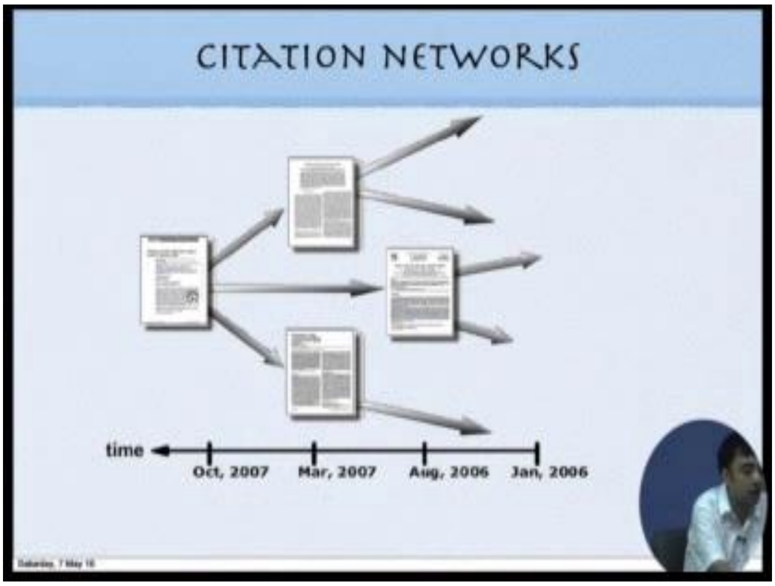
Complex Network: Theory and Application
Prof. Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 18
Citation Analysis – I

So, we have come to the last lap of this lecture series, and as we already said, that we will discuss a bit about a case study on citation networks. So, for the rest of this lecture, we will discuss mostly; how one can builds such build big networks and extract the various interesting properties from these networks. We have already defined the idea of citation networks in one of the earlier classes, in the initial lectures, but in order to recap. So, let us look at the slides.

(Refer Slide Time: 00:56)



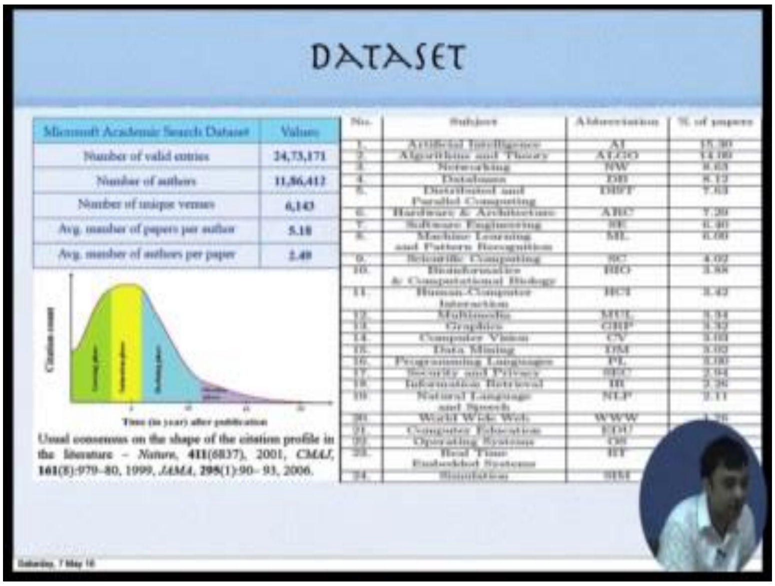
CITATION NETWORKS

So, here I draw a citation network, where each node represents paper or a scientific article. And if a particular scientific article refers to another scientific article, then you draw directed edge like this. So, this particular scientific article has, in it is reference list the name this particular scientific article, and in this way you construct this network.

Basically this is a directed graph, which has both the concept of out degree; which is the total number of edges or citation edges that go out from this node and also in degree; that is the total number citation edges that enter into this particular node. So, here the out degree for example, for this particular node, is equal to three, where as the in degree for this particular node is equal to one. Moreover another interesting property of this citation networks is that, they can be constructed over a time line.

Basically papers get published at a particular time, or citations can be made to only the earlier papers. So, it is not possible to site some paper which has not been written so far. So, that is why only edges point backward in time. you see here is the time, but time is in this direction. So, edges only point backward in time, because you cannot actually refer to something which has not been written so far, and therefore, these graphs does not usually have a directed cycle. So, with this basic notion, we will try to see what can be a way to construct large citation networks, and then do meaning full citation analysis on this network.

(Refer Slide Time: 02:44)



DATASET

Microsoft Academic Search Dataset	Values
Number of valid entries	24,73,171
Number of authors	11,86,412
Number of unique venues	6,143
Avg. number of papers per author	5.18
Avg. number of authors per paper	2.49

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.

No.	Subject	Abbreviation	% of papers
1.	Artificial Intelligence	AI	15.30
2.	Algorithms and Theory	ALGO	14.00
3.	Networking	NW	8.63
4.	Databases	DB	8.12
5.	Distributed and Parallel Computing	DIST	7.63
6.	Hardware & Architecture	ARC	7.20
7.	Software Engineering	SE	6.40
8.	Machine Learning and Pattern Recognition	ML	6.00
9.	Scientific Computing	SC	4.02
10.	Bioinformatics & Computational Biology	BIO	3.88
11.	Human-Computer Interaction	HCI	3.42
12.	Multimedia	MUL	3.34
13.	Graphics	GRP	3.32
14.	Computer Vision	CV	3.03
15.	Data Mining	DM	3.02
16.	Programming Languages	PL	3.00
17.	Security and Privacy	SEC	2.94
18.	Information Retrieval	IR	2.36
19.	Natural Language and Speech	NLP	2.11
20.	World Wide Web	WWW	1.76
21.	Computer Education	EDU	[illegible]
22.	Operating Systems	OS	[illegible]
23.	Real Time Embedded Systems	RT	[illegible]
24.	Simulation	SIM	[illegible]

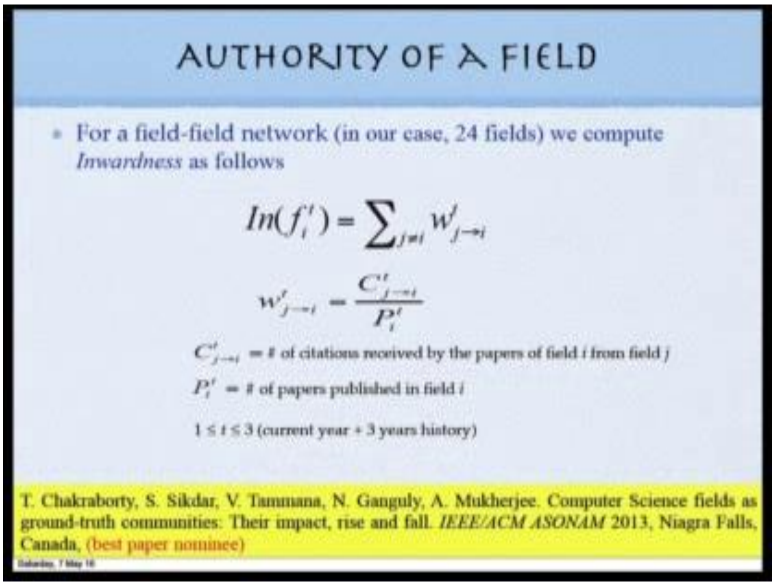
Saturday, 7 May 16

So, to start off, the data set that we will consider, is scrolled, or is built from the Microsoft academic search data set, which is publicly online repository from where, at least 2.4 million entries where scrolled 2.4 million papers; papers like these scientific articles, 2.4 million such papers where scrolled. The total number of authors across all these was where around 11 lacks number unique venues, journals or conferences where these papers were published, is like around 6000, average number of papers per author is like 5, and average number of authors per paper.

So, usually in computer science, these data is mostly related to the computer science field. Usually in computer science what is observed is that, most articles are written by 2 or 3 authors. So, usually you do not see papers written by 10 20 authors, as you seen other domains like biology.

So, further these data is also segregated into fields. Basically along with each paper, you also know to which field that paper belongs to. To which field that scientific article belongs to. So, for instance, the 24 fields that are considered in this data set are these; artificial intelligence algorithms and theory networking, data bases distributed in parallel computing and so on and so forth. So, there are 24 such fields of computer science, and this is the proportion of papers, that are available in the data set from each different field.

(Refer Slide Time: 04:49)



AUTHORITY OF A FIELD

- For a field-field network (in our case, 24 fields) we compute *Inwardness* as follows

$$In(f_i^t) = \sum_{j \neq i} w_{j \to i}^t$$

$$w_{j \to i}^t = \frac{C_{j \to i}^t}{P_i^t}$$

$C_{j \to i}^t$ = # of citations received by the papers of field i from field j

P_i^t = # of papers published in field i

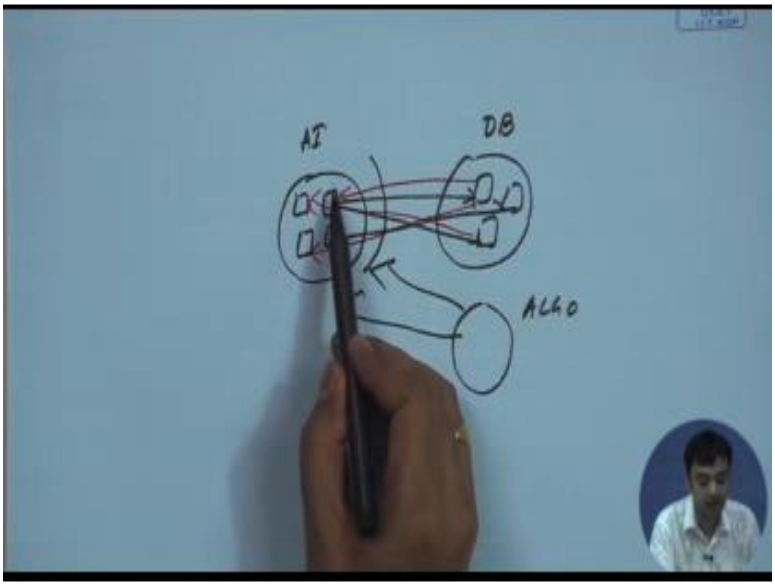
$1 \leq t \leq 3$ (current year + 3 years history)

T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, A. Mukherjee. Computer Science fields as ground-truth communities: Their impact, rise and fall. *IEEE/ACM ASONAM* 2013, Niagra Falls, Canada, (best paper nominee)

Saturday, 7 May 16

Basically what you have, along with this particular graph you also know, that whether this particular paper belongs to algorithms, or it belongs to artificial intelligence, or it belongs to data bases, or whatever field it belong. So, you have the field information for each particular paper, along with the simple citation information. So, with this we start of the preliminary analysis. So, the first thing that we define is called the authoritativeness of a field. This is basically nothing, but the total number of citation links, that are coming to particular field from other fields. So, let us take an example and see; so for instance.

(Refer Slide Time: 05:14)

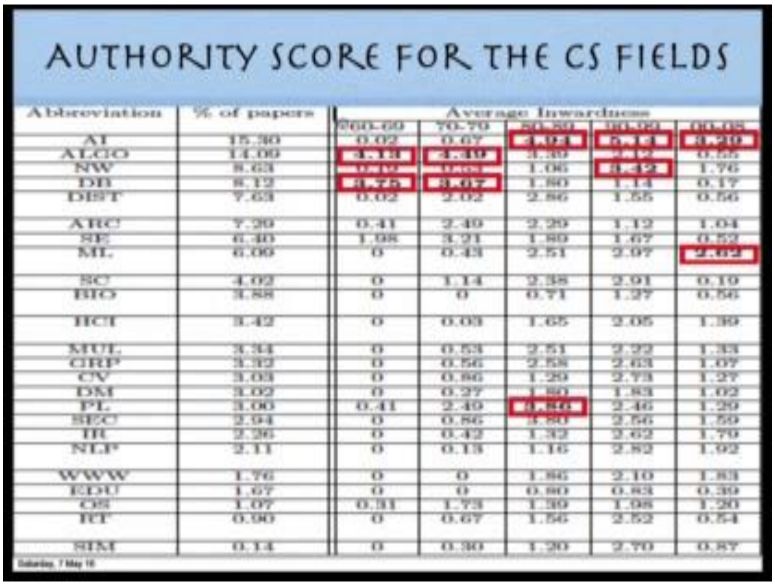


If you have papers from two different fields; say this is a i; for artificial intelligence, and this is data base is. And now there are certain citations going from a i to data bases, and there are certain citations, that are coming from data bases to a i. So, you basically find out by the authority measure, you find out what is the total number of citations that are coming to a i from all other fields; say data bases, algorithms.

So, what is the fraction of citations that are coming from all other fields to a i, normalized by the total number of papers that are published in a i. So, that is basically what is the fraction of citations, that each paper in a i is enjoying from other fields that is what we want to quantify. So, that would tell you what is the authority of a i. Basically that would tell you, in some sense the importance of the field a i.

So, that is telling you; what is the total number of in citations that the field of a i is enjoying from all other different fields. So, you are trying to estimate, what is the fraction of citations that each paper in a i gets from all other different fields, not including a i. So, this tells you what is the authoritativeness of a i over all other fields. So, this is a very simple measure that we define. And we see that this simple measure actually gives certain very interesting signals as we see in the next slide. If you plot, or if you write down the values of these authority scores for the different decades; so as I said, the citation network is time line network. So, there is time information.

(Refer Slide Time: 07:16)



AUTHORITY SCORE FOR THE CS FIELDS

Abbreviation	% of papers	Average Inwardness				
		60-69	70-79	80-89	90-99	00-08
AI	15.30	0.02	0.67	4.94	5.14	3.29
ALGO	14.09	4.13	4.49	3.39	2.12	0.55
NW	8.63	[illegible]	[illegible]	1.06	3.42	1.76
DB	8.12	3.75	3.67	1.80	1.14	0.17
DIST	7.63	0.02	2.02	2.86	1.55	0.56
ARC	7.29	0.41	2.49	2.29	1.12	1.04
SE	6.40	1.98	3.21	1.89	1.67	0.52
ML	6.09	0	0.43	2.51	2.97	2.62
SC	4.02	0	1.14	2.38	2.91	0.19
BIO	3.88	0	0	0.71	1.27	0.56
HCI	3.42	0	0.03	1.65	2.05	1.39
MUL	3.34	0	0.53	2.51	2.22	1.33
GRP	3.32	0	0.56	2.58	2.63	1.07
CV	3.03	0	0.86	1.29	2.73	1.27
DM	3.02	0	0.27	1.80	1.83	1.02
PL	3.00	0.41	2.49	3.86	2.46	1.29
SEC	2.94	0	0.86	1.80	2.56	1.59
IR	2.26	0	0.42	1.32	2.62	1.79
NLP	2.11	0	0.13	1.16	2.82	1.92
WWW	1.76	0	0	1.86	2.10	1.83
EDU	1.67	0	0	0.80	0.83	0.39
OS	1.07	0.31	1.73	1.39	1.98	1.20
RT	0.90	0	0.67	1.56	2.52	0.54
SIM	0.14	0	0.30	1.20	2.70	0.87

Saturday, 7 May 16

So, if you look at the citation network of 60 to 69, 1960 to 69, then you can look the next decade 70 to 79, 80 to 89, 90 to 99 and 2000 to 2008. So, in each decade, you see that there are some fields that have very high authority scores. For instance, in the first decade 60 to 69, the two fields that have the largest authority score, are algorithms and data bases. So, immediately you observed that this was the time when computer science was in it is infancy, computer science was still growing, and this was the time when the basic principles of computer science were being developed. So, these principles were mostly developed in the fields' algorithms and data bases; that is why these are two most authoritative fields of that decade.

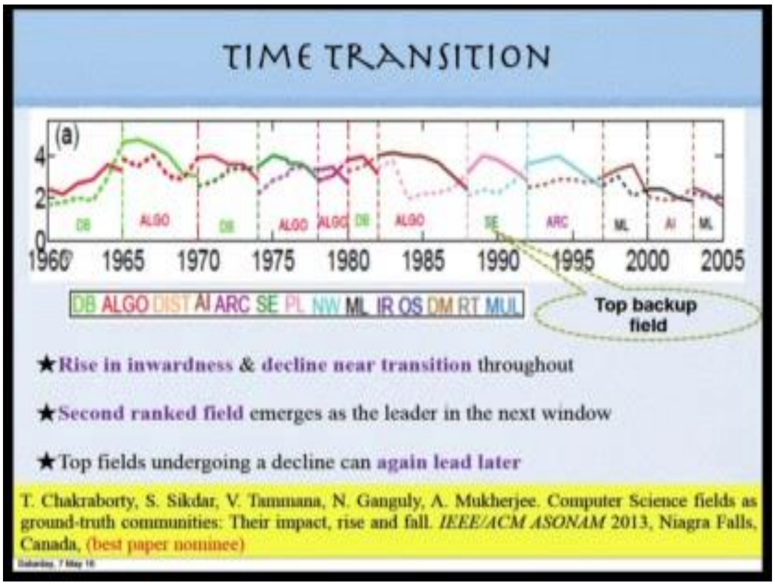
In the next decade, the same observation continues; algorithms and data bases, so still that field is developing. in the third decade you already see a difference. So, you see that a i has received a very high authority score, as well as the other field that has received a very high authority score is the programming language. So, a i has received a a very high authority score, because this was the time 80s was the time when computer science or the a i boo. People were trying to do everything solve all problems using a i techniques. So, that is why a lot of researches were working in this area.

A lot of people papers being published till. There was a lot of interest actually in this area, and that is why the authorities score has risen up. The authority score of p l has risen up, because this was the time, this was the decade when lots of programming languages where written like C, C++ etcetera, the compilers were being constructed in this particular decade.

If you look at the next decade a i is still at the top, but then there is another field, another interesting observation, that networks has entered in. So, network has also received a very high authority score; no word that this is the decade when, we had the internet boon. So, internet has had had script at this particular decade, and that is why there was a lot of research going on in networks; in internet and networks. So, that is why the authority score of networks has also come up. And in the last decade, you see that a i is still at the top, but then there is another field that has gained a lot of importance which is, machine learning, and this is very operate.

Now every other school every other top school has machine learning course, and has one or more machine learning courses, and these as become one of the hottest topics of the computer science area. So, these simple analysis of the citation networks based on the field information, and the authority score that we defined in the last slide, actually clearly shows; how the authority of the different fields has changed over time, how the importance of the different fields have changed over time. So, in order to further look at these interesting behaviors at a more micro level, we draw something called the time transition diagram.

(Refer Slide Time: 10:33)



TIME TRANSITION

★Rise in inwardness & decline near transition throughout

★Second ranked field emerges as the leader in the next window

★Top fields undergoing a decline can again lead later

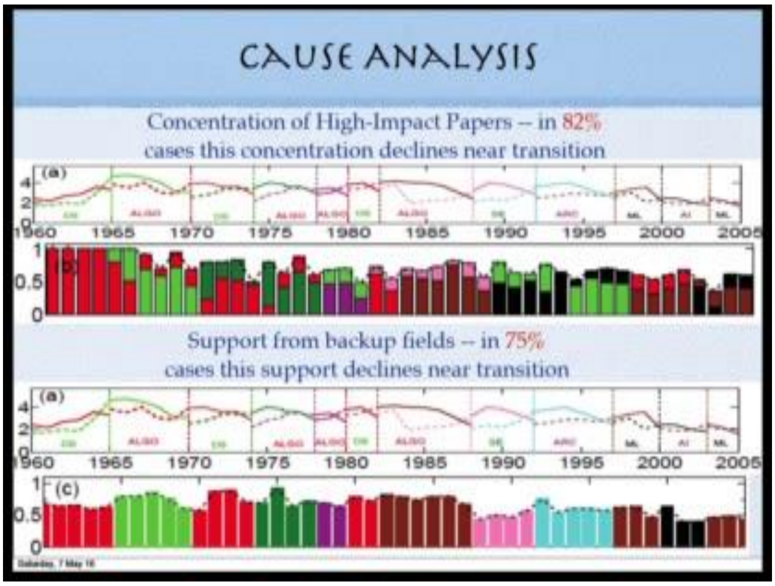
T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, A. Mukherjee. Computer Science fields as ground-truth communities: Their impact, rise and fall. *IEEE/ACM ASONAM* 2013, Niagra Falls, Canada, (best paper nominee)

So, this is a diagram that is usually drawn in various physics experiments. So, what you do here is on the x axis you have the time line from 1960 to 2005. And on the y axis you plot the authority score for the two top ranking fields, based on the authority score. So, the top one and the second one; so the top one in the first few years, as we see is algorithms, followed by data bases. Then in the next few years it is data bases, followed by algorithms and it keeps on going like this. So, one very interesting observation that you has, from this particular slide is that.

Look in any particular time window, if some field is in the second rank, then in the second time window, then in the immediate next time window, that field becomes the top rank field. Basically this is a very interesting observation, and this holds for hundred percent cases. So, what you observe in a previous time window, if there is a particular field which is at the second rank of authority score, then it comes at the top rank in the next particular time window. So, this tells you that. These actually gives you an indication that, if you know the current second ranking field in terms of authority score, then you already know who is going to be the topper in the next set of time window.

So, that actually, can also give you an hint of which field you should perceive your research career in. basically the field which is like the hot field, which has a very high authority score, you would always hope to perceive a research career in that particular field. And if you want to pre estimate what field that would be, you have to look in to the second ranking field of the current time window. If you can tell what is the second ranking field of the current time window, then you can tell with 100 percent guarantee, that these guy is going to be top ranker of the next field, and this is the field that I should work on, because this is going to be the hottest field of the next time window, and that is happening in at least hundred percent cases in a our data set.

(Refer Slide Time: 17:00)



CAUSE ANALYSIS

Concentration of High-Impact Papers -- in 82% cases this concentration declines near transition

Support from backup fields -- in 75% cases this support declines near transition

So, that is very interesting observation that we make. So, now, the next immediate question that we have is that; why do we see such a property, why certain fields have high authority value compare to other fields. Can there be same systematic analysis of what would be the ground reasons for a particular field to have a high authority score. So, for that we study two properties one; is the property of the high impact papers, and the other is the support from the back up fields. So, I will define each of them one by one.

So, what do we mean by high impact papers? So, we take, in every time window we take the papers which are top sited papers. So, that I have received the largest number of citations; that is the highest in degree papers, of that particular decade. we take the top ten percent of such papers. So, if you rank all the papers based on their citation counts, based on their in degree values, and take the top ten percent of these set, then that is what we called the high impact papers.

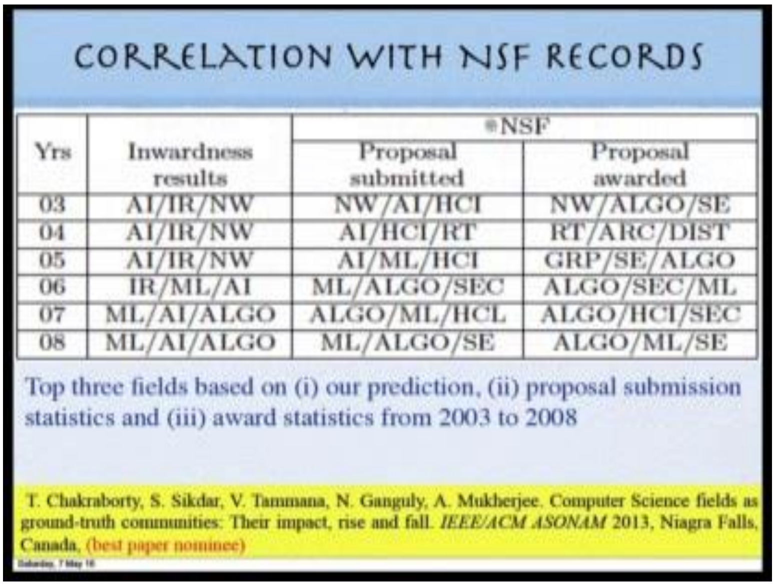
Now, we see that what is the proportion of high impact papers from a particular field, and as you see that in the first time window, where algorithms was at the tops, the largest fractions of, all the papers in the top ten percent is from the algorithm fields itself. You see the authority score of algorithms is highest, and also we see that the top ten percent highly sited papers in computer science, all of them are from algorithms area. Basically what I want to pass on, is that if you have a particular field which is high (Refer Time: 28: 33) and an authority score; that means, it has a large fraction of high impact papers.

Similarly in the next time window you see, that data bases has a large fraction of high impact papers, and these continues these observation continues, and this actually holds true for at least eighty two percent cases. similarly there is this concept of backup fields. Basically if you are in degree is high, then that in degree is actually derive from somebody else out degree. So, somebody else is citing you; that is why you are highly cited right. So, somebody else is actually citing you that are why you are highly cited.

So, if you assume these to be true, then those people who are citing you, to make you highly cited are called the backup fields. So, and we see that, if a particular authority field, is having a high authority score, then it is enjoying a lot of support from it is backup fields. So, there are many other fields that provide you with citations. So, there are a lot of backup fields, which support your citation basically. And what happens towards the transition, the support from the backup field gets reduced, that happens at actually in seventy five percent of the cases in our data.

Basically the backup fields actually re track their votes from you. Somehow, the interest of the papers in the backup field gets reduced in the papers from the high authority field. So, that is why they retracted their citation from the high authority field, and immediately the score of the high authority field falls. So, this is an observation that we see that holds true in seventy five percent of the cases.

(Refer Slide Time: 16:44)



CORRELATION WITH NSF RECORDS

Yrs	Inwardness results	NSF	
		Proposal submitted	Proposal awarded
03	AI/IR/NW	NW/AI/HCI	NW/ALGO/SE
04	AI/IR/NW	AI/HCI/RT	RT/ARC/DIST
05	AI/IR/NW	AI/ML/HCI	GRP/SE/ALGO
06	IR/ML/AI	ML/ALGO/SEC	ALGO/SEC/ML
07	ML/AI/ALGO	ALGO/ML/HCL	ALGO/HCI/SEC
08	ML/AI/ALGO	ML/ALGO/SE	ALGO/ML/SE

Top three fields based on (i) our prediction, (ii) proposal submission statistics and (iii) award statistics from 2003 to 2008

T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, A. Mukherjee. Computer Science fields as ground-truth communities: Their impact, rise and fall. *IEEE/ACM ASONAM* 2013, Niagra Falls, Canada, (best paper nominee)

Saturday, 7 May 16

So, we also further did some more interesting analysis. So, I would now introduce to you the concept of proposal submission. So, national science foundation is a body which actually keeps funding for doing research. So, it gives founding to different academic institutes in the United States for doing research. Similar body in India is the Department Of Science And Technology or Department Of Information Technology. they give you funds in order to do meaning full research.

So, now, if we do an analysis of the areas in which proposals were submitted in the different years from 2003 to 2008, we see that in 2003, the top three areas in which proposals were submitted, in the national science foundation, top three areas in which research proposals were submitted by different academic institutes, are networking, artificial intelligence, and human computer interfaces.

Now, among these if you see in 2003, according to our authority scores. The score that we define in the last slide theased on that authority score the top three ranks fields are a i information retrieval and networks. you see two out of these three fields are also present here. So, this shows that there is a there is really lot of interest in these areas at that particular point in time. So, actually the proposal submission data is very much similar to the data that is ranked by the authority score, and that is true for all the years. the proposal awarded means where money flows in.

The proposal submitted is very much correlated, where as the proposal awarded depends on many other factors. So, there you mostly see that there is only one overlap. So, it is not only based on the volume of proposals submitted, but the merit of the proposal, and there are many other things that actually goes in to, which proposals will be finally, funded. So, that is why there you do not have a very strong correlation, with the authority scores, but in terms of the interest, which is shown by the number of proposals submitted, you see that there is a high overlap. So, this is another interesting observation that one can make.

(Refer Slide Time: 19:12)

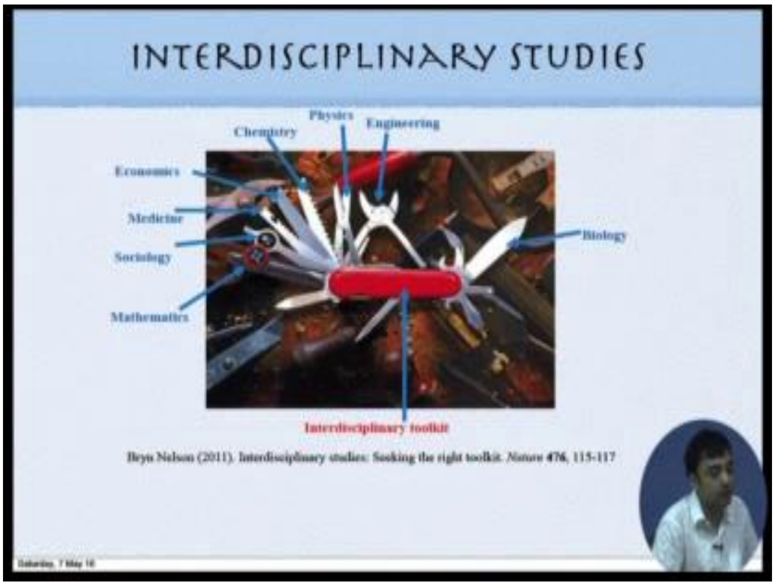
HOW TO SEEK THE RIGHT TOOLKIT

Nature **476**, 115-117(2011), doi:10.1038/nj7358-115a

Saturday, 7 May 16

So, the next concept that we will discuss, is nicely introduced in this particular figure. So, what you see in this picture, this is a very interesting picture. You see that there is one particular tool, which has many others small tools included in it, in the same framework, whereas there are other tools that are separate in themselves. So, this actually, is very nicely correlated to one of the rising or leading concepts in various scientific disciplines, which is called interdisciplinary.

(Refer Slide Time: 19:53)



INTERDISCIPLINARY STUDIES

Bryn Nelson (2011). Interdisciplinary studies: Seeking the right toolkit. *Nature* 476, 115-117

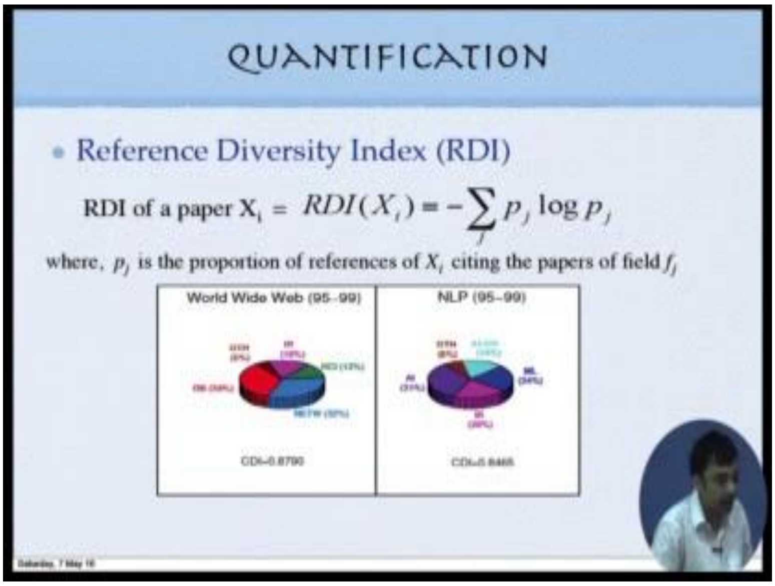
Saturday, 7 May 16

Basically the way people proposed to do research in current times, is through interdisciplinary practices. Basically the idea is that, now you look you cannot do well in research if you look into one single field. You have to like cross your boarders and look at various different disciplines, borrow concepts from various different disciplines.

Even while your discipline is computer science, you have to borrow concepts from say physics, from chemistry, from biology, in order to do meaning full and cutting edge research that is the mantra of interdisciplinary practices in research. So, and interdisciplinary is the talk of the current day, is the way that all the experts suggest; one should do research in an interdisciplinary framework. So, that is why studying interdisciplinary with in computer science is also very important.

So, the question is; like how one can actually describe or quantify mathematically the concept of interdisciplinary; that is what we will see in the next few slides. The basic idea is very clear, if you want to become a good researcher, you have to be an interdisciplinary researcher. You have to have knowledge in depth in one particular area, but you also have to have knowledge in breath in many other areas, in order to do successful research.

(Refer Slide Time: 21:25)



QUANTIFICATION

- Reference Diversity Index (RDI)

RDI of a paper X_i = $RDI(X_i) = -\sum_j p_j \log p_j$

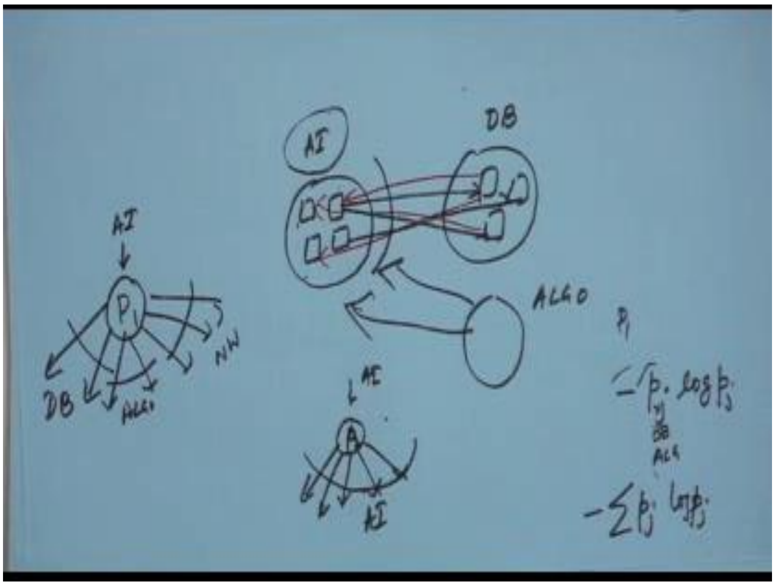
where, p_j is the proportion of references of X_i citing the papers of field f_j

Saturday, 7 May 16

So, the point is that, now the question that we try to ask is, whether we also see certain signals of interdisciplinary in computer sciences itself, and it turns out that we actually do see such signals in computer sciences also. So, in order to define the concept of interdisciplinary, we will bring in certain matrix into definition. So, the first metric that we will talk about, is the reference diversity index or r d i.

Basically what you try to measure here, is take a paper from a particular field, and you try to see what is the reference set of this paper, what are the other papers that this paper that this particular paper is referring to, and you see what is the fraction of references, that is going to a particular field.

(Refer Slide Time: 22:25)



Suppose there is a paper p 1, and it has say 5 6 references like this. Now the first two references say it goes to data bases. Say, this paper is from the field of a i. the first two goes to databases, the second two goes to algorithms, the third three goes to networks. Suppose you have a one paper like this. Whereas, there is another paper p 2, also from the field a i which again has some references, but all of them going to a i itself. So, immediately you see that there is a difference in the referencing style of the paper p 1 from the paper p 2.

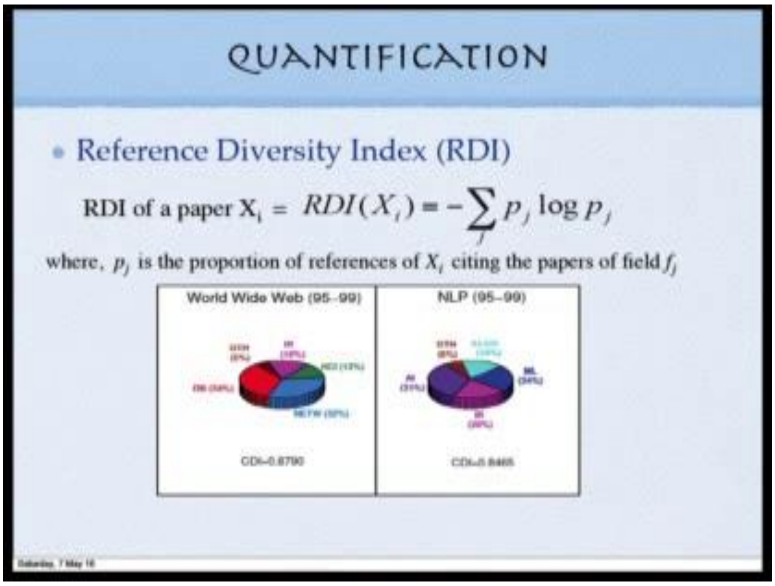
So, we can see that paper p 1 actually invokes a lot of references from different fields. So, there is a notion of interdisciplinary here. So, paper p 1 actually refers to various different fields, while writing the paper. Whereas, paper p 2 actually invokes all it is references from one signal field; and therefore, the chance is that it is not very highly interdisciplinary. So, the question is that whether one can actually quantify this particular concept. Now we try to quantify this concept in terms of all papers present in a field. You try to find out what is the fraction of references that are going for every individual paper.

What is the fraction of references going to each individual field? Now if you say that fraction, the fraction of references going to field j is p j, then you basically find out p j log p j, which is the entropy. You find out the entropy; that is the fraction of references that is going to a particular field j. say for the paper p one fraction references going to d b, fraction references going to algo and so on and so forth.

So, find this fractions you find out p i p j log p j, and then you do a sum of p j log p j which is nothing, but the entropy. Basically if this entropy value for a paper is high, then you know that this paper is actually making reference to various other fields. So; that means, the indication that it is more interdisciplinary, whereas, if this value is low, then this paper is actually not highly interdisciplinary.

Basically based on this these value of interdisciplinary, this value of entropy, you can have shallow notion of interdisciplinary. So, you find out the proportion of references that goes to each field, from a particular paper in a particular field, and then you find out the fraction, and then you express this fraction as an entropy, and if this entropy is high. Then you call this particular field to be interdisciplinary; whereas, if this value is low, then you call this particular field to be not highly interdisciplinary.

(Refer Slide Time: 25:57)



QUANTIFICATION

- Reference Diversity Index (RDI)

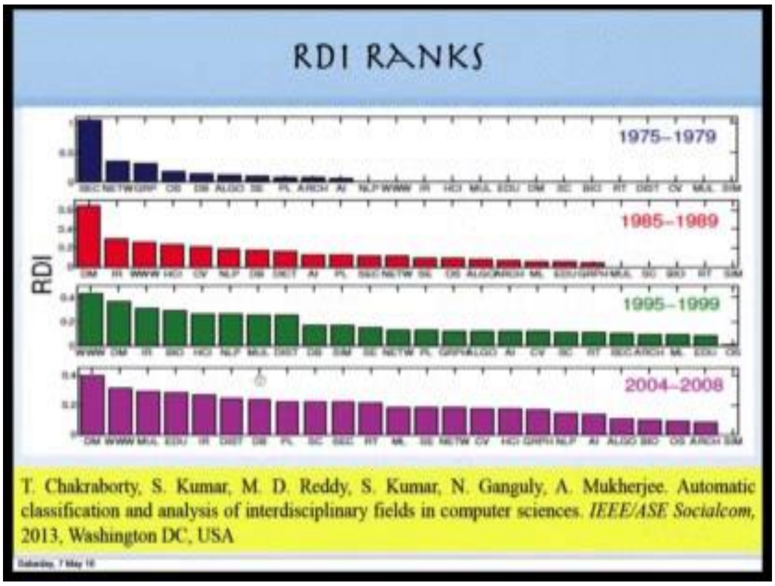
RDI of a paper X_i = $RDI(X_i) = -\sum_j p_j \log p_j$

where, p_j is the proportion of references of X_i citing the papers of field f_j

So, now if you take all papers from the World Wide Web, and you see what is the fraction of references; that is going to other different fields. So, World Wide Web is a field, which includes mostly graph theory, social networks, analysis from twitter face book and things like that. Now if you see the fraction of references that are going to other fields is, like - two information retrieval it is the total fraction of references is 16 percent, to human compute interaction it is 13 percent, network 32 percent and so on and so forth.

Similarly, now for the field natural language processing this is the distribution or the proportion of references from the papers of n l p, going to the papers of other fields. So, now, even there is definition of RDI you can very nicely calculate, from this proportions you can very nicely calculate the entropy values.

(Refer Slide Time: 26:54)



RDI RANKS

T. Chakraborty, S. Kumar, M. D. Reddy, S. Kumar, N. Ganguly, A. Mukherjee. Automatic classification and analysis of interdisciplinary fields in computer sciences. *IEEE/ASE Socialcom*, 2013, Washington DC, USA

Saturday, 7 May 16

Now, based on this entropy values you can basically rank the different fields. So, if you look at the window 1975 to 79, you see ranking like this. So the fields that are at the top of the rank are networks OS etcetera. Whereas, as you go on over the time line and you come to 2004 to 2008, some of the fields that you see are determining multimedia World Wide Web and things like that. So, these actually are really interdisciplinary fields, because if we look at the World Wide Web, or the determining community.

These communities actually do lot of references to many other different communities;

whereas, communities like operating systems architectures. They do most of their references; most of the papers refer to the papers of operating systems, and of architecture only. Hardly, they would refer paper to papers from other fields. So, that is why they come at the bottom of the rank list. Whereas, fields like data mining and World Wide Web, come at the top of the rank list.

In the later lectures, we will look at into some other ways of quantifying interdisciplinary, and then we will see a complete framework by which we can find out what fields in computer science are really interdisciplinary, and what fields are not interdisciplinary more sort of a code. So, this is what we will do in the next lecture.

Thank you.