Proposed method of facilitating fruitful cross discipline connections using numerical values mined from research papers and an ongoing feedback process to improve matching algorithms

Mark Waterman

Context and Overview

Historically, many significant scientific breakthroughs and technological leaps forward have come from cross-fertilisation of ideas in what were ostensibly previously unconnected areas of research.

The point in time in which one human being was able to keep abreast of latest developments in science (or what was previously called the arts) was probably passed five or six hundred years ago.

The cutting edge of scientific knowledge has grown from something analogous to a small ball with 12th century university study areas covering law, the arts (science), medicine and theology to a sphere perhaps several hundred thousand times as big today, as shown in figure 1.

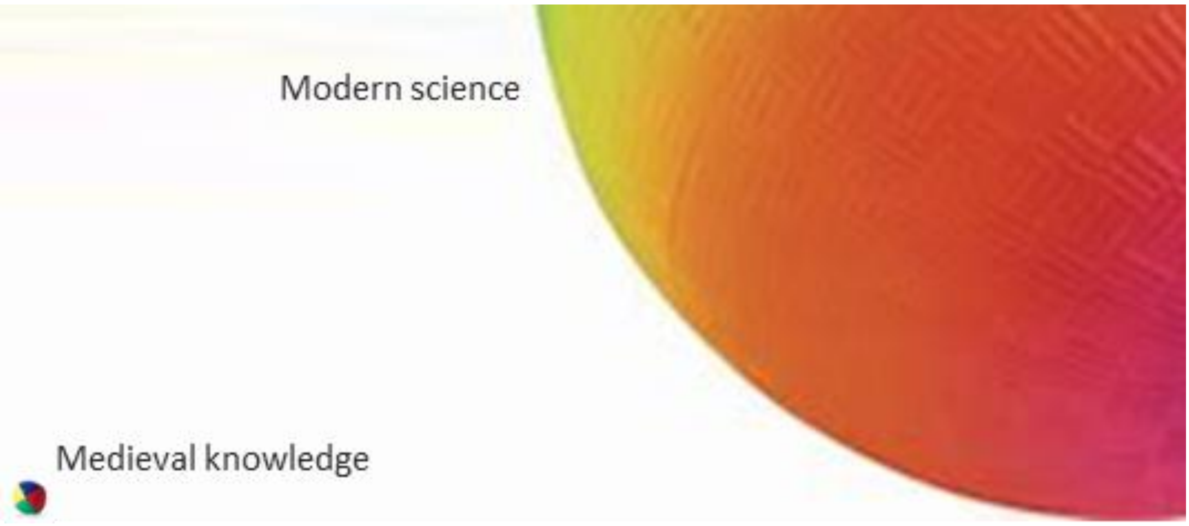


Figure 1. *Metaphorical representation of the extent of human scientific knowledge*

To further illustrate the point, and as a proxy for scientific knowledge, data on the annual number of patents granted is shown in figure 2. With an acknowledgement that although not all knowledge is necessarily cumulative (this point could be debated), a large amount certainly is and a very rough estimate might be, based on the cumulative number of patents granted, that the human species is <u>collectively</u> several hundred thousand times more educated than it was just two centuries ago.

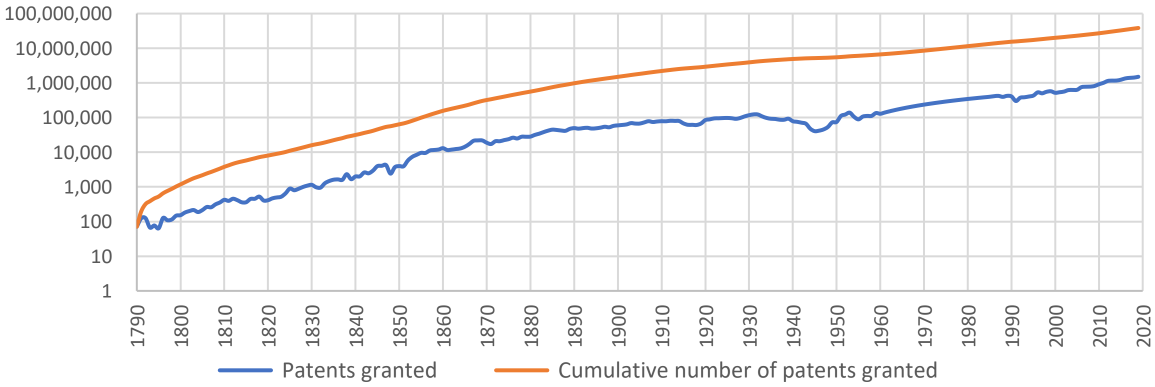


Figure 2. *Annual and cumulative global patents granted since 1790 (note, logarithmic scale)*

It might have been possible 600 years ago to gather together in one room the leading thinkers of the time and discuss in some depth what each was currently working on and perhaps share ideas on the applicability of research in one area to that in another.

Today, the sheer volume of different ongoing research fields and the depth of technical understanding required makes such an undertaking a practical impossibility. It is difficult enough to keep up with scientific advancements in one field let alone across the myriad of other unrelated disciplines.

There is certainly an enormous opportunity here in finding areas of combined benefits, the question is how to best facilitate what would only otherwise only happen by chance.

Many centres of learning make attempts to bridge this apparent impossible challenge by encouraging cross discipline sharing events, but the results will always be constrained by the numbers attending and their respective fields. There are nascent efforts underway to use AI technologies to mine research papers for meaning and by doing so connect together researchers in different fields such that complementary ideas can be shared. However, there are some non-trivial technical challenges in this undertaking, not least language, meaning and highly specific technical terms often used.

The alternative approach outlined in this paper is far less technically ambitious (although probably ultimately less useful than a fully AI methodology), but it is thought to be faster, easier and therefore cheaper to implement and rapidly scale and ultimately be net positive with respect to making useful connections that would otherwise be unlikely to occur.

Rather than try and glean meaning from text, research papers are mined for numerical values and their associated units of measure and these patterns are then used as a method of comparison.

The process is language, subject and context agnostic; a paper can be written in any language, about any subject, however esoteric and still be mined for data, providing numerical values and units of measure can be identified.

It is hypothesised that similar dimensional values might be useful to researchers in different topics for at least three potential reasons:

1. Sharing methods of measurement

The progression of science has gone hand in hand with the progression of technology capable of more accurate measurement of various physical properties.

The capability to repeatably and with high confidence, measure experimental outcomes against hypothesised values is core to the scientific method.

Where measurements are taken at the extreme end of currently available measuring technology, the techniques used may well be novel and knowledge of the processes, experimental apparatus and set up useful to other researchers working in the same units and scale.

2. Sharing algorithms or calculations associated with specific combinations of dimensions and scale

It is possible that highly bespoke mathematics underpinning research into somewhat obscure topics have cross discipline applications where similar variables and scales are involved.

3. Direct applications in different disciplines

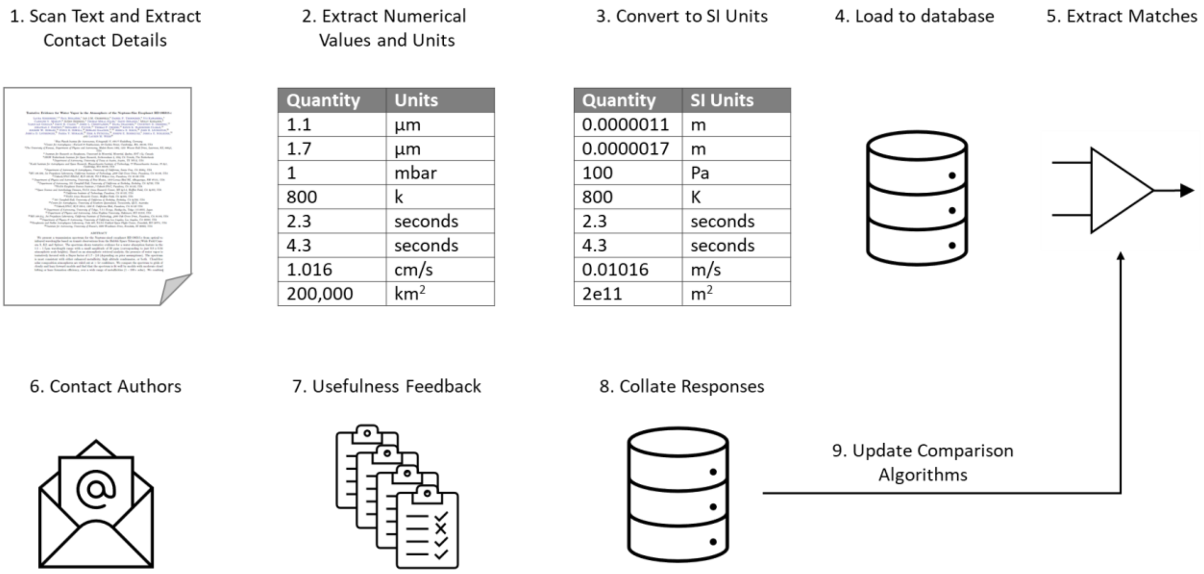
Potentially, there may be direct cross field applications that might otherwise only be spotted by chance leading to the development of new products or services.

For example (not based on any known connections), research into nanobots might be applicable to certain biological processes that have the same underlying scale of movement or forces involved.

Similarly, the same nanobot technology might (hypothetically) lend itself to the production of paints that reflect particular wavelengths of light or fabrics that display certain mechanical properties, etc. etc.

Outline Methodology

The process described in figure 3 shows the high-level steps used to generate connections between research papers based upon numerical data and associated units of measure. No context or knowledge or the subject is required.



Quantity	Units
1.1	µm
1.7	µm
1	mbar
800	k
2.3	seconds
4.3	seconds
1.016	cm/s
200,000	km^2

Quantity	SI Units
0.0000011	m
0.0000017	m
100	Pa
800	K
2.3	seconds
4.3	seconds
0.01016	m/s
2e11	m^2

Figure 3. *Flow chart overview of variable capture, matching, feedback and learning process*

1. Scan Text and Extract Contact Details

Tagging information is extracted from research papers, including the research institution, the lead and contributing authors and contacts email addresses. This information is used later in the process to connect authors and illicit feedback.

2. Extract Numerical Values and Units

Using relatively simple sampling code, values and units of measure are extracted from research papers. In the example of selected text from various research papers shown in Figure 4, the units of measure are shown in red and the corresponding values in blue.

> *…. Subsequent exposures used the G141 grism, which covers the wavelength range from* ***1.1 – 1.7µm*** *….*
>
> *…. The solar water abundance corresponds to the chemical equilibrium water volume mixing ratio for a solar composition gas at* ***1 mbar*** *pressure and* ***800 K*** *….*
>
> *…. During the testing, from* ***2.3 - 4.3 seconds****, the springs and foam pads exert a maximum combined force on the rail as the proximal link begins to side off at* ***4.295 N****, (see Figure 45) for max force ….*
>
> *…. This zero-g climbing approach is slow (****1.016 cm/s****) and limited in its capability due to the need of specialized handrails for the Robonaut hands to grasp and the limited number of handrails available around the exterior of the space station ….*
>
> *…. In addition, well-excavated contexts belonging to the fourth and early/mid third millennia are extremely rare across the whole central Anatolia (an area covering some* ***200,000 km***^{2}*) ….*

Figure 4. *Example quantities and units of measure "read" from text within research papers*

The values extracted are then initially stored in a tabular format as shown in Table 1.

Quantity	Units
1.1	µm
1.7	µm
1	mbar
800	k
2.3	seconds
4.3	seconds
1.016	cm/s
200,000	km^2

Table 1. *Example extracted values and units of measure*

3. Convert to SI Units

The extracted values are converted into a consistent format as shown in table 2.

Quantity	SI Units
0.0000011	m
0.0000017	m
100	Pa
800	K
2.3	seconds
4.3	seconds
0.01016	m/s
2.00E+11	m^2

Table 2. *Example extracted values converted into standard SI units*.

4. Load to database

The converted values are then loaded into a suitable database containing a full suite of potential physical characteristics as shown in appendix 1.

It should be noted that in any research paper there are likely to be multiple values for some units of measure and there is no inferred correlation between values across measurement unit aside from being linked to the same paper.

5. Generate Matches

The key determining factor in the efficacy of the process in terms of finding useful matches rests upon the methodology used to prioritise the relative importance of different combinations of measures and values.

To put this problem into context, there are at least 79 core units of measure, an indeterminate number of values for each and an infinite number of different alternatives to prioritising the relative weighting of the different combinations.

A pre machine learning approach to such a problem would involve coding complex algorithms based upon best assumptions or hypothesis of likely useful connections. The drawback of this methodology is that any feedback data would be inherently problematic to interpret and relate back to how the model needed to be changed for better results.

A machine learning approach has no such assumptions or hypothesis and operates on purely input pattern and output results basis.

Once the process is running feedback data can then be used to periodically to improve the outcomes by refining the machine learning matching algorithm.

There is however a necessary first step manually created algorithm in order to generate enough feedback to drive the first iteration of a machine learning algorithm. It is anticipated that this will probably only need to be a relatively simple weighting system based upon similarity of values and number of common units of measure, matches being defined in percentage terms.

All papers will be compared to all other papers and depending on an initial cut-off, which will be based upon the profile of results generated, the relevant authors will then be contacted.

The volume of comparisons is expressed by Metcalf's law used in network telecommunication systems to determine the total number of potential connections with n network nodes:

Total possible connections = $n(n-1)/2$

The number of comparisons grows as a proportion of the square of the number of nodes, or in this case the number of research papers.

Whilst the connection algorithm stays constant, new papers added require (n-1) comparison connections to be determined. If not in real time, and depending on the number of papers in the database, this can probably be easily handled on an overnight batch basis.

Every time the algorithm is updated, all of the comparisons (minus any where matches have been identified and emails sent) must be rerun.

Given that there are an estimated 75 million published research papers and c. 3 million papers (and growing) are published each year, at full scale this is probably a significant processing challenge.

6. Contact Authors

Using the contact details extracted from the research papers, emails can then be sent to the authors of both papers with short and simple messages outlining the following key points:

1. Objectives: to connect researchers across different disciplines by means of correlating similar units of measure and value ranges [being captured in the research] to potentially lead to sharing of measurement methodologies or other collaborative opportunities.
2. Match criteria: Units and value ranges with percentage match between papers
3. Summary: Summary text from matching papers
4. Feedback request: please take 2 minutes to rate the usefulness of this match to help us improve the matching algorithm

7. Usefulness Feedback

A simple online tick box scoring feedback mechanism to collect minimal data with the minimal of effort required is likely to garner the most responses. Key questions being:

1. Did you make contact? (y/n)
2. If yes, was it useful? (0-5) 0 = not useful, 5 = extremely useful
3. If no, why not? a. couldn't see the point b. the contact didn't respond c. the contact email failed d. time or other constraints
4. If the contact was useful, in what area? a. measurement process b. algorithms / calculations c. new application d. new line of potential research e. other (please specify)

8 & 9. Collate Responses and Update Comparison Algorithms

By collecting a large enough number of responses, it will be possible to use the dataset to train a machine learning algorithm to improve the efficacy of the matching. This process can be run either periodically or volume dependant.

Governance and funding

In an ideal world, the management, hosting and administration of the system would operate under non for profit conditions and in an open source collaborative environment, Wikipedia being the role model.

This should (and must) be an altruistic endeavour and not open to commercial exploitation or influence.

Potentially there might be a very small sustaining revenue stream build around commercial access to research papers with a pre-determined dimensional search criterion. Beyond that, ongoing funding needs to come from charitable sources

As AI technology progresses, a hybrid matching process utilising a combination of numeric and context based comparisons would seem to be inevitable.

Initial start up and pilot testing needs funding and technical resources, both are beyond the current abilities of the author.

It is hoped that the ideas outlined in this paper will be picked up and road tested by an institution or organisation that has a purely altruistic science based agenda.

Appendix 1, *Physical measure with standard SI units*

Physical Quantity	Symbol	Unit in SI	SI Unit Symbol
Acceleration	**a**	meter/second^2	m/s^2
Amount of substance	*n*	Mole	mol
Angular momentum	**L**	kg-m^2/s	
Angular speed	**ω**	rad s^{-1}	–
Area	***A***	square metre	m^2
Capacitance	C	Farad	F
Chemical potential	μ	Joule/ mol	J mol^{-1}
Current density	J	ampere/meter^2	A/m^2
Density, mass density	ρ	kg/cubic metre	kg/m^3
Dynamic viscosity		Pascal-second	Pa·s
Electric charge	Q	coulomb	C
Electric current	I	ampere	A
Electric potential	V	Volt	V
Electrical resistance	R	ohm	Ω
Energy	E	Joule	j
Entropy	S	Joule/Kelvin	J K^{-1}
Force	**F**	Newton	N
Frequency	*f*	Hertz	Hz
Heat	Q	Joul	J
Inductance	L	henry	H
Length	*l*	Meter	m
Luminous intensity	*Iv*	Candela	Cd
Magnetic field	**B**	tesla	tesla
Mass	M	Kilogram	kg
Momentum	**p**	kg-m/s	kg-m/s
Permeability	μ	henry/meter	H m^{-1}
Permittivity	ε	farad/meter	C^{2}N^{-1}m^{-2}
Plane angle	Θ	radian	rad
Power	P	watt or (Joule / second)	W
Pressure	P	pascal (N/m^2)	Pa
Refractive index	μ	unit less	–
Solid angle	Ω	steradian	sr
Specific heat capacity	c	J kg^{-1} K^{-1}	
Speed, velocity	***v***, *s*	meter/second	m/s
Surface tension	Y	N m^{-1} or J m^{-2}	
Temperature	T	kelvin	K
Time	*t*	Second	s
volume	*V*	Cubic meter	m^3
Wavelength	λ	Meter	m
Wavenumber	*k*	reciprocal metre	m^{-1}