Complex Network: Theory and Application

Prof. Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

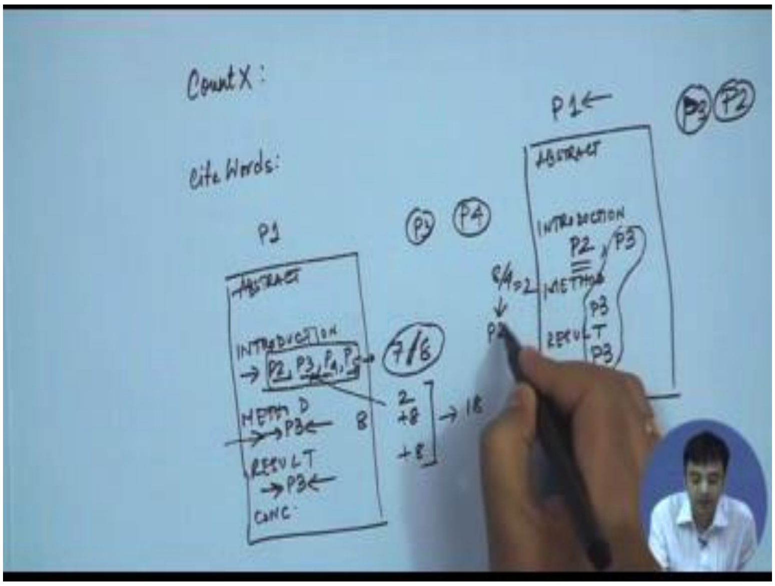
Lecture - 21
Citation Analysis – IV

Welcome to this last lecture of this complex network theory and application series. So, we have been looking into the properties of citation network analysis and last day, we have seen that the citation profile of a paper can be categorized into at least 4 different categories and then we have seen how we can use this idea in order to do citation count prediction efficiently.

So, today we will further develop on this idea and what we shall see today, especially is like all this time we have been looking into the link features mostly that is we have been looking into the in degree and the out degree of the citation network, but then if you remember that each scientific article actually is a collection of a large volume of text.

So, a scientific article can be thought of a corpus of text basically and today, we will see that if we can further use some text features in order to enhance the citation prediction model that we have already proposed the stratification model that we have proposed in the last lecture. We will try to further enhance that model using certain interesting text level features. So, the 2 text level features that we will talk about today are one which we called count x and the other is which we call cite words.

(Refer Slide Time: 01:44)



So, basically if you look into the scientific document, if you look into a paper, you will see that the paper actually talks about some other paper at difference in different points in paper. Suppose you have a paper p 1.

Now, this paper p 1 has, say an abstract, then there is an introduction section, then there is a methodology section, then there is a result section and then finally, there is a conclusion section. Now, let there be 2 other papers; p 2 and p 3. Now, say for instance paper p 1 talks about p 2 in introduction whereas, it talks about p 3 in introduction in method and as well as in results.

So, this immediately tells you that paper p 3 is actually more important for paper p 1 than paper p 2, paper p 2 is referenced only once whereas, paper p 3 is referenced at least 3 times. So, this actually is the basic idea behind the formulation of count x, basically count the number of times that a particular paper in this case p 3 is being invoked by some other paper in the entire set of references. This actually defines the idea of count x.

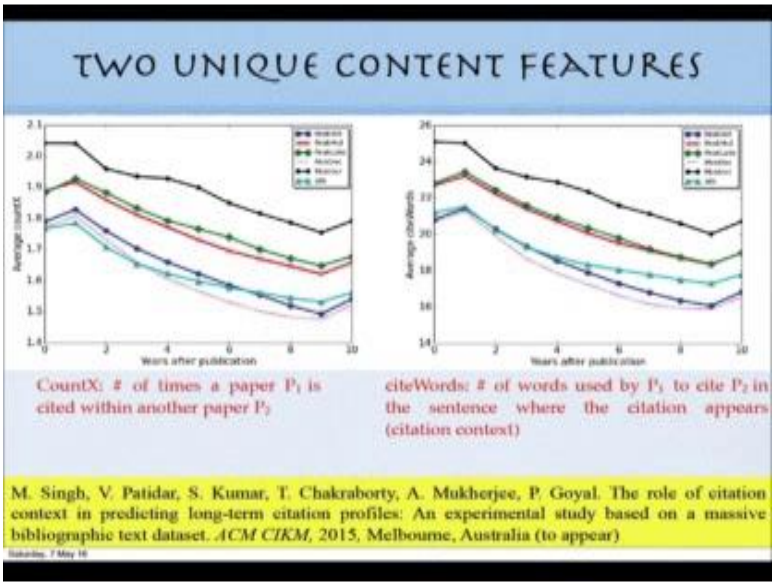
So, the next idea, cite words is basically also interesting. So, what you try to observe here again you can take 2 examples, say p 3 and p 4 and say you have abstract introduction method result and conclusion. Now, say the paper p 2 is invoked only in introduction with many other papers and there is one sentence which paper p 1 actually writes about all these papers together whereas, paper p 3 actually is invoked in the method section and there is one full sentence that is dedicated to paper 3.

Again in the result section there is one full sentence that is dedicated to paper p 3. So, you basically find out what is the average number of words that is dedicated to paper p 3 whereas, what is the average number of words that is dedicated to paper p 2.

Now, here if there are like 7 words or 8 words in this particular sentence which invokes p 2, p 3, p 4, 5 then this 7 or 8 words are actually shared between all the references and you say that p 2 on an average gets a share of 8 by 4 that is 2 words. If there are 8 words in the sentence, p 2 gets a share of 2 words whereas, p 3 probably gets a share of 2 words from the introduction, say if here in the method section again say if there are 8 words dedicated to p 3. So, 2 plus 10 and again in the result section if there is a sentence with 8 words. So, then another 8 words are dedicated to p 3.

So, basically the total number of words that are dedicated to p 3 is 16 plus 2; 18 whereas, the total number of words that are dedicated to p 2 is just 2 so; that means, again this is an indication that p 3 is probably a more important paper for p 1 than p 2. Now, the question is given this information can we do our citation on prediction even better?

(Refer Slide Time: 06:26)



TWO UNIQUE CONTENT FEATURES

CountX: # of times a paper P_1 is cited within another paper P_2

citeWords: # of words used by P_1 to cite P_2 in the sentence where the citation appears (citation context)

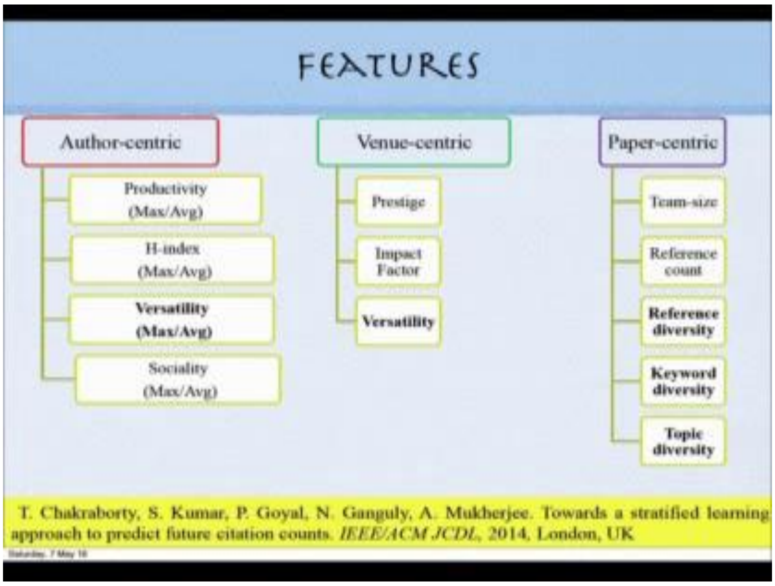
M. Singh, V. Patidar, S. Kumar, T. Chakraborty, A. Mukherjee, P. Goyal. The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. *ACM CIKM*, 2015, Melbourne, Australia (to appear)

Saturday, 7 May 16

So, what I show you in these 2 slides is basically, if you take papers from the different categories that we introduced last year peak in it multiple peak late monotonically decreasing monotonically increasing you see that the count takes value that is the total number of times that a paper is being referred in the text by any other paper.

So, the category is monotonically increasing and late peak these are the 2 categories that are at received the highest count x values. So, these are the papers which are invoked by other papers a lot in their text a lot compared to all the other categories of papers. So, basically again you see a discrimination between these 2 sets of categories compared to the other different categories. So, in terms of count x, you see that the count x values are usually larger for these 2 categories of papers. Similarly, in the cite words you see the monotonically increasing as well as the peak, let papers have on an average larger cite words than other categories of papers. Now, you introduce these 2 additional features in your set of features that we have already discussed.

(Refer Slide Time: 07:49)



FEATURES

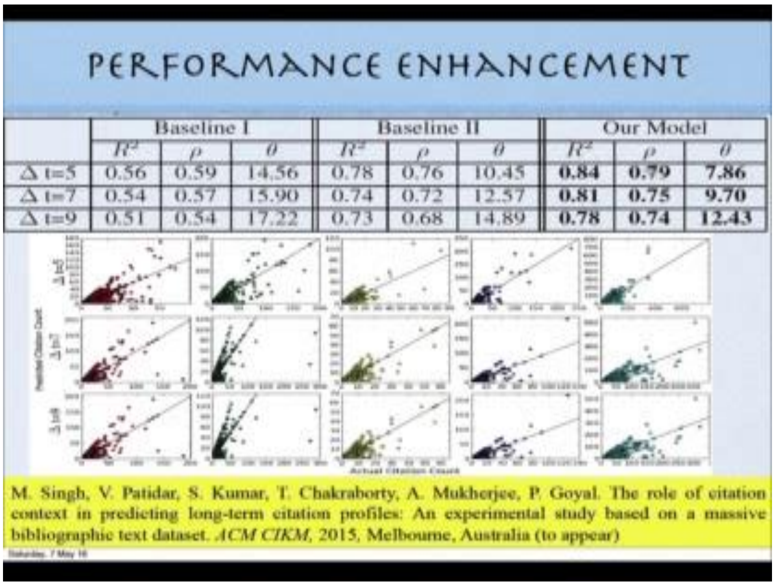
- Author-centric
  - Productivity (Max/Avg)
  - H-index (Max/Avg)
  - **Versatility (Max/Avg)**
  - Sociality (Max/Avg)
- Venue-centric
  - Prestige
  - Impact Factor
  - **Versatility**
- Paper-centric
  - Team-size
  - Reference count
  - **Reference diversity**
  - **Keyword diversity**
  - **Topic 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

Saturday, 7 May 16

If you look at the slides that we already discussed and there are some author centric features, some venue centric features and some paper centric features which you use to train your regressor. Now, in addition to these features, if we add the 2 interesting and very, very different features that we have just now defined the cite words and the count x what we observed is.

(Refer Slide Time: 08:12)



PERFORMANCE ENHANCEMENT

	Baseline I			Baseline II			Our Model		
	R^2	ρ	θ	R^2	ρ	θ	R^2	ρ	θ
Δ t=5	0.56	0.59	14.56	0.78	0.76	10.45	**0.84**	**0.79**	**7.86**
Δ t=7	0.54	0.57	15.90	0.74	0.72	12.57	**0.81**	**0.75**	**9.70**
Δ t=9	0.51	0.54	17.22	0.73	0.68	14.89	**0.78**	**0.74**	**12.43**

M. Singh, V. Patidar, S. Kumar, T. Chakraborty, A. Mukherjee, P. Goyal. The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. *ACM CIKM*, 2015, Melbourne, Australia (to appear)

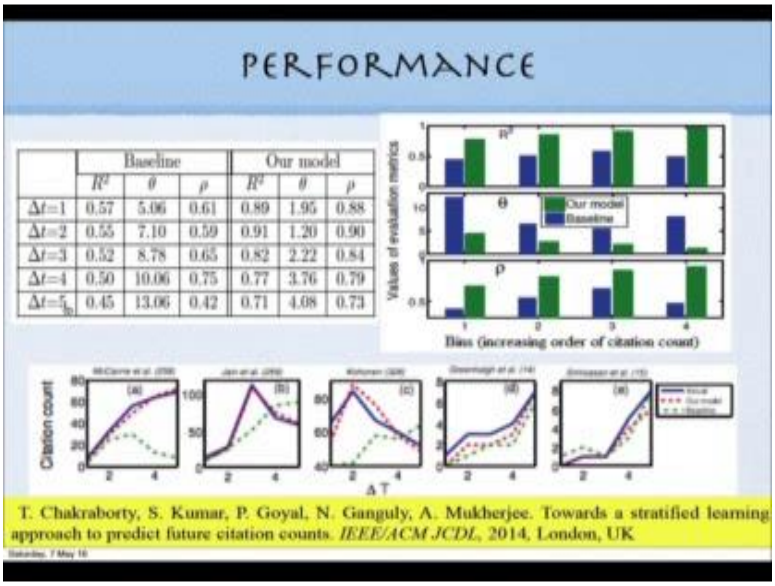
Saturday, 7 May 16

That these 2 extra features allows us to obtain the r square, the rho and the theta that we also obtain earlier the things that the values that we obtain using these 2 set of features, using these 2 additional features actually agree with our ground truth data far better than the 2 previous models.

Remember, we had the first baseline, which had no stratification, the second baseline which had the stratification of the training sample into 4 categories and then we have this third model which has the stratification plus additionally these 2 features the cite words and the count x. So, including these 2 features what we observe is that we further improve the r square the theta and the row values

So, the agreement with the ground truth data is actually further enhanced if you use this very 2 simple with this very 2 simple features based on text properties and not only that, but we also observed. So, you should remember in our last results, we had looked at a time gap of 1 year, 2 years and up to 4 years.

(Refer Slide Time: 09:32)



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

4 years down the line, what should be your predicated citation history, but inclusion of these 2 simple features actually allows us to predict reliably even after 9 years of publication. So, start stay a standing at the current point in time you can do a prediction

of the citation count even after 9 years quite reliably.

If you just use these 2 simple features these 2 features are very, very simple, but then since they are extracted from the text properties they are very strong discriminative feature features and they actually give you a very good intuition about the citation count of a particular paper in the long time. So, that actually gives you a brief idea of how the notion of categories of citation profiles of papers as well as these 2 simple text properties can enhance the task of citation prediction which is used as I said in various different tasks like academic recruitment's university rankings etcetera.

(Refer Slide Time: 10:46)

FATE OF "ANCIENT" PAPERS

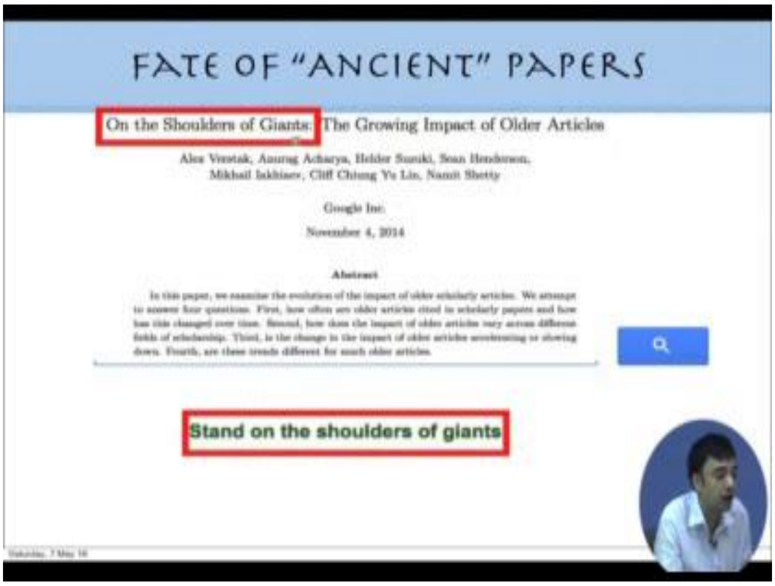
Stand on the shoulders of giants

So, in the last lap of citation analysis, we will try to look into the citation history or the citation count or the citation behavior of what we call ancient paper which are pretty old say 10 years old, 20 years old. So, we will try to see what is the citation behavior of these set of papers usually in the community it is thought that as a paper becomes older and older the number of citation that it gets becomes lower and lower and that is what we also saw in our citation profiles more a majority of the papers actually classify themselves into these peak in category.

As we have already seen that is why the average behavior also resembles the peak in its category. So, mostly in the first few years, we will again allow a large number of citations and then it will actually decay over time. So, that is what we have observed.

Now, in the rest of the analysis, we will try to dwell a little bit deeper into studying this citation behavior of older peoples, is it really like that or are there certain interesting further observations that 1 can make actually I guess all of you are aware of this particular portal. So, is the standard Google scholar portal and or of late Google scholar has started writing this particular slogan on it is portal stand on the shoulders of the giants. So, basically 1 would wonder why such a slogan is being reflected on the page of Google scholar.

(Refer Slide Time: 12:38)



FATE OF "ANCIENT" PAPERS

On the Shoulders of Giants: The Growing Impact of Older Articles

Alex Verstak, Anurag Acharya, Helder Suzuki, Sean Henderson,
Mikhail Iakhiaev, Cliff Chiung Yu Lin, Namit Shetty

Google Inc.

November 4, 2014

Abstract

In this paper, we examine the evolution of the impact of older scholarly articles. We attempt to answer four questions. First, how often are older articles cited in scholarly papers and how has this changed over time. Second, how does the impact of older articles vary across different fields of scholarship. Third, is the change in the impact of older articles accelerating or slowing down. Fourth, are these trends different for much older articles.

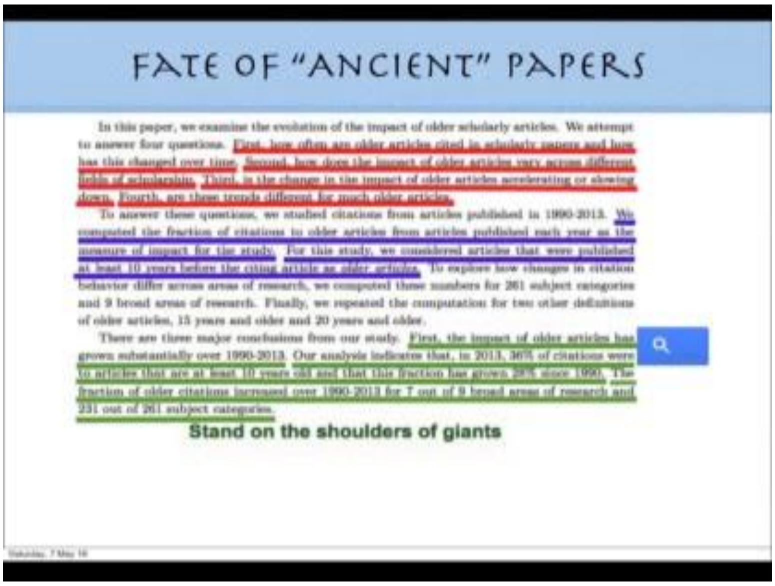
Stand on the shoulders of giants

So, this actually is connected to this paper which was written by a set of authors who actually are the lead developers of Google scholar, Anurag Acharya is the leader of this team and they wrote a paper on the shoulders of giants, the growing impact of older articles.

So, in this paper what they show is that older papers actually are gaining citations over time. So, in general community believes that older papers over time are losing citations. They are going into obscurity slowly and slowly whereas, this is a paper which tries to show that it is not always true papers, which are 10 years old, 20 years old may get a large number of citations. So, accrue a large number of citations in general and that is what they try to do using some quantitative data analysis and show that there are a large number of old papers, who are getting a huge volume, who are attracting a huge volume of citations.

So, to make the difference clear the usual observation in the scientific literature is that papers that grow older in time actually go into oblivion and obscurity. They continuously loose citations, whereas this is one particular work which tries to show that this is not true. They try to challenge this idea and try to show that papers which are 10 years old or even 20 years old actually accrue a large volume of citations. So, this work actually attracted a lot of media attention, since they the teams of authors were trying to tell something which is very, very different from what is observed so far. So, this work actually attracted a lot of media attention.

(Refer Slide Time: 14:41)



FATE OF "ANCIENT" PAPERS

In this paper, we examine the evolution of the impact of older scholarly articles. We attempt to answer four questions. First, how often are older articles cited in scholarly papers and how has this changed over time. Second, how does the impact of older articles vary across different fields of scholarship. Third, is the change in the impact of older articles accelerating or slowing down. Fourth, are these trends different for much older articles.

To answer these questions, we studied citations from articles published in 1990-2013. We computed the fraction of citations to older articles from articles published each year as the measure of impact for the study. For this study, we considered articles that were published at least 10 years before the citing article as older articles. To explore how changes in citation behavior differ across areas of research, we computed these numbers for 261 subject categories and 9 broad areas of research. Finally, we repeated the computation for two other definitions of older articles, 15 years and older and 20 years and older.

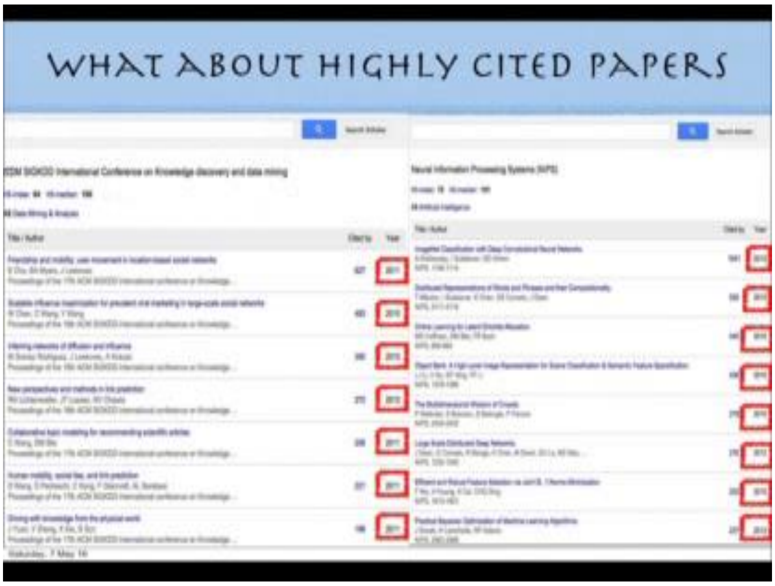
There are three major conclusions from our study. First, the impact of older articles has grown substantially over 1990-2013. Our analysis indicates that, in 2013, 36% of citations were to articles that are at least 10 years old and that this fraction has grown 28% since 1990. The fraction of older citations increased over 1990-2013 for 7 out of 9 broad areas of research and 231 out of 261 subject categories.

Stand on the shoulders of giants

So, we try to see like, and they make a claim which I am trying to state here in this particular. So, I have just extracted the claim that they have made. So, what they should say is that 36 percent of the citations it is were 2 articles that are at least 10 years old. So, basically what they saw is that large fractions of citations are going to papers which are old which are pretty old which are like 10 years old. So, this is what their primary hypothesis was and they use interesting data analysis techniques to establish this fact, but then we try to look into a little bit more details about this analysis. So, we were kind of thinking like how these 2 contradictory ideas can be existing at the same point in time.

While the literature so far have said that older papers grow try to lose citations over time and they go into oblivion and obscurity these authors are telling that it is not true and that older papers actually attract a large volume of citation. So, how to reconcile these 2 opposing observations, the first thing that we did was we tried to look into 2 top tier conferences in the computer science area.

(Refer Slide Time: 16:11)



WHAT ABOUT HIGHLY CITED PAPERS

ACM SIGKDD International Conference on Knowledge discovery and data mining

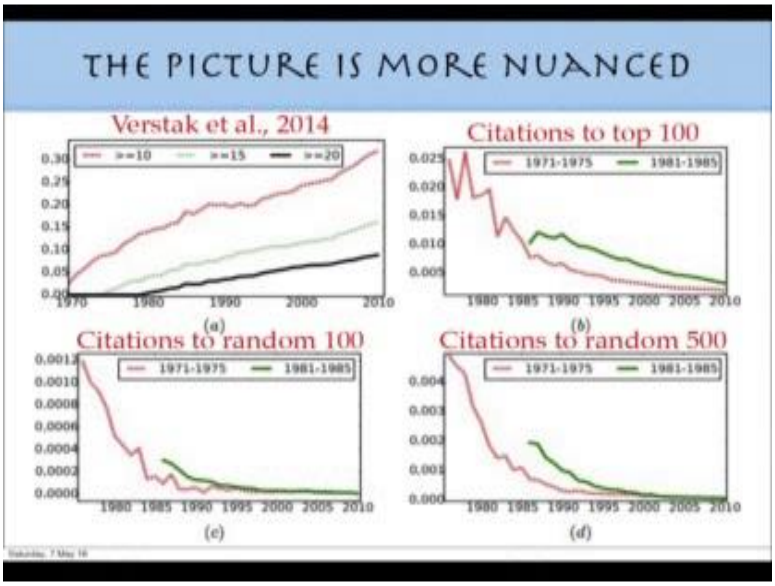
Neural Information Processing Systems (NIPS)

So, one of them as I already told you is the data mining conference ACMC International Conference on knowledge discovery and data mining and which is the largest and the most important conference in data mining in the area of data mining and you have another conference, the neural information processing systems which is a top tier conference in machine learning.

So, if you look at the most highly cited papers in these 2 conferences, what you see is that most of these papers are not very old most of these papers. So, the highest cited paper is like was authored in 2011, the next highest was authored in 2010. So, these are like only 4 or 5 years old. So, the claim that the authors of Google scholar were making that most of the citations the largest 1 the high volume of citations is going to ancient papers that are 10 years old is actually not being observed at least in this simple data analysis that we do.

So, the same holds for the other conference also you see for the other conference, the papers that get the largest number of citations are even more recent 2012, 2013, etcetera. So, where lays the fallacy? So, why do the scholar Google scholar authors observe that there is an increasing volume of citations that is going to older papers? And what we essentially see is actually summarized in this particular slide.

(Refer Slide Time: 17:48)



THE PICTURE IS MORE NUANCED

So, Google scholar authors were observing that the volume of citations that were going to older papers which were like 10 years or old 15 years old or 20 years old that is actually rising over time. Now, what they did not consider is that the volume of publication is also growing exponentially over time. So, this factor they did not count in. It is not only the decitation of a paper is growing, but also the total number of papers published in a particular field in computer science is growing over time.

So, if you look the profile of publications is actually also exponentially rising over time. So, these factors has not been taken into account in their analysis and therefore, you get an observation like the one that they have seen that is the total number of citations that is going to the older papers is rising because the volume of older papers is actually rising over also rising over time. So, since this volume this rising volume is not factored into the analysis properly that is why you observe that there is a steady growth in the fraction of citations that is going to older papers.

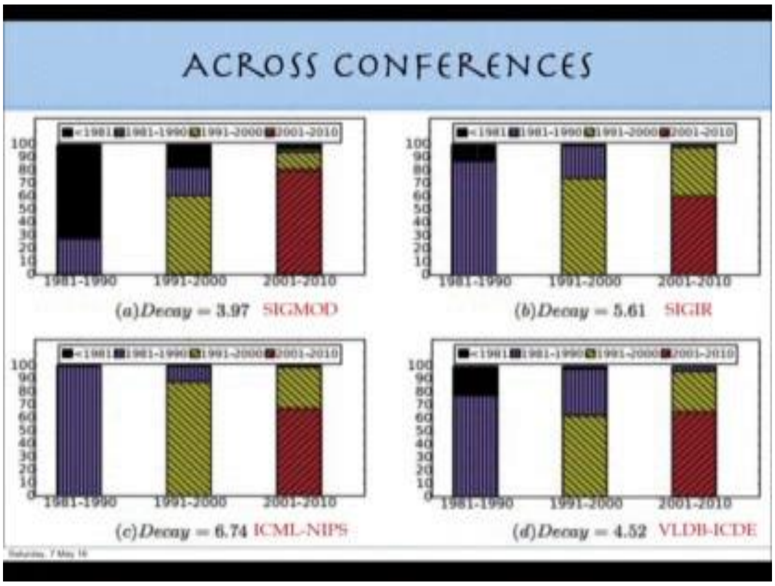
So, in order to factor in this idea of growing number of publications over time what we try to do is we try to fix a set of papers and we try to see like, let us take papers which are 10 years old and only top 100 cited papers of that time say papers which the top 100 cited papers which are 10 years old take those papers and see what is the fraction of citations that are going to these top 100 papers not the entire set of papers that are 10 years old not that entire exponentially rising volume of papers, but to a fix set of papers only the top 10, top 100 highly cited papers which are 10 years old.

If you look at that profile you see that, actually the fraction of citations going to such papers is actually declining which is opposite to what the Google scholar authors observed and it is not only true for the top 100 cited papers, it is also true for any random set of 100 papers. If you take a random set of 100 papers random, but fix set of 100 papers, which are 10 years old, you observe that the fraction of citations going to those papers is actually over a decline.

So, basically the hypothesis that people have already observed in the literature long back that older papers tend to usually on an average go into oblivion, and obscurity is actually holding from our data analysis the glitch that was there in the analysis made by this Google scholar team is that they did not factor in the growing volume of number of publications. So, that is, basically when you are doing large scale data analysis that the analysis could be. So, this particular observation gives you an idea how tricky such analysis could be and you have to be very, very cautious that you are doing your analysis appropriately and correctly especially while normalizing things.

So, then what we try to do is like. So, the Google scholar team observed that there is a rising volume of citations whereas, we observed that there is a steady declined in the rate of citations. So, can these 2 things be nicely summarized into 1 picture, so that is what we want to now do? So, Google scholar team actually saw if you look at the unlimited volume of publications were exponentially growing volume of publications you observe that there is a rise in the number of citations that is going to older papers, however, once you make the set fixed you observe the defraction of citations going to these fix set of old papers basically is on a decline. So, now, can you put this 2 information together in 1 single quantitative plot? That is where we will try to do in the next slide.

(Refer Slide Time: 22:35)



ACROSS CONFERENCES

$(a) Decay = 3.97$ SIGMOD

$(b) Decay = 5.61$ SIGIR

$(c) Decay = 6.74$ ICML-NIPS

$(d) Decay = 4.52$ VLDB-ICDE

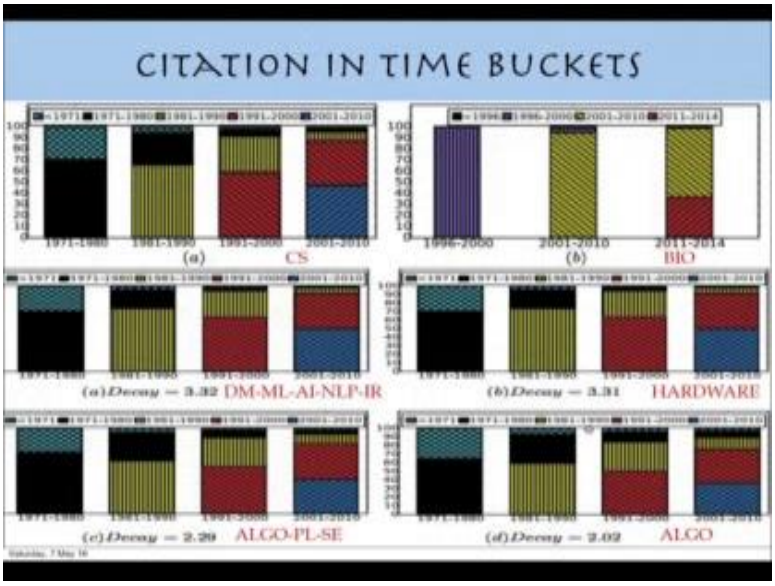
Saturday, 7 May 16

So, here we look into different venues in Computer Science; SIGMOD is a venue which predominantly publishes in databases and data mining, SIGIR is a venue which predominantly publishes works in information retrieval, ICML-NIPS is predominantly are venues which publishes works on machine learning and VLDB-ICDE again are venues which publish work on large scale databases. Now, for each of these venues what we try to do is we segregate papers based on decades. So, there is as we show, we segregate papers into 3 decades 81 to 90, 91 to 2000 and 2001 to 2010. Now, in each decade what we try to plot is the fraction of citations that is going to that particular decade and the immediate previous decade, the next immediate previous decade and so on and so forth.

So, for instance, let us look at this particular bar. So, what we observe here this is the height of the yellow color actually tries to show, what is the fraction of citations that is going from the papers of 1991-2000 to the papers of 91 and 1991-2000 itself then the purple box tries to show, what is the fraction of citations that is going from the papers of 1991-2000 to the immediate previous bucket that is 81 to 90 and the black bucket tries to show, what is the fraction of citations that is going to papers beyond 1981.

So, if you try to see in each of these plots, we try to estimate these fractions and then plot it as stat bar charts across different fields across different venues of computer science. Now, what you see here immediately that, in a particular decade what we have seen in a particular decade. If you look at a particular decade the fraction of citations that is going to the total fraction of citations that is going to the immediate set of previous decades, if you take the purple plus the black. This is the total number of citations that is going to the previous decades and here you have yellow plus purple plus black that is going to the that makes the that constitutes the total fraction of citations that is going to the immediate previous decades.

(Refer Slide Time: 26:33)



CITATION IN TIME BUCKETS

Now, what you see that the total height of purple plus black here is larger than the total height of yellow plus purple plus black. So, you see, this is what actually the sticks paper was observing the total volume of citations going to older decades is actually higher. So, here the total volume of citations that is going to the older decades is higher than here. So, this is the total volume of citation that is going to the older decade, but then what they missed is, what is the fraction of citation that is going to the current decade?

So, that actually constitutes the major fraction that is what they did not observe they only looked into this part this part of the data. So, and that is why they were observing that this is actually on a rise whereas, they actually completely did not observe this particular fraction which is going to the immediate, which is going to the current decade itself and these now, we extend this analysis to the entire computer science field to the another domain of biomedical documents to a bunch of sub fields like data mining, machine learning, artificial intelligence is taken together. So, this is the data mining, machine learning area, and then you have the hardware area, algorithms area, algorithm programming language and software engineering taken together.

In each of these, what you observe is that say, for instance, take the entire computer science field, the entire computer science domain, look at the decade 1981 to 1990, the total fraction of citations that is going to the older decades is black plus this blue whereas, the total fraction here in 1991 to 2000, here is yellow plus black plus blue. Now, the height of yellow plus black plus blue as you see here is larger than the height black plus blue and the height of red plus yellow plus black plus blue is even larger than the immediate previous decades.

So, that is the total fraction of citations that is going, the total volume of citations that is going to the previous decades they all taken together is actually growing overtime. So, this height is smaller than this height and this is yet smaller than this height. So, this is what actually was observing, but then there is another fraction which is going to the decade itself and that is actually the largest chunk and this is what actually (Refer Time: 28:09) picture does not show us at all.

So, we here try to show both these fractions together. So, this picture actually reconciles the 2 facts; the fact that there is a growing number of citations that is going to the older fields because the height is actually rising, but then if you look at the total number of citations going to the current decade itself that is actually very, very high compared to what is going to the previous decade.

So, recent papers actually are getting considerably large number of citations consider compared to the older papers. So, it is not really true that ancient papers are growing, are getting larger fraction of citations over time. So, this actually brings us to the end of this particular course. We have tried to see many different aspects starting from the theory of complex networks, where we have looked into the basic properties like centrality measures, social network, roles etcetera to community detection which is another theoretical prospective of complex networks and then finally, we ended up with a case study where we tried to show applications of all the methods that we have come across.

So, far and how one can meaningfully use this in order to do various quantitative analysis on large scale networks and do various predictions like citation count prediction and prediction of like whether really older papers are getting, ancient papers are getting larger and larger a number of citations.

So, we would end here and I hope that you have enjoyed the different parts of this course and I hope that you will find that this course interesting and useful in various applications that you take up from now onwards.

Thank you very much.