Lecture 26 : Science Mapping - II

Hello learners, welcome again. In the previous lecture I have discussed about the concept of science mapping, what science mapping is. We have discussed that science mapping is a visual representation of the structure of science. Then we have discussed about what is network and the network is made up of connected things. These things are nodes and the connection between these nodes are known as edges. After that we have discussed about the various application of network analysis and why we used to do network analysis.

Then we have also seen that a network can be either directed or undirected. And in the last we have discussed about the adjacency matrix and few of the centrality measures which measure the strength of the network. So in this particular lecture I will be discussing about the different techniques of science mapping. So these techniques are bibliographic coupling, co-citation analysis, then co-authorship analysis and co-word analysis.

So let us first discuss what bibliographic coupling is. So in 1960 M.M. Koechler from MIT have studied the references in the documents and found the relationship between the two documents. And Koechler have published 7 research reports.

Some of the key studies were an experimental study of bibliographic coupling between technical papers, bibliographic coupling between scientific papers and many more. So what was that relationship? So the relationship between the two documents that sharing the common references. So the concept of bibliographic coupling says that if two documents shares at least one common reference between them then those two documents are bibliographic coupled. And the strength of the bibliographic coupling between those two documents will be calculated based on the number of common references they share. So let us understand this particular concept with an example.

Say for example we have this is document A and it has four references. Reference 1, reference 2, reference 3, reference 4. So we have already discussed about that what a document is and in the document we have some references in the end. So we are assuming that this particular document 1 has four references. Now another example document B which have say for example another references.

So document B has four references and it has reference 1, reference 5, reference 6, reference 7. Now take another document C which have three references, reference 8, reference 9, reference 10. So we have another document D which has six references. So references 1, references 2, references 5, 5, reference 7, reference 6, reference 9. So these are ten references of different papers.

So what exactly this particular case is saying that document A is a published document which has four references. So the document is published and it has four references. So another document B has another four references, reference 1, reference 5, reference 6, reference 7. Then we have document C which have three references 8, 9, 10 and we have a document D which have six references, reference 1, reference 2, reference 5, reference 7, reference 6, reference 9. So all these ten references are different.

Now to see whether a document are bibliographic couple or not what we have to do? We have to see the two documents and then we have to calculate what are the common references they are sharing. If we see this particular case we have four documents. So how many pairs of different documents we have to test? So we have to check one A and B whether they are bibliographic couple or not. Then we have to check B and C. Then we have to check C and D.

So this is one set, this is one set, this is C and D one set and then we have A and D and then we have B and D and we have to see A and C. So these are the six pair of documents we have and now we have to calculate the bibliographic coupling. So if we see A and B document, so we have common references is only this. So we can say that A and B are bibliographic couple. Then we will calculate for B and C.

If we check B and C there is no common references this year. So we will say that it is not a bibliographic couple. So it is not bibliographic couple. Then we will see document C and D. So for document C and D there is like reference one common reference is there.

So we will say that C and D are bibliographic couple. So what that reference is? Reference nine. So C and D is also bibliographic couple. Then we will see A and D. For A and D there is like two common references.

One is this, one is this and here one and two. So A and D are bibliographic couple. So it is also bibliographic couple. Then we will see B and D. So for B and D there is a common reference one is this.

So we have this here. Then five we have. Then we have six. Also six we have.

Then we have seven. We have seven here also. So B and D are also bibliographic couple. A and C, they are not sharing any common references. So we will say that A and C are not bibliographic coupled. So our bibliographic couple documents are now, if we say bibliographic couple document, we have A and B.

Then we have C and D. Then we have A and D. And then we have B and D. Now we will calculate the strength between these bibliographic couple. So for A and B there is only one common references.

So its strength is one. Now for C and D there is a common reference of one. So we will say that its strength is also one. For A and D we have two common references. So one and two, one and two. So we will say that strength of A and D is two.

For B and D we have one. Then five we have. Then we have six. Then we have seven. So we have four documents. So the strength of bibliographic couple between B and D is four.

So this is how we calculate the bibliographic coupling between the two documents. So the study of bibliographic coupling is considered as measure of forming the cluster having similar research interest. So here we have a different clusters. So these are different cluster have been formed. So now you guess that what kind of graph will be formed between the bibliographic coupling of these documents.

So will it be a directed graph or a undirected graph? So if we see here it will be undirected graph because if we are saying A and B is bibliographic coupled that means B and A are also bibliographic coupled. It is a mutual relationship. So what we have discussed about the undirected graph. So if we see here that this is A, this is B. So these are bibliographic coupled.

Then C and D. So there is C is there and then this is D. So they are also bibliographic coupled. Then A and D. This is also bibliographic coupled and B and D.

This is also bibliographic coupled. So this is how our network representation of this bibliographic coupling will be shown. And if we see here the strength of these bibliographic coupled what exactly it is. So A and B it is 1. C and D it is 1.

A and D it is 2 and B and D. So if we have to show the higher bibliographic coupling strength between the two documents how we will do? We have already discussed that we will make this edge thick. So this is the thick edge and then this will be the another thick edge and these edge will be thin edge. So this is what a bibliographic coupling is. So the strength between the bibliographic coupled documents is static. So why it is static? Because once the document is published those reference will be fixed.

It won't be updated later. So we are saying that document A and B bibliographic coupled and there is strength is 1. So it is static. Because the document is published so bibliographic coupling is directly proportional to the number of common references they have. More the number of common references more will be the strength of bibliographic coupling. So for A and B if we have 10 common references then this strength of this particular bibliographic coupled will be 10.

So the range of strength of bibliographic coupling between the two documents will be from 1 to the number of common references they have. If the bibliographic coupling strength is 0 that means the documents are not bibliographic coupled and if it is 1 so the documents are bibliographic coupled and their strength is 1. So if A is having 50 references and B has 40 references and say for example this document B has all the 40 references common to A so the bibliographic coupling strength of A and B will be 40. It can't be 50 because B doesn't have those 10 references. So this is about the strength of bibliographic coupled documents.

Now see another example of a document index in a Scopus database. So this is an interface of Scopus. We have already discussed that what Scopus is and how we extract the data from this particular database. So I am searching here Facebook and topic and okay. So what I am saying here is that that in Scopus search all the documents where the titles have Facebook topic and modeling all those three words in the title itself.

So these are the nine documents which have been retrieved from Scopus. Now if we take the example of this let's take the example of this one and then this. So if we see these documents so somehow this it is not loaded but for bibliographic coupling we don't need the other details here we need the references here only. So if we see these two documents so what all two documents we are seeing in so we are looking for this LDA based topic modeling on hospital Facebook post and then the second document is our this the strategic usage of Facebook by local governments a structural topic modeling analysis. Okay now we have to check whether these documents are bibliographic coupled or not.

So what we will do we will search this. So like we have this is one document this is another document by this author and this. So now we'll see whether we have. So the same paper we have here in this document also and then in in this document also like by this author published in 2003 journal of machine learning and research and if we see here the same document. So what exactly it shows here is that these two documents are bibliographic coupled and their strength is at least one. Okay so we haven't analyzed the all the references of the first document and the second documents also we can say that that this particular document has 89 reference and another this document has 18 references.

So we can say that the bibliographic coupling strength of these two documents will not be higher than 18. Okay so the maximum thing it can be like all those 18 references whatever in the second document can be also in the first document if it is not so the bibliographic coupling strength will be lesser than 18. Okay so this is what about the bibliographic coupling is what this I have taken up his air is by just an example and these all things will be visualized with the help of different software which we will be covering up in the next week. So let us now discuss another similar technique known as co-citation analysis. So it was given by Henry Smoll in 1973 that Henry Smoll is also discussed about the relationship between the documents what that relationship was.

So the earlier one the bibliographic coupling studied by Kessler was where the two documents are analyzed whether they are sharing the common references or not. Okay but in co-citation analysis we have to analyze whether the two documents are co-cited in a single document. Okay and how many these documents are there which are citing these two documents together in their references. Okay sometimes people get confused with this bibliographic coupling and co-citation but this is the difference. Okay so what bibliographic coupling is bibliographic coupling analyze the documents that are sharing the common references.

Okay and what co-citation analysis is so the co-citation analysis analyze the two documents which are cited together in other document. Okay so Henry Smoll discussed this relationship in the paper co-citation in the scientific literature a new measure of the relationship between the two documents and it is measured between two documents that are cited together in the published literature. The cluster formed by co-citation analysis describe the structure of science expert in the area and how fields evolved over time. Okay so the analysis helps in knowing the influential publication and then prominent authors and journals and drawing the thematic clusters. Okay so let us take a practical example of what co-citation analysis is.

So let us assume that we have six papers. Okay so what these papers are so let us assume that this is paper one then this is paper two then we have paper three paper four paper five and paper six. Now each paper have references we have already seen that so let us assume that it has many references but it also has references of document one and document two then paper two has also many references and it has document two and document three. So these document one and document two are another papers okay then document two and document three then we have paper three which have all the references and document two and document three then we have paper four we have document two document one then we have paper five which have document two and document four and then we have paper six document one and document two. Okay now if we have to analyze the co-citations analysis of the documents but we have to see we have to select the two documents so these are the two documents and then we have to see that how many papers are there which are cited these two documents together. So we see here document one and document two has been cited in paper one so this is where one and two we have then we have here also one and two and then we have here also one and two.

So document one and two we have three papers paper one paper four and paper six. Okay so one and two has been co-cited together in these three documents then if we see document two and three it is cited here so document two and three has been cited here is in paper two and then paper three and if we see document two and four it has been co-cited together in paper five. Okay so similarly we can give the strength also of this co-citation so we see the strength of document one and two is three and then document two and three is two because it has been co-cited in two documents and then document two and four is one because this particular set of documents have been cited only once together. Okay now similar to bibliography coupling what kind of graph will be there for co-citation analysis it will be a directed or an undirected graph so it will be a undirected graph. Okay so how undirected graph so document one is there document two so we have because if we are saying that document one and document two are co-cited together or document two and document one are co-cited together these are the same meaning the mutual relationship is there.

Okay so document one and two is co-cited together it will edge is there then document two and three we have a edge and then document two and four. Okay so this is a network graph of co-citation analysis when we are doing for the documents. Okay so when we were discussing about the bibliography coupling the strength of bibliography coupling is static because once the document is published reference will not be further changed but in the case of co-citation analysis when we are talking about the two documents they are co-cited together the strength will be dynamic. Okay like say for example now we are taking the example here is of document one two and then two second and third and second and fourth document. So if we see the strength of second and four is only one but maybe in future like there will be many other documents will be published which co-cite together these two documents then the co-citation strength of this particular set of document will be more compared to the other thing.

So that is why the strength of co-citation analysis is dynamic in nature. Okay so the similar way what we have discussed about the Scopus example of bibliography coupling. Let us now discuss about the example of co-citations analysis of documents. Okay and if we see these were our two documents we discussed about the bibliography coupling of this particular document and this document. Okay so now we are doing the analysis of co-citation analysis of documents.

So how it will go? So let us see we are taking these two documents one is that latent richlet allocation and another is this probabilistic topic models. Okay and if I search on this document we have this probabilistic topic models then this latent richlet allocation. Okay so the co-citation exist here because these two documents latent richlet allocation and probabilistic topic models have been co-cited together in this document also. Okay and similarly so here also those two documents have been cited together. So how we will say that co-citation analysis of this particular two documents? So now we will say that the documents which has the title probabilistic topic models and the latent richlet allocations are co-cited together in two documents.

Okay but when we were saying the bibliographic coupling we were saying that the strategic usage of Facebook by local governments a structural topic modeling analysis this paper and this paper LDA based topic modeling on hospital Facebook post. So these two papers are bibliographic coupled. Okay so this is the common difference many times we get confused in the bibliographic coupling and co-citation analysis. Okay so from here we can easily get the idea of how co-citation can help us to know the structure of science. So taken an another example that this is a document one maybe on R package.

Okay then we have document two on LDA. So LDA is a popular text mining algorithm for topic modeling we will be discussing in the upcoming weeks then this is a document three which is on python package. Now to do the LDA analysis we have different package in different kind of environment. So we have the package for LDA in R also and python also and if we do analyze the co-citation of this and this. Okay and say for example I am just taking an example that R package and LDA co-citation is around like say for example 100 and if I take the co-citation strength is 100 and for this LDA and python package the strength of this co-cited togethers are maybe 15.

Okay so I am just taking a hypothetical example. Okay so what I am saying here is that that there are 100 documents which have co-cited R package and LDA and there are 15 documents which have cited LDA and python package. Okay so from here it can be easily underlined that the LDA analysis using R package is quite popular and the researcher are doing the research on LDA analysis using R package. So this is how the co-citation analysis helps you in identifying the similar kind of documents. So now other than the co-citation analysis of documents there are two other types of co-citation analysis we can do. So one is that author co-citation analysis and then the journal co-citation analysis.

So what author co-citation analysis is so it exactly measure the relationship between the two authors if they are co-cited togethers in the publications. Okay so similar to the documents so earlier when we were doing the co-citation analysis of documents we were doing the analysis of documents. Now we will see whether two authors are co-cited together and the similar way we have the third co-citation analysis of journals or maybe we can do for the source also and we have already discussed about journals as they are considered key formal channels of science communication and if we do the analysis of this journal co-citation we get to know about that what are the influential or the prominent journals are there which have been co-cited togethers in the publications. So this is all about the co-citation analysis so we have discussed about the co-citation analysis of documents and similar to that we can do the analysis of authors and then for the journals.

So let us take one example of journal co-citations. So let us take that so these are the 58 documents on this tweets based topic modelling. So let us take this an example. So this is one publication where we are doing the analysis of journal co-citation and this is another paper for our journal co-citation analysis. So for doing the analysis of journal co-citation what we have to do is we have to see these references so I will just see in the result format and similar to here also. So it has 77 references so we will take all those journals what all journals have been here so I will just take it down and let us take this filter it by j.

Similarly here we can see different journals for 133 references so I will just then roll and I will just go with j. So if you note here that the journal now we have the 77 reference in one paper and 133 references in another paper so based on this we can make the pair of journals so in these 77 references these many journals are there and so totals journals are these many journals so I am just filtering it down with j. So here we can make multiple pairs we have to do the co-citation analysis so what I am doing here I am making one pair is that jMIR public health and surveillance so this is one journal and another is journal of medical internet research. So this is one is a co-citation of this particular pair is all because it is already cited co-cited together then similarly if we go here in this particular paper we found that this is a journal jMIR public health and surveillance and then journal of medical internet research. So what we can say here is that that there is a two journals are there which are jMIR public health and surveillance and another is journal of medical internet research both the journals have been co-cited together in two documents.

So one is this document evolution of COVID-19 tweets about Southeast Asian countries topic modeling and sentiment analysis and another paper is this. Analyzing public reaction perception and attitude during the AMPOCs outbreak findings from topic modeling of tweets. So this is how we can calculate the co-citations of the journals but because we have lot many journals that and this is a small example where we have taken the only two sources. But at the same time we have to handle the multiple journals of all the publication so for that we need a some kind of analysis software. So there are like different software's are there which we will be discussing in the upcoming week.

So this is all about the co-citation analysis. So what we have discussed we have discussed about the co-citation analysis of documents. So then we have discussed about author co-citation and also we have discussed about the journal co-citation. So now let us move to another technique called co-authorship analysis. So co-authorship analysis is an important technique in the science mapping where we do the analysis of collaboration. So this collaboration can be between even the authors or maybe the institution or maybe the countries.

So how exactly they are collaborating the particular research to generate the new knowledge. And we know that collaboration in research is important that helps in establishing the communication networks share ideas and generate new knowledge. So the study of the collaboration of author, institution and countries how exactly it is propagating between them is a kind of a major that helps in knowing the structure of collaborative research. So the study of author, collaboration, institution, collaboration and countries collaboration and how it is propagating in the academic literature helps in knowing the structure of collaborative research. So we can do the analysis of co-authorship on either of the authors then in the institution or on the countries.

So to do the co-authorship analysis what kind of data we need? So we need the authors and their affiliation based on that we can assess easily. So we see here we have this author data. This is based on this we can do the analysis of co-authorship that this G. and Jordan have collaborated on research one research paper and then more the number of papers they have for their collaboration we can easily assess that okay this must the strength of the link between them. Then similar to this if we see the affiliation that this is the University of California then another is from Stanford University.

So this is how we can do the analysis of institution co-authorship and similar way if we have the different countries authors so we can have the collaboration analysis of this different countries okay. So the analysis of co-authorship is very straightforward so we can easily see it. So the publication which have only single author they do not have any collaboration with other authors but the publications which have either two or more authors so they have co-authorship for that publications okay. Let us now discuss about the institutional co-authorship.

So how exactly in institutional co-authorship is let us take example of this. So we see here the authors of this particular publications are from two different institution one is a University of Athens and another is a University of Paris. So this is how we can establish the network between these two institution that University of Athens and the University of Paris are co-authoring for this particular publication and the more the number of publication they are collaborating more the their link strength will be there. So similarly if you see here there are four authors are there and in similar way we can draw the co-authorship network of these four authors and the more papers they publish together their co-authorship network will be stronger okay. If you see in this particular data we cannot do the analysis of co-authorship of countries because all the authors belong to the Greece only on single country.

So to understand the country wise co-authorship let us see this example. If we see this particular example so one of the author is from United States and another is from India. Here we can easily map that relationship that there is a co-authorship relationship between India and USA then more will be the publications from these two countries more will be the link strength between these USA and India and similar to the other countries who are collaborating on any of the research. So this is all about the co-authorship analysis. So we can do the co-authorship analysis on either of the authors then we can do for institution then we can do for countries. So how the particular authors are collaborating or how the institution are collaborating on a research topic or research area and how countries are collaborating on particular research to generate the new knowledge.

So till now we have discussed about bibliographic coupling which assess the intellectual structure by analyzing the two documents sharing the common references then we have discussed about the co-citation network which assess the strength of two publications that are cited together in other documents. So then we have discussed about co-authorship where we assess the collaboration between the authors and or either the institution or the countries. So co-auth analysis is another similar kind of technique related to these bibliographic coupling or co-citation or co-authorship where the thematic clusters are made based on the relationship between the words and how two words are mentioned in different documents. So it shows the relationships of words occurring together where we can see the words in a publication. So if we see our data set so this is a title here words are there then we have the words in abstract also then we have the words in author keywords.

So author keywords are the keywords which are provided by the authors during the submission of the publication and then we have index keyword. So index keyword based on MESH medical subject heading or may be these engineering control terms. So and there are like different indexing terms are there. And then we have the words in full text but whenever we are downloading the data from either of the bibliographic data sources we are getting the words only in the title. So we are getting the title then we are getting the keywords and then we are getting the words in abstract but we have the full text also.

So how to extract the words from abstract or title or may be from full text we will discuss when we will be talking about the text mining but for understanding the co-word analysis consider the words from author keywords and index keywords. So co-word analysis is based on a idea that occurrence of a words together shows the research themes in a particular discipline. So first the pair of words are identified when the pair of words are selected after that we search in the text corpus so whatever our data set is we search in that particular data set that these pair of words are occurring or not and if these pair of words are occurring we will collect all those documents and we will make it a cluster. So this is what co-word analysis is where we group the documents based on the findings of the similar words on those documents and this grouping of these documents are done on some clustering mechanism. So the simpler way of collecting the keywords are from like either of the author's keyword or the index keyword but if both keywords are not given then we have to switch to collection of the word from either of the title or abstract or maybe the full text.

So this is one publication and another is this publication. Now let us assume that we have selected the pair of words is big data and Twitter. This is a one pair of word which are co-mentioned in this particular paper then if we go here so we can see that there are like another pair of words the same word big data and Twitter what we have mentioned in this pair. So based on this analysis we can conclude that these two papers are related to big data analytics of Twitter data. So this is how the co-words analysis helps in understanding the research topic of a particular document. So in this whole week we have discussed about the science mapping and we have discussed about how we can do the mapping of science using the network analysis.

We have discussed about the different kind of a network then we have discussed about the different majors to calculate the strength of a network and in the last we have discussed about the few of the technique of science mapping that is bibliographic coupling then co-citation analysis co-authorship analysis and co-word analysis and just we have discussed the this whole this technique with small examples but in reality we have a huge data and analyzing that data can only be done with the help of some software and visualizing it because network analysis is done by visualization. So in the next week we will be discussing about all those visualization technique why exactly we need the visualization what are the important things in data visualization and what are the different software we have how we can use these different software for doing the data visualization and for our analysis. Okay, so see you in the next week. Thank you.