Amazon.com uses item-to-item collaborative filtering approach. In this approach the items purchased by the user are matched with the rated items and these two lists are observed to combine and form a recommendation list. A similar items table is built by observing the most affinity of the customer to purchase an item. A product-to –product matrix is also built but the inefficiency is that it does not have similar customers. By using an iterative algorithm similarity is computed between two items. In [36] Xiao Yu, Jiawei Han et al proposed an entity recommendation approach based on meta-path where user feed-backs are diffused across the meta-path. The diffusion scores corresponding to each user versus item is called as the diffused user preference matrix. If we traverse across the matrix we can identify the set of users and their preferences which also give the diffusion scores to propose the recommendations considering and applying the facts of the global model. Not only global recommendations but also personalized recommendations can be built using user implicit feedback. The users are clustered using K means algorithm with cosine similarity based on their preferences also matrix factorization techniques are used when clusters are formed. On these clusters formed the Bayesian ranking based optimization or stochastic gradient descent [37] method is used to learn the personalized entity recommendation. The figure 6 shows how recommender systems use the meta-paths and the diffusion matrix to suggest the recommendations.

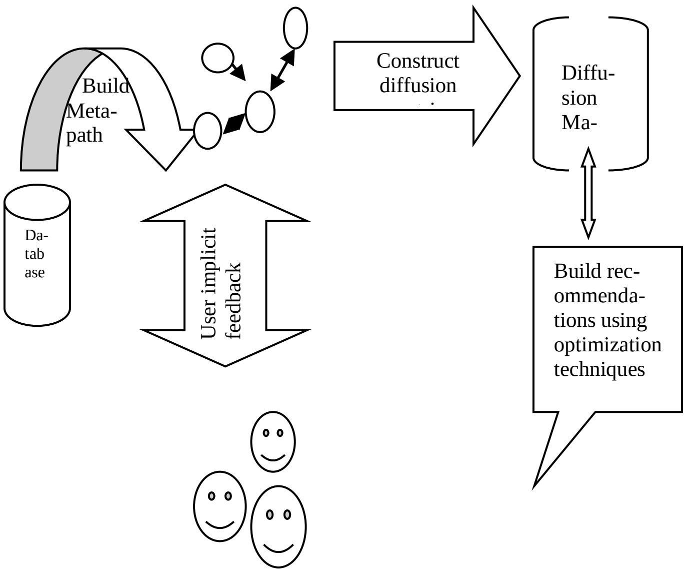

Figure 6: Recommender Systems based on Meta-path.

7. Conclusion

In this paper the authors discussed various Data Mining techniques for addressing information mining of homogeneous as well as heterogeneous information networks. Mining techniques based on similarity, ranking, clustering, classification are discussed in detail. The prominence of the recommender systems is increasing in the present day purchases, suggestions for product recommendations, venue recommendation, hotels and restaurant recommendations etc. The recommender systems are also discussed for both homogeneous and heterogeneous networks. Application and usage of the techniques are also tabulated. There is a lot of scope for research in any of the above discussed techniques by identifying the semantics and the structural behaviour of the objects interacting with each other. The research areas like image retrieval based on texture is a very challenging area of research where we need the information networks to identify the texture, another area of research based on similarity includes identification of patterns in continuous data which is an information network based on text documents. Shape based similarity and classification is also an interesting area of research. The readers can visualize various areas of research as different perspectives of usage areas are presented here.

References

[1] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, Philip S. Yu,” A Survey of Heterogeneous Information Network Analysis”, in Journal Latex class Files, Vol 14,no 8,August 2017.

[2]Glen Jeh, Jennifer Widom,”SimRank: A measure of Structural-Context similarity” in KDD, pp. 538–543, 2002.

[3] A. Blum, T.-H. H. Chan, and M. R. Rwebangira. A random-surfer web-graph model. In ANALCO ’06: Proceedings of the eighth Workshop on Algorithm Engineering and Experiments and the third Workshop on Analytic Algorithmic and Combinatorics, pages 238- 246, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics.

[4] Peixiang Zhao, Jiawei Han, Yizhou Sun,” P-Rank: a Comprehensive Structural Similarity Measure over Information Networks”.CIKM’09, Hong Kong, China. November 2–6, 2009

[5] H. G. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, Science, 24(4):265-269, 1973.

[6] M. M. Kessler. Bibliographic coupling between scientific papers. American Documentation, 14:10-25, 1963.

[7] R. Amsler. Application of citation-based automatic classification. Technical report, The University of Texas at Austin Linguistics Research Center, December 1972.

[8] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, Tianyi Wu, “PathSim: Meta path Based TopK Similarity Search in Heterogeneous Information Networks” *Proceedings of the VLDB Endowment,* Vol. 4, No. 11 2011.

[9] Wensi Xi, Benyu Zhang, Edward A. Fox, SimFusion: A Unified Similarity Measurement Algorithm for Multi-Type Interrelated Web Objects, in the www conference May10-14, 2005.

[10] C. Shi, X. Kong, Y. Huang, S. Y. Philip, and B. Wu, “HeteSim:A general framework for relevance measure in heterogeneous networks,” IEEE Transactions on Knowledge & Data Engineering, vol. 26, no. 10, pp. 2479–2492, 2014.

[11] Sadhana Kodali,Madhavi Dabbiru,Kamalakar Meduri,”Constraint based approach for minging Heterogeneous Information Networks” 6th IEEE IACC 2016 ,27th -28th February 2016.

[12] Jeffrey Dean and Sanjay Ghemawat ,” MapReduce: Simplified Data Processing on Large Clusters” ,OSDI 2004 11th March.

[13] Jiazhen Nian, Shanshan Wang, and Yan Zhang,”HN-Sim: A Structural Similarity Measure over Object-Behavior Networks”, Part I, LNAI 8346, pp. 48–59 ,ADMA-2013.

[14]J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” in SODA, pp. 668–677. 1999.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, “The Page Rank citation ranking: Bringing order to the web.” Technical report, Stanford University Database Group, 1998.

[16] R. LEMPEL and S. MORAN, SALSA: The Stochastic Approach for Link Structure Analysis, ACM Transactions on Information Systems, Vol. 19, No. 2, April 2001.

[17] Taher H. Haveliwala, Topic Sensitive Page Rank WWW May 7–11, Honolulu, Hawaii, USA, 2002.

[18] G.Jeh and J. Widom, “Scaling personalized web search,” in WWW, pp 271–279, 2003.

[19] Gyongyi Z, Garcia-Molina H, Pedersen J Combating web spam with TrustRank. In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, VLDB Endowment, VLDB '04, pp 576-587, 2004.

[20] Balmin A, Hristidis V, Papakonstantinou ObjectRank: authority-based keyword search in databases. In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, VLDB Endowment, VLDB, pp 564-575, 2004.

[21] Yoav Freund, Raj Iyer, Robert E. Schapire, Yoram Singer, An Efficient Boosting Algorithm for Combining Preferences, Journal of Machine Learning Research 4 pp.933-969,2003.

[22] Kazawa, H., Hirao, T.Maeda,: Order SVM: a kernel method for order learning based on generalized order statistics. Systems and Computers in Japan pp 35–43, 2005.

[23] Herbrich, R., Graepel, T., Bollmann-Sdorra, P Obermayer, K.: Learning preference relations for information retrieval. In: ICML-98 Workshop: Text Categorization and Machine Learning.pp 80–84, 1998.

[24] Nie Z, Zhang Y, Wen JR, Ma WY (2005) Object-level ranking: bringing order to web objects. In: Proceedings of the 14th international conference on World Wide Web, WWW '05, pp 567-574,2005.

[25] Hai-jiang He, A Co-Ranking Algorithm for Learning Listwise Ranking Functions from Unlabeled Data, journal of computers, vol. 6, no. 11, november 2011.

[26]Zhirun Liu,Heyan Huang,Xiaochi Wei,Xianling Mao, Tri-Rank: An Authority Ranking Framework in Heterogeneous Academic Networks by Mutual Reinforce,26 th IEEE International Conference on tools with Artificial Intelligence,2014.

[27] 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,pp. 565–576, 2009

[28] A. P. Dempster; N. M. Laird; D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. , pp. 1-38, 1977.

[29] Y. Sun, Y. Yu, and J. Han, “Ranking-based clustering of heterogeneous information networks with star network schema,” in KDD, pp. 797–806, 2009.

[30] Ling Chen,XueLi,Jiawei Han “MedRank: Discovering Influential Medical Treatments from Literature by Information Network Analysis” Twenty-Fourth Australasian Database Conference (ADC2013), Adelaide, Australia,2013.

[31]R. Wang, C. Shi, P. S. Yu, and B. Wu, “Integrating clustering and ranking on hybrid heterogeneous information network,” in PAKDD, pp. 583–594, 2013.

[32] R. Angelova, G. Kasneci, and G. Weikum, “Graffiti: Graph-based classification in heterogeneous networks,” in WWW, pp.139–170, 2012.

[33] M. Ji, J. Han, and M. Danilevsky, “Ranking-based classification of heterogeneous information networks,” in KDD pp. 1298–1306, 2011.

[34] C. Luo, R. Guan, Z. Wang, and C. Lin, “HetPathMine: A novel transductive classification algorithm on heterogeneous information networks,” Advances in Information Retrieval, vol. 8416, pp. 210–221, 2014.

[35]L. Ungar and D. Foster, “Clustering Methods for Collaborative Filtering,” Proc. Workshop on Recommendation Systems, AAAI Press, 1998.

[36] Xiao Yu, Jiawei Han et al."Personalized Entity Recommendation: A Heterogeneous Information Network Approach"WSDM’14, , New York, New York, USA, February 24–28, 2014.

[37] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. "A limited memory algorithm for bound constrained optimization." SIAM Journal on Scientific Computing, 16(5):1190–1208, 1995.