Complex Network: Theory and Application
Prof. Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 20
Citation Analysis – III

In the last lecture we have been discussing about the citation profile of a given paper. So, what we discussed was that when a paper is published then after the point of publication, the paper initially starts getting a citation at a particular rate which accelerates then it goes to a steady state in between 2 to 5 years and then there is a exponential d k in the rate at which the citation is gain by a particular paper.

Now, based on this observation metrics like impact factor etcetera where designed and these observation actually observation actually date back to 2000, 2001, but then given our data set we actually did an extensive analysis and we would show that apart from the average characteristics, average behavior where you have you observe the single peak in the citation profile within 3 to 5 years.

There are 4 other different citation profiles of, but of a paper that is possible depending on which citation profile a paper actually follows or depending on the citation profile of a given paper you can categorize the paper into at least 5 different classes; one class is our known class which was also earlier stated in the literature that is the acceleration then a steady state and then there is a decline.

However, there are four other classes also the second class as we saw was the multiple peak class where you can have more than one peaks in the citation profile. The paper gets some, gain some citation. There is an acceleration gain some citation there is a peak it declines then again there is a rise in this number of citations, then there is a decline and this can continue as a ripple effect.

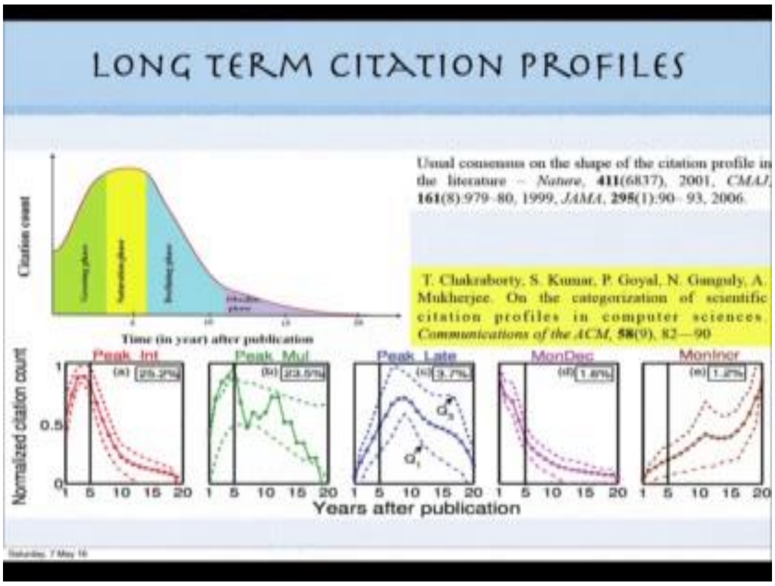
There is a third type where the peak is no where seen between the first 5 years only, if you see late quite late in time you observe that there is a peak in the citation. So, these papers initially did not manage to get high numbers citation they were not popular as such in the initial years after their publication, but then suddenly be due to probably development of certain techniques they became very, very popular and their citation grew up.

Then there is another class which is kind of the most unfortunate class actually where like you see that the citations only d k over time. The rate at which the citations are acquired is actually going down and down declining over time and as a researcher you would never wish to be in this particular category and then there is this last category which is a very interesting category where you observe that the citation only rise and rise there is no decline at all its monotonically increasing over time. So, these are papers which are probably seminal papers from the very beginning. So, they have from the time point they were written they have been heavily applied and they are still being applied into research that is why they are citation never falls over time.

So, as soon as we observe that apart from the citation profile that was recorded in the literature, we have four other different citation profiles also immediately one needs revisit and relook into the definition of bibliographic measures like impact factor. So, since the impact factor definition was based on this 3 year time in time year window.

This will not hold for papers say in the multiple peak category or papers in the monotonically increasing category. Therefore, there is a need to revisit. So, this is word of caution they needs to with people researcher need to think more thoughtfully or they take more thoughtfully the cases of papers which are in other buckets like monotonically increasing bucket or the multiple peak bucket or the let peak bucket.

(Refer Slide Time: 04:32)



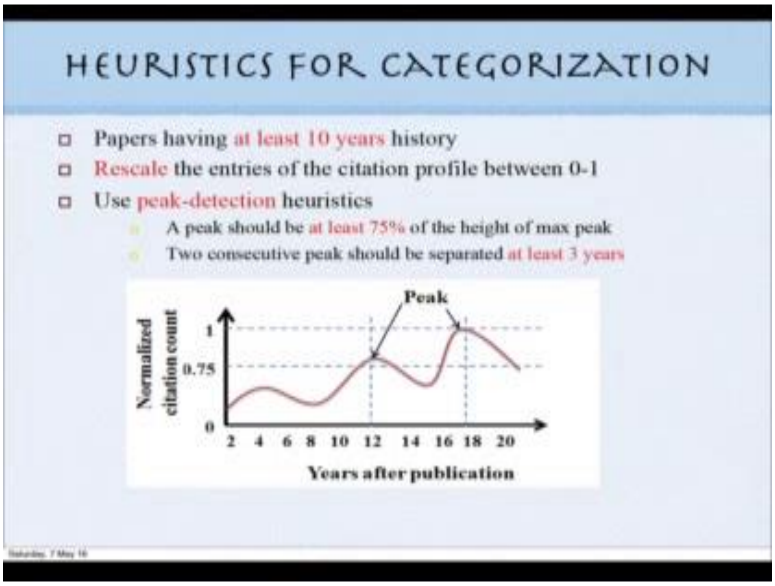
LONG TERM CITATION PROFILES

Usual consensus on the shape of the citation profile in the literature – *Nature*, **411**(6837), 2001, *CMAJ*, **161**(8):979–80, 1999, *JAMA*, **295**(1):90–93, 2006.

T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, A. Mukherjee. On the categorization of scientific citation profiles in computer sciences. *Communications of the ACM*, **58**(9), 82—90

Saturday, 7 May 16

(Refer Slide Time: 04:33)



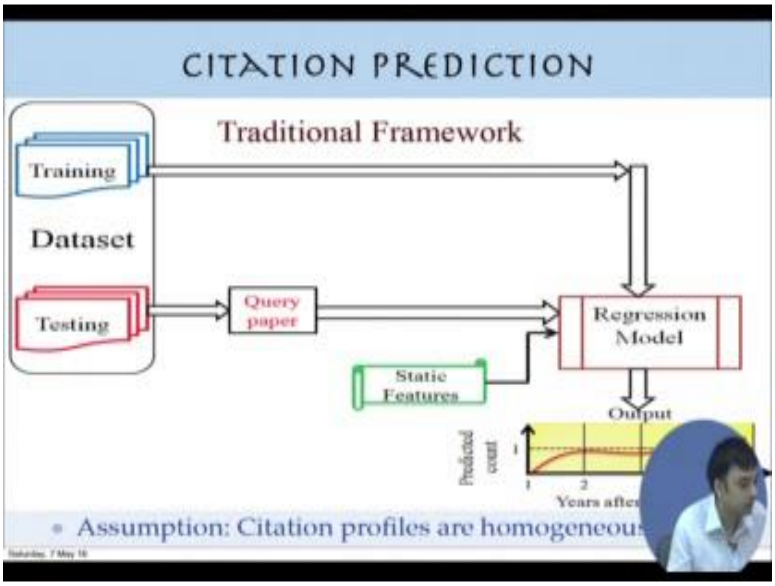
HEURISTICS FOR CATEGORIZATION

- □ Papers having at least 10 years history
- □ Rescale the entries of the citation profile between 0-1
- □ Use peak-detection heuristics
  - A peak should be at least 75% of the height of max peak
  - Two consecutive peak should be separated at least 3 years

Saturday, 7 May 16

Now, we will see some more interesting properties of these different categories.

(Refer Slide Time: 04:38)

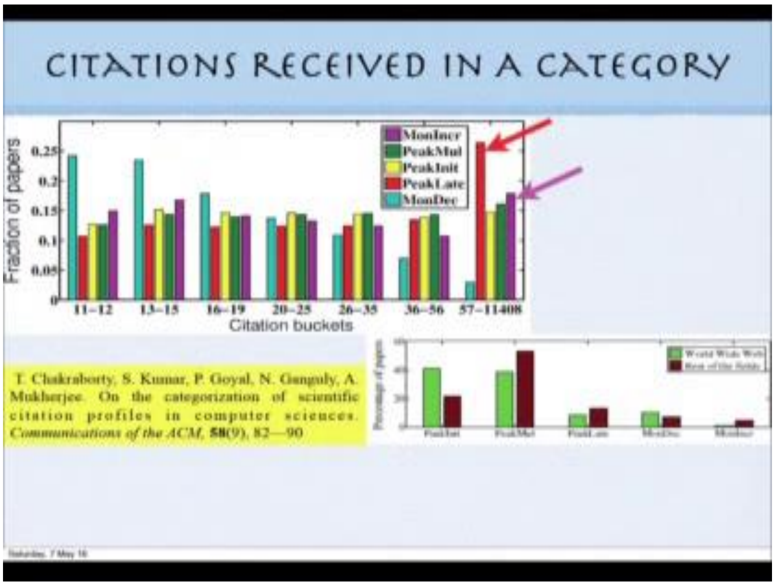


CITATION PREDICTION

Traditional Framework

- Assumption: Citation profiles are homogeneous

(Refer Slide Time: 04:40)



CITATIONS RECEIVED IN A CATEGORY

T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, A. Mukherjee. On the categorization of scientific citation profiles in computer sciences. *Communications of the ACM*, **58**(9), 82—90

The categories recall they are monotonically increasing which we call one mon incr m o n l i n c r standing for monotonically increasing; peak mul standing for multiple peaks peak unit standing for initial peak this is like the one that is that has been usually observed in the literature. So, for peak late it is peaking at a later point in time after 5 years at least and then monotonically decreasing mon dec.

So, if you look at all these categories you see certain interesting observations, you immediately make certain interesting observations. So, on the x-axis here I plot the citation value. The number of citations that are particular paper in from each of this category gets in this citation value is between 11 and 12, this is the number of papers from the different categories. If the citation is between 13 and 15 this is number of papers from different categories. If this citation is between 16 and 1919, in this way you break the entire range of citations in to different buckets into different citation buckets.

And you see that the right most buckets the one which has which actually consist of papers which are mostly citation which are highly cited papers. So, from 57 to 11408, this is the bucket which consists of papers which are highly cited and immediately if you look a little bit carefully, you observe that in this particular bucket in this bucket which has which actually hosts the papers, which have the largest number of citations the majority of papers are contributed by monotonically increasing and the peak late category.

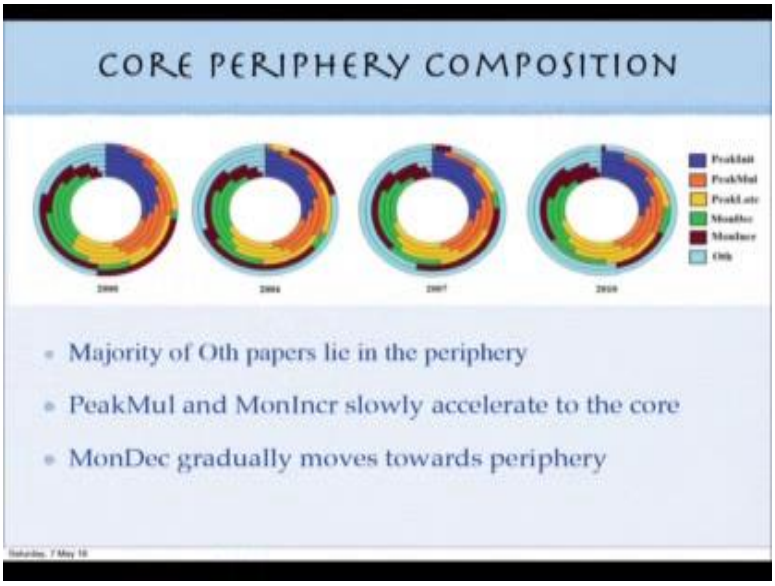
So, these are the categories which are which people, for the peak late category researchers usually do not understand the value of research in the very beginning, but after a point in time say after 4 or 5 years the of the existence of the paper, there is suddenly a search in the use of that technique. So, this peak late paper and the monotonically increasing papers seen to be accruing the largest number of citations, these are the two categories of the papers in which the largest number of citations actually are attracted.

Another interesting observation is that if you look into conferences like World Wide Web and if you look into rest of the other fields, what you observe is that most of the papers that are published in conferences like World Wide Web; they fall in the peak in category. So, that is papers which are published in conferences like World Wide Web what happens is that they have an initial accelerating growth of their citations, then there is a steady face and then there is a decline. So, these papers what I want to say is this papers usually are not very long impact papers.

They do not have a long lasting impact, basically after 4 or 5 years their citation tend to be decline, whereas if you observe the monotonically increasing class most of the papers in this class are published in computer science journals. So, there is always a debate in computer science whether to publish in conferences or journals. So, this is one shallow indication which tells you that journals papers mostly are monotonically are classified into the monotonically increasing class journal, many of the journal papers actually fall in the monotonically increasing class that is journal papers tend to have rise and rise in their citation, whereas conference paper like typically like the world wide web conference the citation is like the growth of citation is like only initial then there is a study phase and then there is a decline.

The point that I am trying to make here is that journal papers tend to classify themselves most in the monotonically increasing class which is like a co-weighted class right. So, you only gain citations there overtime, whereas conference papers are not like that they stay for a short time. They are short timers like they gain citations in the initial years, but then after 4-5 years their citations becomes steady and then finally, decline over time. So, this is a very interesting difference between the journal and the conference papers which is app which becomes apparent from the analysis of our data.

(Refer Slide Time: 09:20)



CORE PERIPHERY COMPOSITION

- Majority of Oth papers lie in the periphery
- PeakMul and MonIncr slowly accelerate to the core
- MonDec gradually moves towards periphery

Saturday, 7 May 16

There is another thing that we study and that is the actually the core periphery analysis.

Basically what you do you can imagine a citation network or a reference. So, one is the contrary of the other, in a basically in a citation network you try to find out what are the nodes, which have like one single citation there in degree is basically equal to 1. You remove all those nodes then once you have removed all these nodes you try to. So, this forms your first cell as if you are like trying to break the network into cells.

The first shell comprises the nodes which have only one in degree, once you have remove them you then remove the next iteration of nodes which have in degree two then you remove nodes which have in degree three in this way you do kind of composition of the whole network and the hypothesis or the intuition is that nodes that belong to inner cells are actually the core of the network form the core of the network these nodes which are actually a part of the inner cells actually determine the core of the network. So, they actually determine the backbone structure of the network, whereas those which are at the periphery like they have 1 or 2 shallow connections and they do not actually define the core structure or the core backbone of the network.

So, if you try to do this core periphery analysis you see certain interesting things happening. So, what you observe here, here we draw the proportional of different categories like each category as I have already told you peak in it multiple peak, late monotonically increasing monotonically decreasing and the last category which is which is like none of the above 5. So, if you look at the proportion of papers from each category in the different cells.

So, these are the different cells of the citation network what you see is that papers which are in the monotonically increasing group and papers which are in the peak late group. These are the most significant groups actually in terms of citations as we have already seen. So, these papers are fast accelerating towards the core of the network. In the initial years, you see so many of them the monotonically increasing paper in the outer most cell, however, as time progresses they migrate inner and inner into the inner most core of the network.

Whereas on the other hand like papers from the monotonic papers from the peak in it group or monotonically decreasing group, they move towards the periphery. Basically papers which are monotonically increasing over time papers or citations are monotonically increasing over timer papers which see usually, see late peak those papers actually are accelerating towards the core of the computer science citation network where as papers which classify themselves into peak in it or monotonically decreasing category they are actually first accelerating towards the periphery of the computer science citation network.

The core the backbone is majorly defined by the monotonically increasing as well the peak late papers. So, that is message that I am trying to give you through this particular slide. So, now after we have categorized the paper of this based on based on a citation profile we have categorized the paper into one of these categories we thought that can these be used in developing certain applications. So, and one application that is that has become very, very important in the present day academic world is to predict the future impact of a particular paper. So, if you one to judge the future impact of a particular papers standing at the current point in time how can you do it.

Basically, one way to measures the impact is to find out the citations that a paper will get say in 5 years time or in ten years time that is a fact that you try to estimate given the current data given say a paper is published today can you predict with high accuracy what is going to be the citation of this particular paper after say 5 years from now or after 10 years from now can we do this quite accurately. So, this idea is actually called the citation count prediction task and this is actually very, very handy when people try to do recruitment.

In the academic world, whenever there is a new application made by a say freshly graduating PhD candidate, you want to try an estimate the impact of his or her papers, but then this person probably has written papers for 3 or 4 years only and it is very difficult to judge the impact of this person or the impacts of the of the papers written by this person within this 3 to 4 years time.

So, that is why in order to identify what will be the long term citation count of the different papers that are written by this particular candidate we need a prediction model like this. So, where we try to predict the citation count of a particular paper or written by a person say after 5 years 7 years 12 years down the line similarly you can use this citation count prediction framework in order to do rankings of academic institutes.

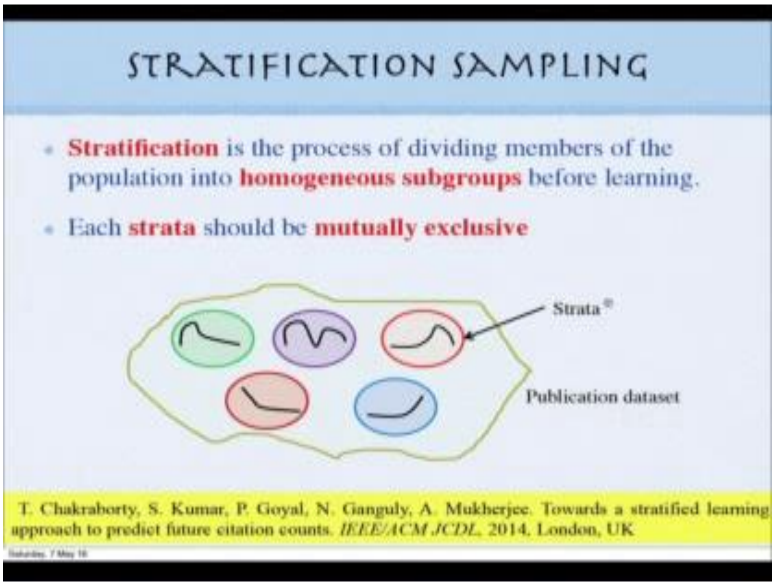
So, if you look at the publication of the current year of the academic institute and try to see what will be each of this publication what will be the citation of each of this publication saying 5 years or 7 years or 10 years down the line you from there you can do a precise ranking of the different academic institutes say some academic institute for some academic institute you see papers which will land up into very high citation after ten years and there are some academic institutes for which you see papers which will not actually land up into large number of citations in 10 years of time. So, you immediately have a way to discriminate between these two academic institutions right.

So, that that tells you why citation counts prediction could be an interesting task in itself. So, this is what a machine learning task where you can use network structure properties. So, this is a unique actually framework where you kind of unite principles from citation network analysis network theory with principles from machine learning by which you do an efficient prediction.

The standard framework the traditional framework for citation prediction is the one that I show over the slide. So, you have a set of training samples there are set of papers that from where you train your model and then there is a set of testing samples from which you draw a query paper now from this training samples you extract a bunch of features and then you fit that into a regression module.

So, this bunch of features are some static features which you extract from the training samples now there is a query paper which you also feed into the regression model and the regression model tells you what should be the citation count of this query paper say after two years three years four years five years of time by looking into the properties of the training sample and trying to find a match probable match between query paper and some of the data points in the training sample. So, that is that is a method where you use a regressor.

(Refer Slide Time: 17:34)



STRATIFICATION SAMPLING

- **Stratification** is the process of dividing members of the population into **homogeneous subgroups** before learning.
- Each **strata** should be **mutually exclusive**

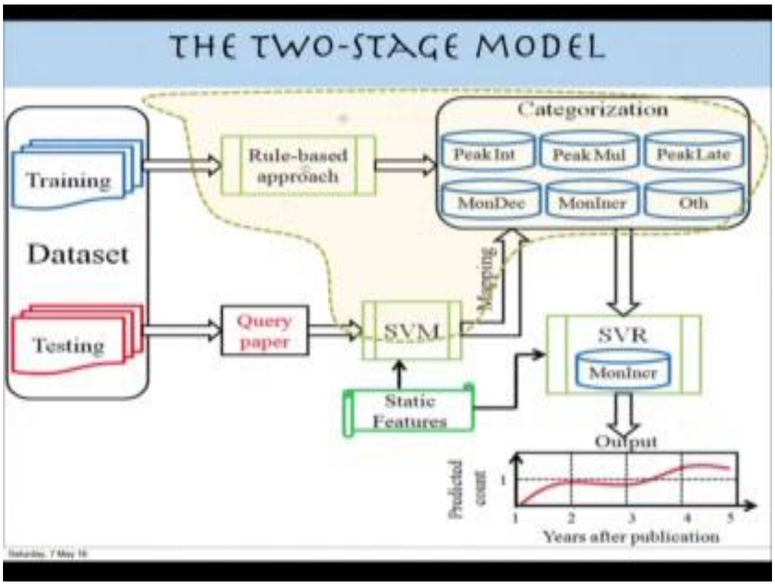
T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, A. Mukherjee. Towards a stratified learning approach to predict future citation counts. *IEEE/ACM JCDL*, 2014, London, UK

So, we try to see like whether this framework of citation prediction using. So, it is a supervise framework as you understand you have a bunch of training samples and then you have a query paper based on the training samples you extract features from recruit training sample and then you use a standard regression model to identify the citation count of a given query paper. So, and you try to do this based on a regressor you try see which paper in the trying sample or which set of papers in the trying sample that your query paper fits well with and you based that you try to predict the citation count of the query paper ok.

So, now we try to like try to advance this particular citation prediction framework by our idea of categories of citation profiles. So, we have observed that a paper can have different citation profiles. And there could be at least 5 different citation profiles. So, a paper could be peak in it a paper could be monotonically increasing a paper could be monotonically decreasing. So, depending on the citation profile we try to divide our training set. So, this idea in machine learning is called stratification.

So, if you look at this slide. Basically you have this training sample which is the publication data set and you divide this training sample into different sets. So, each set corresponds to one category. So, this is like the initially peak this is the monotonically increasing this is the late peak this is the multiple peak. So, you divide the data set into categories of papers. So, you divide the total training sample into strata, each strata is composed of papers of only one single category.

(Refer Slide Time: 19:37)



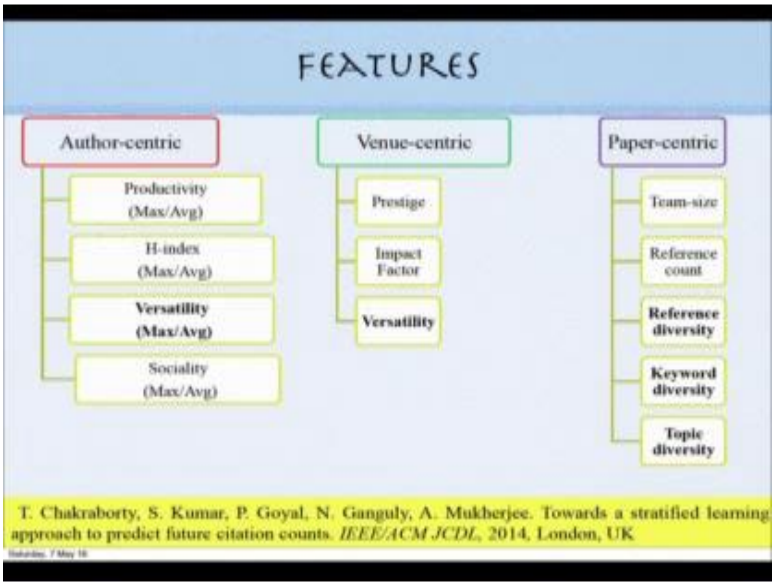
THE TWO-STAGE MODEL

Now, the next task is you have the same training sample, using a rule based approach you divide the training sample into the different categories, initial peak, multiple peak, late peak monotonically increasing and so and so forth. Now, again from the test set you have a query paper. Now, the first step that you do is you classify the query paper using a support vector machine classifier and a set of static features; you first classify the query paper into one of these categories. You try to find out whether this query paper belongs to peak initial or multiple peaks or peak late or monotonically increasing or monotonically decreasing.

Now, if you have classified this query paper into say monotonically increasing class then you take only these training samples. You do not take any other training sample, you take only the class of monotonically increasing papers from this entire data set and train your regressor on that small data set that is smaller data set and our hypothesis is that since by first mapping the query paper into one of this classes what you are trying to do is you are trying to reduce the level of confusion.

So, if a query paper is indeed monotonically increasing and if you have in your data set everything mixed then there is a chance of increased confusion and there is a chance of misclassification, however, if you already know that your query paper actually belongs to the monotonically increasing class then learning the features only based on the monotonically increasing features is more beneficial because in that way you actually deduce confusion. So, we do exactly that and based on only that sub sample only that strata of the training sample, we do our predictions and we see.

(Refer Slide Time: 21:39)



FEATURES

Author-centric

- Productivity (Max/Avg)
- H-index (Max/Avg)
- **Versatility (Max/Avg)**
- Sociality (Max/Avg)

Venue-centric

- Prestige
- Impact Factor
- **Versatility**

Paper-centric

- Team-size
- Reference count
- **Reference diversity**
- **Keyword diversity**
- **Topic diversity**

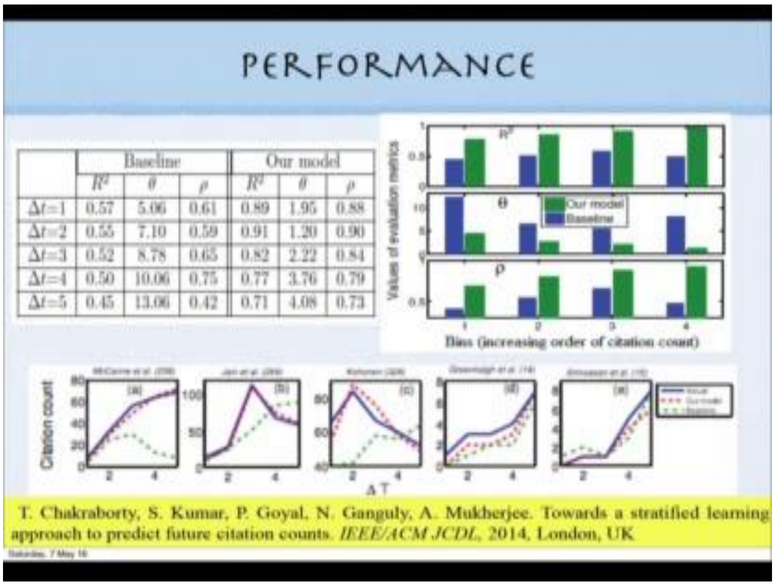
T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, A. Mukherjee. Towards a stratified learning approach to predict future citation counts. *IEEE/ACM JCDL*, 2014, London, UK

In doing this predictions we use this standard features that were used also in the traditional framework. So, there are three different sets of features the author centric feature which includes like the productivity that is the total number of papers written by the author h index of the author the versatility of the author, the sociality or the mixing of author.

Then there are venue centric features like the prestige of the author the impact factor of the venue the versatility of the venue versatility would indicate that what is the difference fields in which papers are accepted in this particular venue then there are paper centric features like what is the generally, what is the number of authors per paper? The total reference count of the paper, the reference diversity, the keyword diversity, the topical diversity of the papers, all this actually goes into defining the static features in classifying or in the classifying the query paper into one of these different strata and then also fitting the regression model.

So, if now, we have one model which do not use this stratification where the training sample is directly used to train the regressor and then there is the base line model and then we have this stratification model, where we have divided the training set into three different categories and we compare the performance of these two modules now. So, we compare the performance based on 3 different metrics, one is the coefficient of determination.

(Refer Slide Time: 23:15)



PERFORMANCE

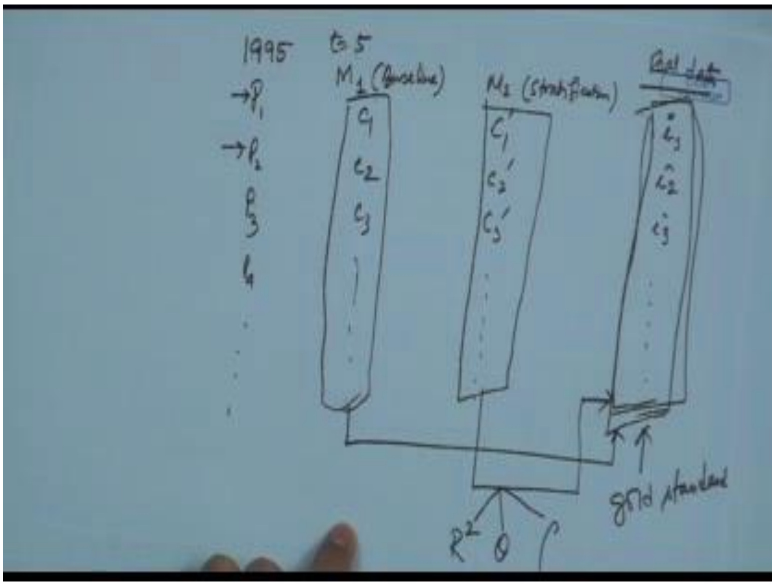
	Baseline			Our model		
	R^2	θ	ρ	R^2	θ	ρ
Δt=1	0.57	5.06	0.61	0.89	1.95	0.88
Δt=2	0.55	7.10	0.59	0.91	1.20	0.90
Δt=3	0.52	8.78	0.65	0.82	2.22	0.84
Δt=4	0.50	10.06	0.75	0.77	3.76	0.79
Δt=5	0.45	13.06	0.42	0.71	4.08	0.73

T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, A. Mukherjee. Towards a stratified learning approach to predict future citation counts. *IEEE/ACM JCDL*, 2014, London, UK

Saturday, 7 May 16

One is the root mean square error and the third is the correlation coefficients. So, these are the three different measures in which you try to find out. Basically what you try to see is your regression model is predicting a citation count and you also know from the real data, what is this citation count of that paper? So, what I am trying to say, suppose you have paper p 1, p 2, p 3, p 4 and so on and so forth.

(Refer Slide Time: 23:44)



Say after 5 years, say t equals to 5 from the model. So, you have model m 1, which is the base line and you have model m 2, which uses the stratification and you have also real data. So, for each individual paper say you are standing at some point 1995. So, at 2000 your base line predict some citation value, say c 1, c 2, c 3, and so on and so forth. So, say for the model, m 2 what is predicted is c 1 prime, c 2 prime, c 3 prime and so on and so forth and the real data. From the real data you already have the citation at say 2000 that is after 5 years is i 1, i 2, i 3 and so on and so forth.

Now, you try to basically see the extent of correlation between this data items and these data items and also the extent of correlation between these data items and these data items. The real data is basically your gold standard and you compare your model predictions, the base line predictions with this old standard data.

The new model based on stratification idea with this gold standard data and we try to do this correlation. We try to estimate this correlation based on three different measures, one is the r square statistic or the coefficient of determination. If you look up the Wikipedia page, you will get a very basic idea of what this measure actually tries to do. It actually tries to plot each point that you get from the model against each point that you get from the real data and then it tries to fit a line against it and the slope of that line is basically the r square fit.

Now, similarly you can have a measure based on root mean square. So, you basically find out the difference of these squares of these two numbers and then you take a square root of that and then you do an average on these two rows and you also find the standard Pearson correlation between these two sets of data items.

Now, based on all these three factors, if you try to see what you observe is that after if you do the prediction after one year then the r square statistic, when you compare the gold standard data with your base line the r square statistic is 0.57, the root mean square is 5.06 and the correlation is 0.61. When you have this stratification included the r square statistic actually increases many folds the root means square decreases, whereas the correlation coefficient increases.

Basically observe that if you have higher r squares statistic that is better. So, that is which this means that your gold standard data is in high agreement with what your model s predicting. Similarly, if you have a high correlation value high Pearson correlation then also you can say that your gold standard data is in high agreement with your model predictions, whereas, if you roots means mean square is low then you can say.

So, while the other two are should be high to indicate that there is a high agreement between the real data between the gold standard data and the model the root for the root mean square, these value should be low indicating a low value would indicate that the gold standard data is in higher agreement with your model predictions and we see that whatever time point you take delta equals after 2; 1 year, after 2 year, after 3 year, after 4 year or after 5 years in all cases the r square and the row values are better for the model that uses stratification then the base line model.

So, this simple idea dividing this trading sample into strata or into groups actually gives you an enhancement in the performance of citation prediction. So, you do your citation prediction much better if you base your predictions on the stratification idea that is the very simple and elegant modification that we being into the traditional framework and this simple modification actually enhances the performance of the system many folds all in terms of r square statistic in terms of theta as well in as in terms of rho.

And on the right hand side I also show the performance in different citations zones. So, if the papers are low cited then also we do pretty well. So, you see these are the bin 1 are all those papers which are low cited papers, bin 2 and 3 are median cited papers and bin 4 is the high cited paper.

So, you see predictions doing predictions for high cited papers is easier while doing predictions for low cited paper because if you have less evidence doing prediction for the low cited papers is usually harder, but what you see over model this stratification idea actually also performs very well in the low citation zone.

In the low citation zone the r square with the ground truth as well as the rho with the ground truth is much higher for the model based on stratification idea, whereas theta is much lower. So, this actually shows the power of this simple modification that we have introduced.

Thank you.