Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 08
Lecture - 39

Lecture 39: Structural Recommendations in Networks

This is last but one lecture. Welcome everyone. So we are on the 39th lecture and here we are going to talk about structural recommendations in a network. It is a completely different kind of system in the sense here we consider the social relationship as an additional input to our recommendation system. So these are the examples of network based recommendation. Network has become very important part of our life.

We participate in various networks. We participate in Facebook, connect to our friends. We have a professional network. We work using computers and which are interconnected. Of course, web is another example and so on and so forth. Now what a network is? A network consists of a number of nodes and edges and it is also called a graph and it is a typically it is a very large graph. In case we are talking about a social network, it is a very large graph where the nodes will be connected to each other. There can be direct connected, they can be directed or they can be undirected. In case undirected means both way movement is possible, both way connectivity is possible.

Now what kind of recommendation we can give in a structure like this? There are basically four situations here. In fact, two broad situations we will be recommending nodes or recommending link. So, when we recommend these nodes, there are three purposes for which we will be recommending nodes. We will be recommending nodes by authority and context. Think of a situation in which you have to know some reputed person.

So, why do you need a reputed person to know or some reputed authentic source of information? How do you know? Typically, in search engine like that of Google and all the modern search engines, they use something called a PageRank algorithm to find out that authentic source of information where many other web pages try connecting. So same concept you can bring into social network and do something called a personalized PageRank, find out something called a personalized PageRank. So why do we do it? So if we find out the reputed node in a social network, what do we do? Suppose a reputed node which you respect tells something; you will believe him. So this can be used as a marketing tool. Second is you can identify nodes by example.

So how do we find out the nodes by example? Usually similar nodes will be doing similar activities. So here there is a concept called homophily. So it says that nodes which

are similar in their nature in a social network will be connected together. So what do we do this? So if I like something, it is very likely that I am friend with somebody and that somebody will also like that. So identifying such similar individuals, we can for a particular product we can do some kind of target marketing.

We can also recommend the nodes via influence and their content. So what happens here is you can find out the nodes who are very quick in disseminating the information. So they can influence others very fast. So those social influencers can now be used for viral marketing. Then comes recommending the links.

One in social network sites you are asked, you are given the suggestions for friends and connections, professional connections maybe. Why do they do that? Because they want to make their network dense. So if they do the network dense, what is the benefit? First of all the benefit comes to you, you get connected to a like-minded individual. And other benefits include as the network becomes dense, propagating information also becomes easy and it helps the company. So first we are going to talk about recommending nodes by authority and context.

Think of a situation like this. A health drinks producer wishes to find a brand ambassador for a newly introduced item. An appropriate person would be whom everyone respects in this domain. So how do we find out such a person in social network? Find out where everybody is trying to connect? So if everybody is referring that person, then he is definitely an important person. And if you make him brand ambassador and he promotes your item, everybody may, everybody is likely to believe that person.

So basis of this is actually Page Rank algorithm which has to be personalized in this context. Because you will be, when you find out certain important person to make you a brand ambassador, it is with respect to a particular item, not with respect to everything. So that to bring that specificity we use personalized Page Rank. We won't be discussing in general details about everything because in this last module we are simply introducing few topics, so we won't be going to very depth. But still foundation of this is your Page Rank algorithm.

$$\bullet \text{ Let } A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix}$$

$$A_{ij} = \begin{cases} \frac{1}{o_i} & if\ (i,j) \in E \\ 0 & Otherwise \end{cases}$$

In a Page Rank algorithm what it does, which was developed in the context of web, it considers each page as a node and the connectivity among them as the links. You can go from this page to this and so on. So in this large graph for a particular page there can be

some in-links which are coming to coming here from various places and there can be some out-links going from this node to other. Now this whole in-link out-link how they are movement is happening, that can be represented in the form of certain how people make transitions. So this situation if you model as a transition probability matrix and the movement we consider as some kind of stochastic movement of people from one page to the other making requests.

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j}$$

$$P = \left(P(1), P(2), P(3) \cdots P(n)\right)^T$$

Many things can be derived from this matrix. And in case this particular structure, this stochastic structure which is represented in terms of state transition probability matrix, if it has certain properties like ergodic, irreducible and aperiodic etcetera which of course we cannot study here in detail. But if these properties are satisfied, then in the long run if we raise it to n where n tends infinity, then what happens every row here becomes same. So there were n rows, so all the n rows will have same value. So basically this thing will represent the probability of staying in a particular state in the long run.

So this long run behavior can be exploited to find out the importance of a particular page where every people is finally converging. However, if this structure is not ergodic, certain arrangements can be made which we are not going to discuss and Page Rank Score can be presented in this form. The problem with this Page Rank is this is independent of any topic. Now if you like to add topic, what do we do? We try making it personalized or make it topic sensitive. So to make it personalized or topic sensitive what we do? We multiply this with another vector 0 1 vector, another 0 1 vector, another 0 1 vector where each entry represents if a topic is if a page is relevant to the topic or not.

So such a vector gets multiplied with the transitional probability matrix. So as a result you get those you get the you involve those pages which are important in a particular context. And among them find out the most prestigious node, most prestigious node and that node becomes your node with prestige. So what was the three situations? So first task was this one, authority and context. So this is how we discover the nodes with authority and context.

Second one was finding the nodes, recommending the nodes by example. So when we say finding the nodes by example, once again let me remind you we are trying to find out the similar nodes. How do we? So, in traditional concept of similarity we are not using in

the sense we may not have the values with respect to each of the attributes that we are considering. And moreover which attribute we should consider is also and also to be thought about. Consider this example again.

You have a manufacturer; you have a manufacturer of golf equipment who wishes to target few nodes for marketing. Now he must select those nodes we are interested in golf. Now this interest can be inferred from their own posts, the likes from the other posts, tagging related items and so on. And these are the people who are explicitly showing their interest. But if you would like to find out the people who are otherwise implicitly also may be interested, who will they be? As per the concept of homophily, they will be the people who are connected to the nodes which are of your interest.

So therefore, the profile, properties and ratings of the neighborhood nodes can be now liberating to recommend such additional nodes. So this process is called recommendation by collective classification. In the context of recommendation by collective classification, the actors of specific interest in the network are specified using certain labels that he is interested. So there can be many labels as well. So these labels indicate he is may be we can say label 1 is for first kind of item, label 2 is for second kind of item and so on.

And there can be let us say some r labels. Now some nodes, node 1 is having label which is explicitly found, node 2 explicitly found and some few other m nodes we have explicitly found. But in the network suppose you have total m, capital N number of nodes. So these number of nodes label we are supposed to predict for these nodes. Now in this prediction problem what will be the regressor variable, what will be the independent variable with respect to n 1, n 2 and so on for which label is known.

Here we make the attributes. These attributes come from two sources. So these attributes come from two sources. So what are those two sources? These two sources are the content-based attribute and link attribute. So you have two kind of attributes here. Content attribute which is derived from the node and you have link attribute, which is derived from the neighbors.

So which means now you have two attributes content and link, content and link. So now you have for every node for which label is known now you have identified the attributes. So now you have a supervised learning, you apply your supervised learning algorithm. But the problem here is again only few nodes will be will have labels. So which means you have to adopt some mechanism to increase the data set size or from this small set somehow determine the values.

So this can be done using many approaches. These two very popular approaches are Iterative Classification Algorithm and Random Work-Based Method. In case of Iterative Classification Algorithm which is a classical one, we are going to look at how to do it.

So we consider a situation in which we have a graph with a number of nodes and links. And we construct some kind of multidimensional feature vector as I discussed.

And total number of unlabeled test nodes are n, t. So out of n unlabeled nodes are n, t rest are labeled. We are supposed to derive a set of link features. How do we get the link feature? In this case a link feature is generated for each class containing the fraction of each class.

How many classes we have? r classes. So in this Iterative Classification Algorithm the link features are those r classes. So you have content feature and you have link feature. In link feature for r classes you have r features for each node. So what these r features are going to take, which value it is they are going to take? Now each node i, for each node i its adjacent node you have to look at and you have to take the weight corresponding to the interest relevant if the interest of that other node is relevant to that particular class. So using certain mechanism you construct the weights.

So that mechanism you can decide or there are certain standard mechanisms which you may follow. So after doing this what do you do? You got the weights, you got the network, then you have the network attributes as well and there you have class labels and so on. You have a classifier. Let us say we have Naive Bayes and so on.

Some classifier you have. This process can be followed by any classifier that is what is meant. So you extract the link features, train the classifier with the current data, predict the level of the test node. How many test nodes were there? Total n, t test nodes were there. So out of this you choose some n, t minus capital T. What is capital T? Capital T is the number of iterations, number of iterations capital T.

You choose that many and again add to the training set. You have with how many items you started with the training set? You had total n number of nodes of which n, t were unknown. So your test training set started with this. So every time to this you add n, t by T number of certain additional records which you predict in each iteration and by the end of your all the iterations every time n by T, n by T, n by T by the end of this iteration all the nodes which were there in this set n, t will be considered.

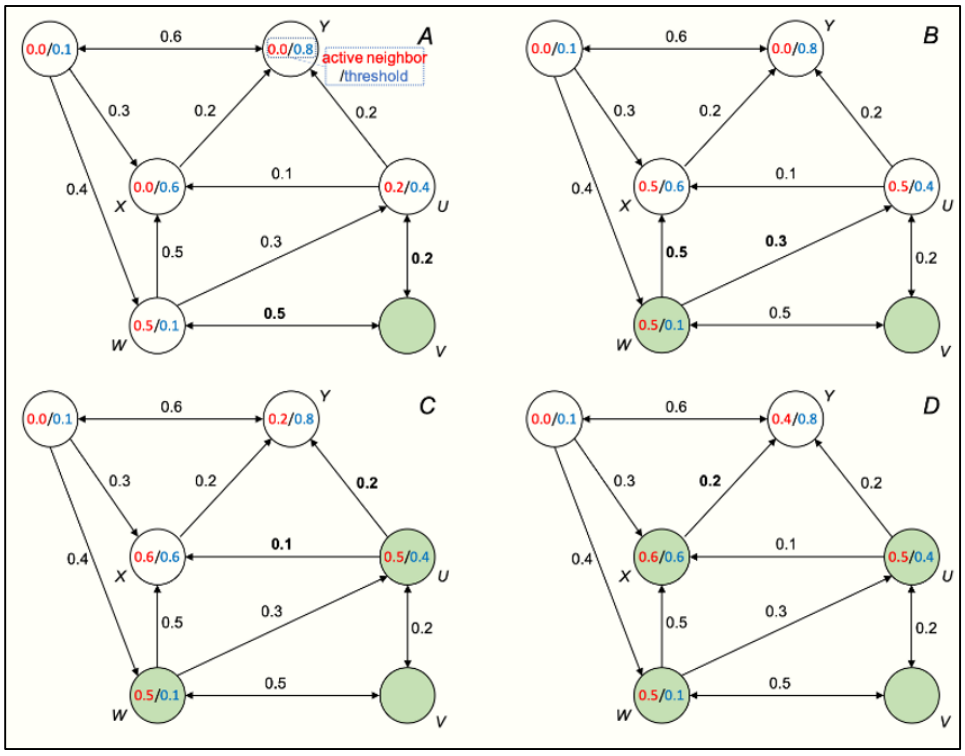
So you continue up to T iterations. Then recommending nodes by influence and content. So now you suppose you want to do viral marketing for your product. So you have to choose few nodes in the network from which you will start propagating. So these nodes will be connected with each other to many other nodes and so on. So you will start check few nodes from which you will start propagating this.

So why do you propagate this? Because you want to spread the information across the network. Now if you spread the information across the network your choice of nodes should be such that they are capable of spreading it. Capable in the sense they have

sufficient influence over their followers. So that whatever they try sharing the other person probably would try sharing that as well or like at least. So now the idea is out of this n you are suppose to choose a subset S.

This subset S you call as your seed set. So this basically is a seed set selection problem. So now while you select this seed set it has to you have to select this using certain function and this function has to follow certain property called submodularity. Now this submodularity property says that if some additional node is added to this set S and same is added to this T, T cannot have larger influence than what S was having. So that set you are supposed to find out. So two common approaches for finding this function in an iterative manner through in a simulation based setting are linear threshold model and independent cascade model.

These are basically for two models for spreading the influence. This is one example of linear threshold model. We are not going to discuss about the other one that is independent cascade model. This is similar but there are certain restrictions that one node cannot influence more than once. So here in this example what is happening? Initially seed set contains this node V.



So this V now belongs to this seed set S. S was initially S was there was nothing in S. So you started with this. So in this network there are many numbers are appearing. Now these numbers indicate how it is going to influence the other. So every node there is a blue thing called see this example is taken from this source.

You can this is a very good source to know more about this influence maximization in social networks. You can study further. So for sake of our understanding here this blue values are the threshold values and red values within this node represent how they are getting activated by their neighbors. And when they will get activated by their neighbors? For example, here V can let us say activate with this value that influence value is 0.2 and it has a threshold of 0.4. So V cannot influence this. But here it can influence and this influence by the way need not always be two ways. It can be one way as well. So if I can influence you, you may not be able to influence me. That is how this social that is what the social influencers are.

$$JaccardPredict(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$$

They actually are able to influence us but we are not able to influence them. So this 0.5 this is here the threshold was this blue one was 0.2 and this was higher than this. So this node got influenced. So now in this this is the situation A situation B this node gets influenced. Now what happens after this node get influenced? This particular node U was having threshold as 4 and this was 2 that is why it was not getting influenced. Now this is giving 2 this is also giving some information which it is the with the value 0.3. So 0.3 plus 0.2 becomes 0.5 which is higher than this 0.4. So this node now is activated is influenced. Now because these three nodes are influenced now what about this node? Now this can be influenced by this person as well as this person threshold is 6 this is 0.5 this is 0.1 so this gets activated. What about this? Here threshold is 0.8 but here it is 0.2 here 0.3 so this cannot be influenced further. So also the case here because here there are only inbound links no only one outbound links so which means here you stop you cannot from this seed set you cannot go further. So how did you add values to seed set? First you added v then you added w then you added u then you added x and beyond this you cannot add things.

So total influence sorry this is not the seed set total influence you made is activating influencing the four nodes and with seed set s v belongs to s you could get this many people influenced. Now suppose along with s you add this node as well are you able to improve the influence are you able to add more number of persons to this set. So you have to now psi taking one node as seed set two combinations of two nodes and so on. So with this now we move to the next topic that is recommending links. Now this recommending links is typically used to show you the possible friends etc the potential friends whom you can connect.

So there are many measures for recommending links. So this while this first three are mostly heuristic type the second two follows some models. So out of this we are simply

for the representative representation purpose the representative will be talking about only neighborhood based measure. Other measures as I told you again I have followed completely from recommender system or text book from Charva Agarwal you can read that to get more insights on other methods as well. There are many common as I told you these are heuristics so they depend on our common sense on this. So the first neighborhood based method is the common neighborhood based method this is the first method.

$$AdamicAdar(i,j) = \sum_{k \in S_i \cap S_j} \frac{1}{log(|S_k|)}$$

So in this what do we do we measure the common neighbors of two nodes i and j. So if s is the set of neighbors of i j is the set of neighbors of j then the intersection is my common neighborhood. So this is a very simplistic measure but the problem here is if there are some I mean if both of them are popular figures then many people will be connecting them so as a result they will have many connections which are common by chance. So to get rid of such a situation we use the Zakad measure. So in this Zakad measure we can take care of the common neighbor immediate common neighbor however the nature of this common neighbors will not be explored.

So as a result if those common neighbors are again popular figures it may create problems so their weights need to be reduced. So that gives rise to another variation of this neighborhood based method which is called Adamic Adder method and this the weights are defined based on how many number of neighbors they have. So which means higher the number of neighbors the weight decreases. So based on this now we can also find out the links to be predicted. So these are my references that figure one example I took from here and most of these as from this book.

So there are four types of structural recommendation in a network. Finding nodes by authority by example and by their power to influence and we can also recommend links. Page rank algorithm is the basis for this first one recommending next by authority. Finding nodes with similar interest we can utilize iterative classification algorithm so this is for the second one recommending by example and recommending by influence we use those methods like independent cascade model and linear threshold model and of which linear cascade model only sorry linear threshold model only we studied and finally for recommending links we studied only about the neighborhood based measures. With this our discussion on structural recommendation in a network is over. Thank you very much.