Escuela Técnica Superior de Ingeniería Informática
UNIVERSIDAD DE MÁLAGA

June 2006

The *h* index, the quest for an optimum metric

http://neo.lcc.uma.es

One of the research lines of the GISUM group

Enrique Alba
University of Málaga, Spain

Index

Introduction

Optimize

- The search for an objective metric to quantify the cumulative impact and relevance of an individual's scientific research output

Subject to

- Should be easy to understand and measure
- Should be applicable to different domains in science
- Should measure the impact of the research
- Should avoid inflating distortion by 'big hits'
- Should be applicable to scientists of different ages
- Should be useful for evaluation and comparison (recruitement, salary complements, grants, awards, etc.)

Quantifying the Knowledge

General thoughts

- The publication data of an individual is the most usual source of information
- But it is not the only one: the citation record of a researcher is also very important

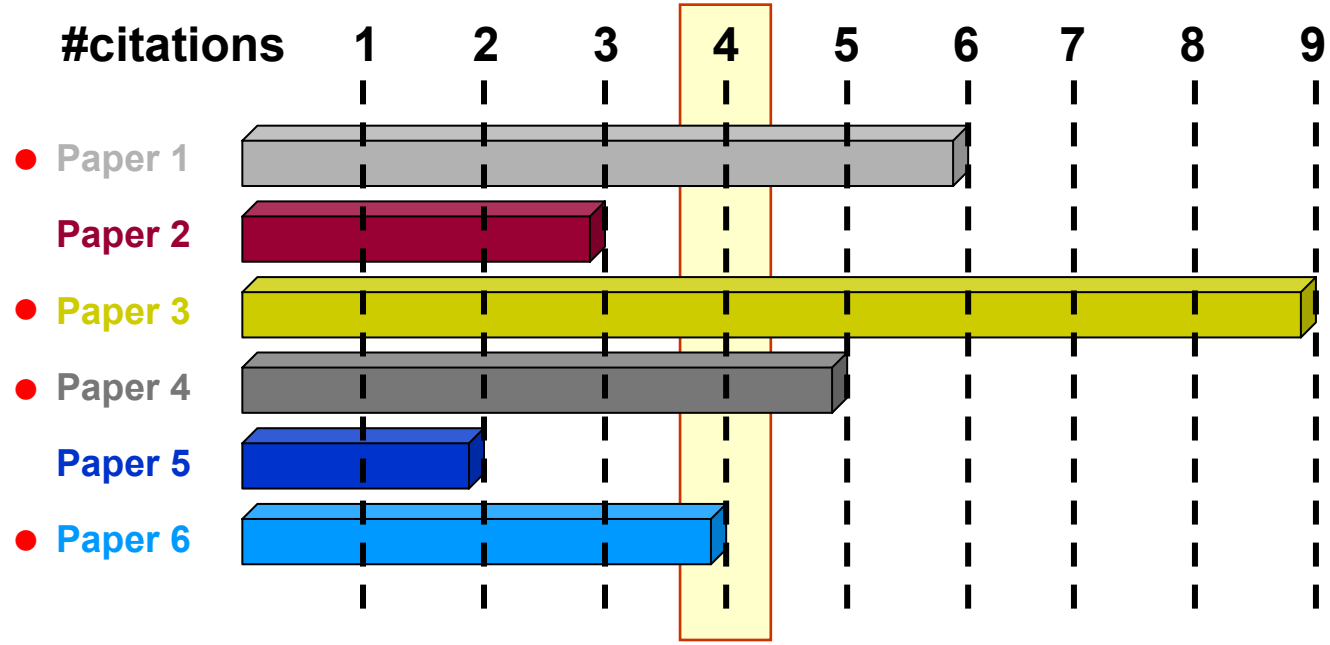
Thus we can quantify relevant information

- N_p : Number of papers published over *n* years
- N_c^j : Number of citations for each paper *j*
- Journals where the papers were published
- Impact factor of a journal (ISI or others)
- Others (trend of a journal, domain rankings, etc.)

The *h* index

Definition of the *h* index

A scientist has index *h* if *h* of his/her N_p papers have at least *h* citations each, and the rest of (N_p-*h*) papers have no more than *h* citations each



Advantages and Disadvantages of the *h* index

Why *h* is preferable to other single-number criteria:

1. Total number of papers (N_p)

PROs: measures productivity

CONs: does not measure importance nor impact of papers

2. Total number of citations ($N_{c,tot}$)

PROs: measures total impact

CONs: hard to find, inflated by 'big hits' or by many coauthors

3. Average citations per paper ($N_{c,tot} / N_p$)

PROs: measures the average impact of publications

CONs: hard to find, rewards/penalize low/high productivity

4. Number of 'significant' papers (> *y* citations)

PROs: no problems like before, it gives an idea of broad impact

CONs: favor/disfavor people arbitrarily, needs tuning for ages

5. Number of citations to each of the *q* most cited

PROs: overcomes many of the precedent disadvantages

CONs: favor/disfavor people arbitrarily, not a single number

Practical Issues

Advantages of *h*-index:

1. Measures the broad impact of an individual's work
2. Avoids most of the mentioned disadvantages (not all of them)
3. Easy to find: rank papers by 'times cited' in Thomson ISI (only this?)
4. Gives a rough estimate of the total number of citations

Practical application:

A. Two individuals with similar *h* are comparable in terms of their overall scientific impact, even if their total number of papers and citations are very different

B. For two individuals (of the same scientific age) with similar number of total papers and of total citation count but with very different *h* values, the one with the higher *h* is likely to be the more accomplished scientist

Mathematical Facts

Direct Implications (I)

1. The total number of citations is $N_{c,tot} = a \cdot h^2$

- The lower bound is h^2, ignoring #citations: above/below h

$a \sim [(1+c/p)^2] / [2c/p]$, p papers earning c citations (per year)

- Parameter a ranges between 3 and 5 (in physics)

2. Assuming linear grow with time: $h \sim m \cdot n$

- For a not too small number of years n, $h = [c / (1+c/p)] \cdot n$
- Slope (m) is useful to compare scientists of different seniority

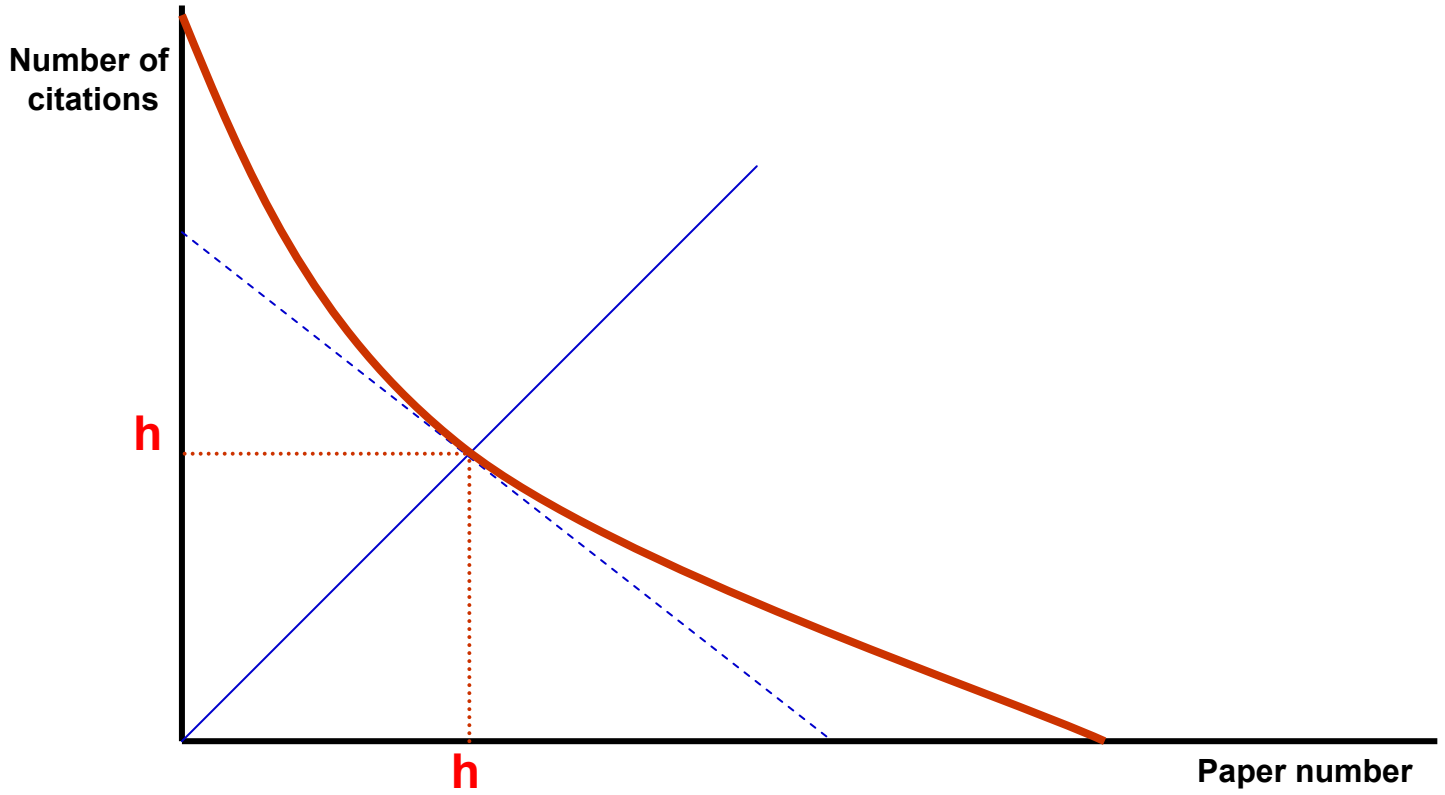
3. The minimum value for a is 2 (when $c=p$)

- Papers with more/less of h citations contribute eq. to $N_{c,tot}$
- The value of a will be larger for both $c>p$ and for $c<p$

4. Most (all) contributions come from highly cited papers when $c>p$

- Note: for $c<p$ *the* (N_p-h) papers are giving the largest contribution to $N_{c,tot}$

Mathematical Facts



The intersection of the 45 degree line with the curve giving the number of citations versus the paper number gives *h*. The total number of citations is the area under the curve. Assuming the second derivative is non-negative everywhere, the minimum area is given by the distribution indicated by the dotted line, yielding *a*=2 (being the slope = -1)

Mathematical Facts

Direct Implications (II)

5. In case a researcher stops publishing

- The linear model makes *h* to grow constantly during *h/c* years and then stays constant, since now all published papers contr.
- In a more reallistic model *h* will level off smoothly

6. Not all papers contribute to *h*

- The ones with fewer than *h* citations won't contribute, like those written late in a career or in not well known media

7. Papers earn citations over a limited period

- Papers contributing to a career are different in diff. moments

8. The *h* index cannot decrease with time

- Papers with *h* citations are in risk of not counting more
- 'Sleeping beauties' are papers that drop and come back again

General Usage

Practical examples ($m = \frac{c}{1 + c/p}$)

1. *m*~1

i.e. *h*=20 after *n*=20 years characterizes a successful scientist

2. *m*~2

i.e. *h*=40 after *n*=20 years characterizes outstanding scientists, likely to be found only at top universities or research labs

3. *m*~3 or higher

i.e. *h*=60 after *n*=20 years, or *h*=90 after *n*=30 years characterizes truly unique individuals

The *m* parameter ceases its usefulness if productivity severely drops, but the *h*-parameter is cumulative and may continue to increase even after a stop

Examples in Physics

Suggestions in physics (assuming *m*~1):

1. *h*~10..12

Might be typical values for tenure (associate professor)

2. *h*~15..20

Might be used to fellowship in the American Physical Society

3. *h*~18

Might be used to promote to full professor

4. *h*~45

Might be used to enter the US National Academy of Sciences

There can always be exceptions to rules, especially in life-changing decisions such as granting or denying of tenure (or civil servants in research)

Fields, Self-citations and Groups

Typical *h*-values will be different in different fields:

- Average number of references in a paper
- Average number of papers per scientist in the field
- Number of scientists in the field

A high *h* means high accomplishment, but the converse is not necessarily always true

- What if a few seminal papers with very large number of citations
- What if papers with many coauthors provoke a large *h* index

Self citations should be eliminated

- Their effect in the *h* index is small (irrelevant for papers below *h*)
- To correct, a paper with *h*+*n* citations and more than *n* self citations should be dropped from the *h* count (*h*--)

The *h* index of a group is larger than individual's ones but smaller than the sum of individual's *h* indices

Hands on the *h* Index

The are are many sources for consulting citations:

- ISI Web of knowledge: http://portal.isiknowledge.com/
- SCOPUS citation tracker: http://www.scopus.com/
- Citeseer archives: http://citeseer.ist.psu.edu/
- IEEEXplore crossref search (Google): http://ieeexplore.ieee.org
- Science direct (SCOPUS): http://www.sciencedirect.com/
- Google scholar: http://scholar.google.com/
- DBLP CS biblio. : http://www.informatik.uni-trier.de/~ley/db/
- ACM digital library citings (Google): http://portal.acm.org

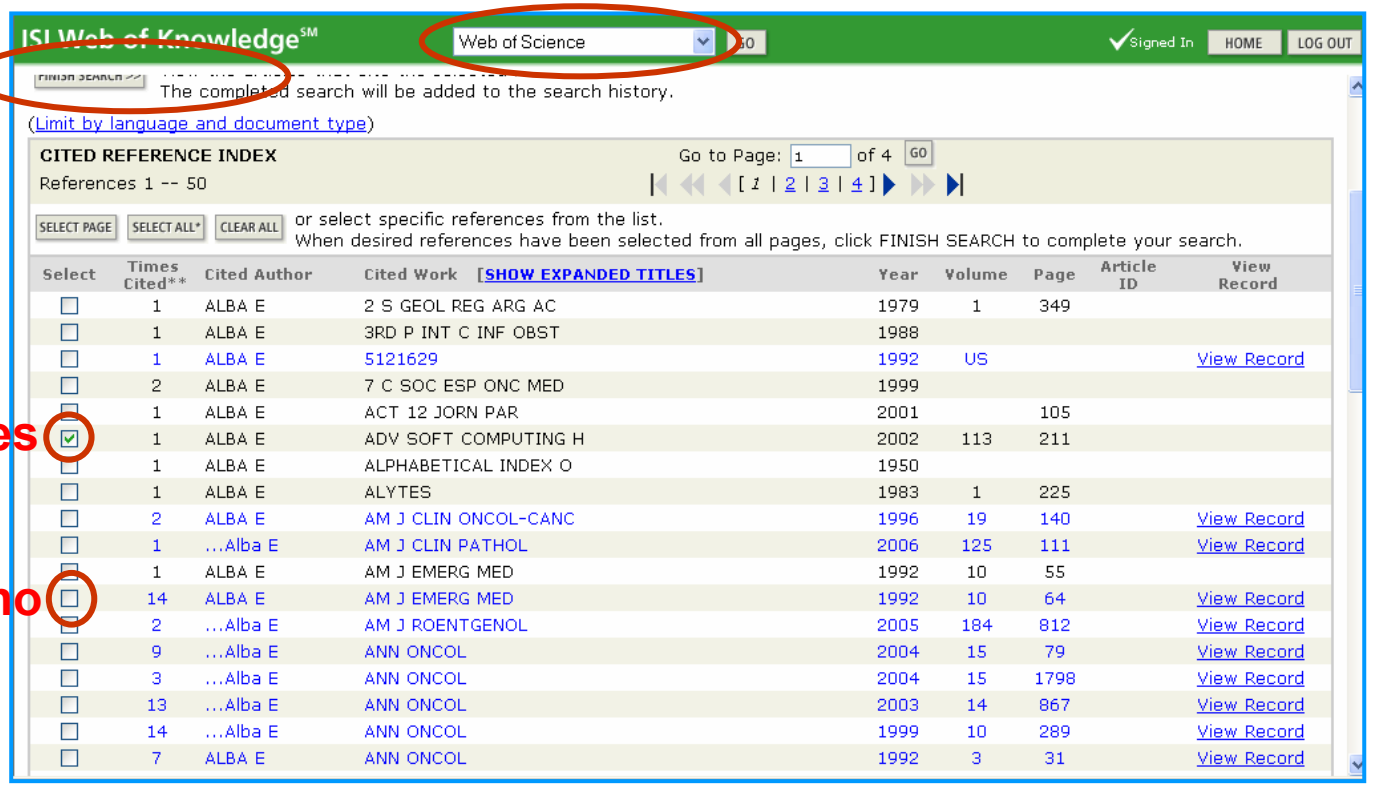
Search tips

- Only ISI and SCOPUS provide some tools for computing the *h* index
- The rest of sources must be seek paper by paper, tracking the cites (!)
- No automatic tool exists to exclude self references (filtered at hand)
- An automatic bibref tool should help, but just a few people do this.
- Search in Google for the exact article names to find new references

Hands on the *h* Index: ISI

Using ISI (I)

- Go to ISI Web of knowledge and select “Web of Science”
- Select “Cited Ref Search”
- Enter your name in the search box (e.g. `alba e*`)
- Select your articles page by page and then click on “Finish Search>>”

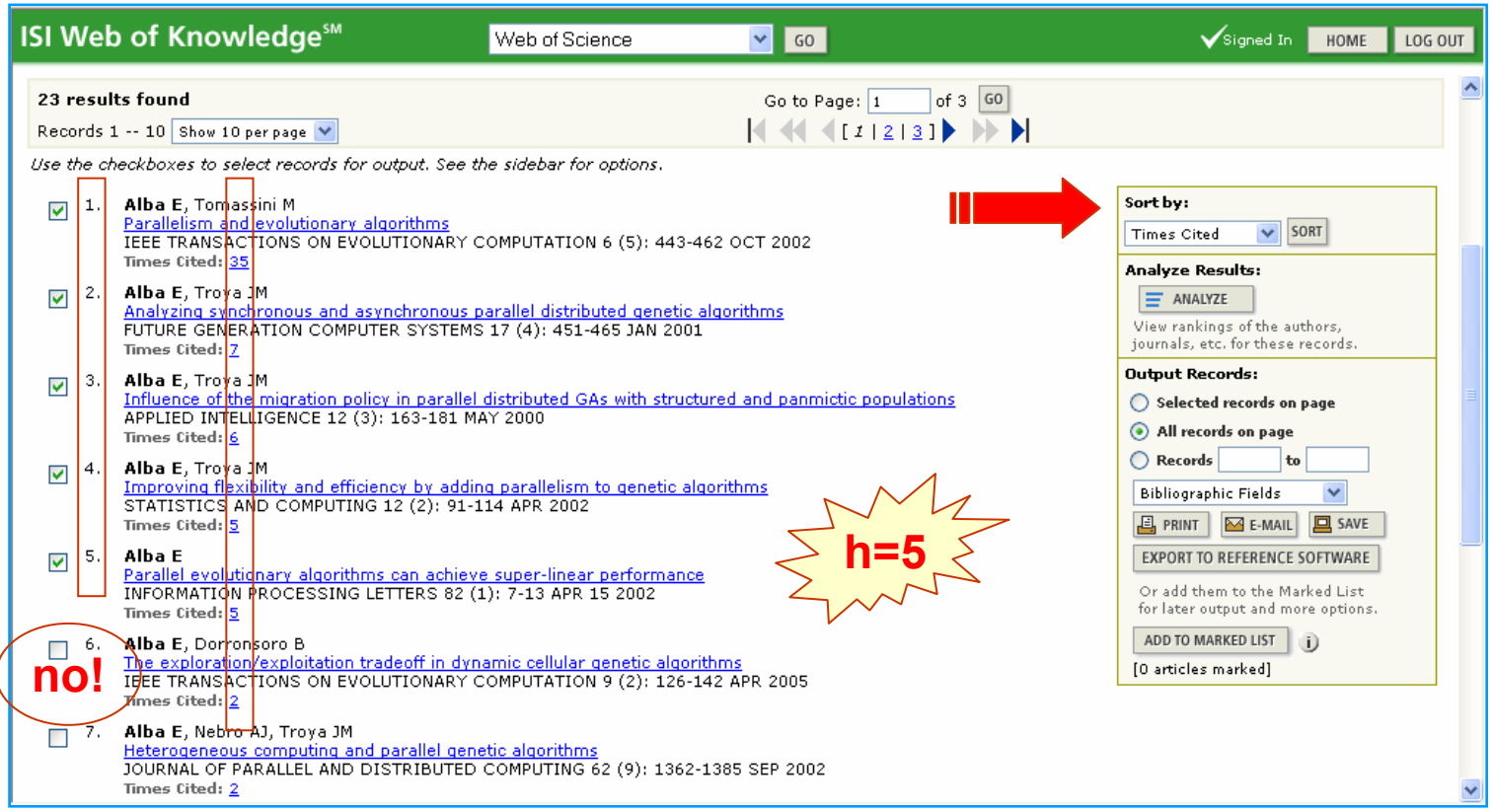


Select	Times Cited**	Cited Author	Cited Work [SHOW EXPANDED TITLES]	Year	Volume	Page	Article ID	View Record
☐	1	ALBA E	2 S GEOL REG ARG AC	1979	1	349		
☐	1	ALBA E	3RD P INT C INF OBST	1988				
☐	1	ALBA E	5121629	1992	US			View Record
☐	2	ALBA E	7 C SOC ESP ONC MED	1999				
☐	1	ALBA E	ACT 12 JORN PAR	2001		105		
☑	1	ALBA E	ADV SOFT COMPUTING H	2002	113	211		
☐	1	ALBA E	ALPHABETICAL INDEX O	1950				
☐	1	ALBA E	ALYTES	1983	1	225		
☐	2	ALBA E	AM J CLIN ONCOL-CANC	1996	19	140		View Record
☐	1	...Alba E	AM J CLIN PATHOL	2006	125	111		View Record
☐	1	ALBA E	AM J EMERG MED	1992	10	55		
☐	14	ALBA E	AM J EMERG MED	1992	10	64		View Record
☐	2	...Alba E	AM J ROENTGENOL	2005	184	812		View Record
☐	9	...Alba E	ANN ONCOL	2004	15	79		View Record
☐	3	...Alba E	ANN ONCOL	2004	15	1798		View Record
☐	13	...Alba E	ANN ONCOL	2003	14	867		View Record
☐	14	...Alba E	ANN ONCOL	1999	10	289		View Record
☐	7	ALBA E	ANN ONCOL	1992	3	31		View Record

Hands on the *h* Index: ISI

Using ISI (II)

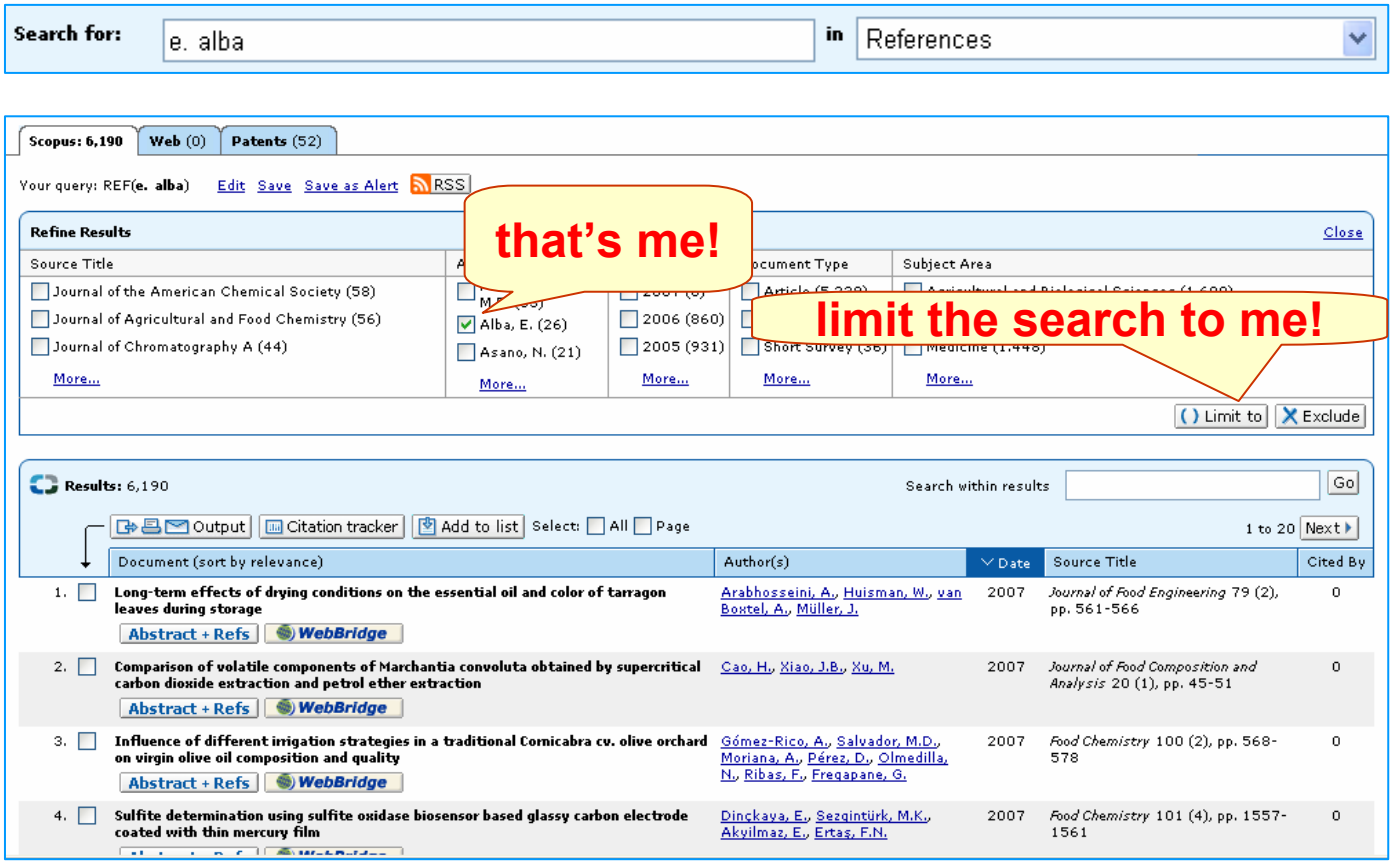
- Select all your the articles (at a time or filter to ensure that it's you)
- On the right hand menu select "Sort by: Times Cited"... and good luck!



Hands on the *h* Index: SCOPUS

Using SCOPUS (I)

- Go to http://www.scopus.com
- Search for your name and be sure to search "in References"

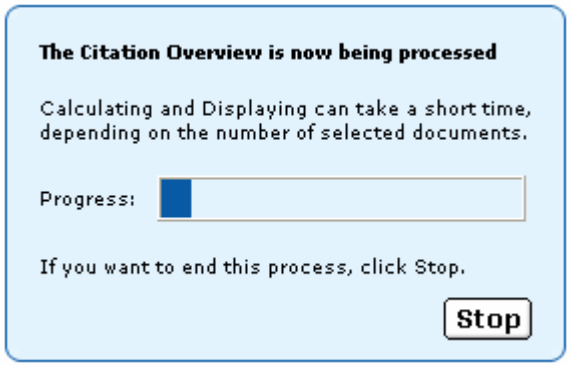


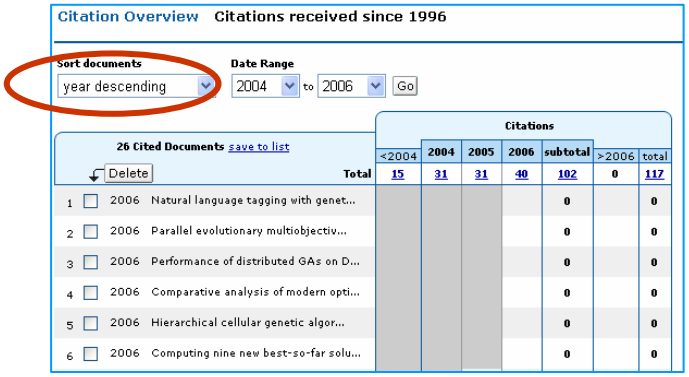
Hands on the *h* Index: SCOPUS

Using SCOPUS (II)

- Be sure no one else is in the list and click on "Select All"

- Click on "Citation Tracker", wait for a while… and that's it! (or not!?)

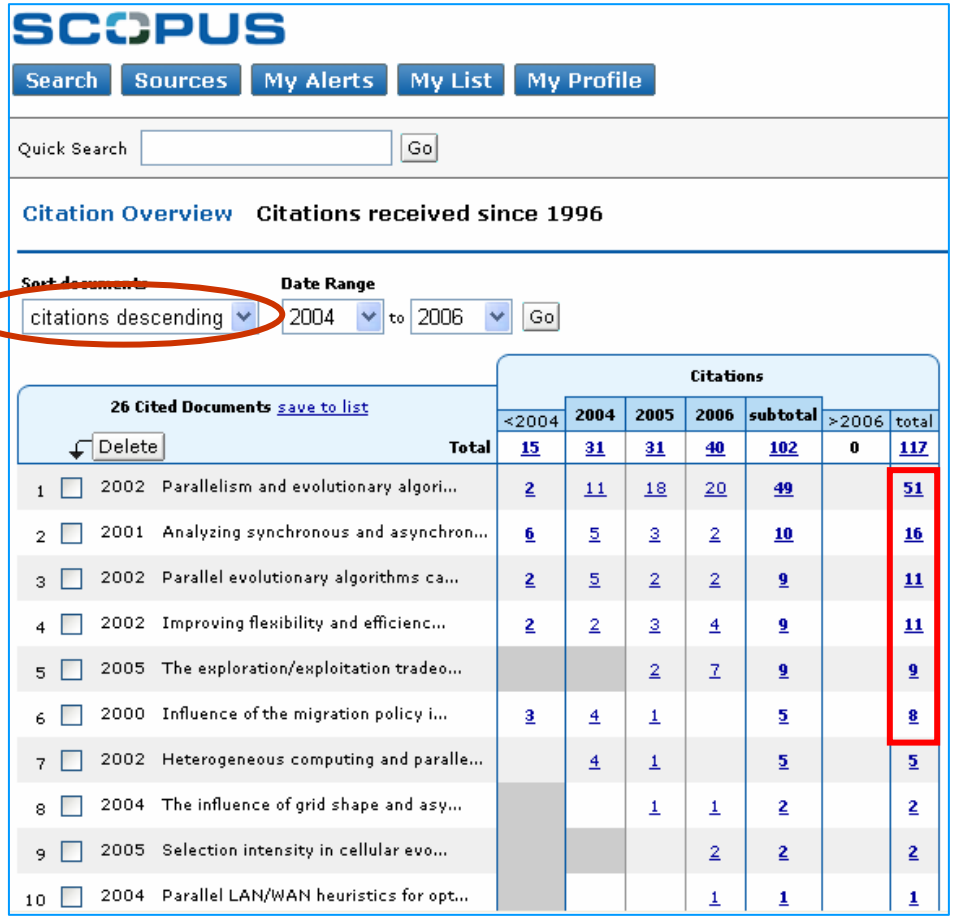




Hands on the *h* Index: SCOPUS

Using SCOPUS (III)

- Just an important final detail left: sort by “citations descending”!



h=6

Index

Final Comments

- **Computing the *h* index is interesting and indicative**
- **The standards of *h* index in computer science have to be still defined (average/minimum values …)**
- **Tools that help in computing it exist, but not directly**
- **Tools for searching citations are very important, but at this moment it's mainly an art**
- **Much has to be discussed on how to purge citations for individuals and for groups**
- **The big deal now is to define how to use *h* index**