Proposal for developing an XML system in scientific papers and articles, to capture the deductive structure of knowledge.

Ameet Sharma

January 30, 2017

ameet_n_sharma@hotmail.com

Abstract

We propose developing an XML-based system to enhance scientific papers and articles. A system whereby the premises of arguments are made explicit in XML tags. These tags provide a link between papers to more clearly exhibit deductive knowledge dependencies. The tags allow us to construct deductive networks which are a visual representation of deductive knowledge dependencies. A deductive network (DN) is a kind of bayesian network, but without probabilities.

1 Introduction

There is a great deal of literature with regards to representing knowledge in the form of bayesian networks in XML files. The problem is that these ideas involve coding the knowledge into some kind of specialized format. What we propose here is relatively simple modifications an author can make to his/her own paper by adding XML tags, which can let us extract the deductive structure of his/her arguments. Now these papers may be in journals, wikis etc; any place where the author would be willing to add these tags, and where the journals would be willing to accept papers with these tags. Note

that with the appropriate viewer/browser, the tags themselves could be made invisible only displaying the original article itself.

The structure of any argument is quite simple. Generally we have "premises", then we use some type of "reasoning or justification" to arrive at our "conclusions". Whether an argument is mathematical, political or literary, this does not change. We have an input-output black box like structure, with the input being the premises and the output being the conclusions. Now, our reasoning at times may be faulty, but even a poor argument is still an argument, and has this black-box structure.

Right now, all of humanity's knowledge is scattered in texts and journals. All this knowledge is interconnected, but the interconnections are far less transparent than they could be. Usually at the end of a paper, there's a bibliography referring to relevant papers. But we usually learn nothing about the nature of the connection between the two papers until they are read.

Making the premises of a paper explicit using tags lets us map out knowledge deductively from premises to conclusions.

2 An example

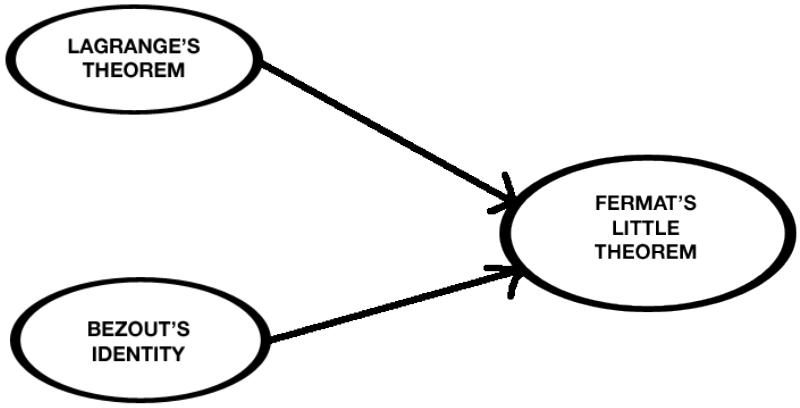
We'll take an example of an article from wikipedia and show how it could be enhanced by such tags. [1] is a proof of Fermat's Little Theorem using group theory labeled 'Proofs using group theory - Standard Proof', on Wikipedia. We'll ignore the fact that the article itself is made using HTML for now. We'll start every one of our tags with a 'd-' to distinguish them from standard HTML tags.

<d-knowledge>
<d-title>Fermat's Little Theorem</d-title>
<d-premise>Lagrange's theorem</d-premise>
<d-premise>Bezout's identity</d-premise>
<d-reasoning>The proof requires the most basic elements of group theory etc. (the text of the article goes in here) etc.
</d-reasoning>
</d-knowledge>

Suppose we call this a 'KNML' file. 'KN' for knowledge, 'ML' for markup language. Now from such a file a computer can easily produce (F.1). Let's call this figure a **Deductive Network** (DN). The DN uses the premises and the title tags but ignores the reasoning tag. The power of a DN is that it gives immediate information about the argument involved without the textual detail of the actual proof. If one doesn't have such tags, then extracting this type of structure would either involve a human reading the text, understanding it, and then producing the figure; or perhaps some extremely sophisticated language analysis software could process the text.

(F.1)

Figure 1:



What about just using wikipedia's own internal links in the actual article? Can those be used to recreate (F.1)? No, because those links may indeed be premises of the argument being presented, but they could also be consequences, or they can be entirely unrelated to the argument, but related topic-wise. Again, unless we have a human reading the text, or sophisticated language processing software, extracting the figure is impossible without our tags. The simple modification of adding XML tags leads to the ability to create a powerful structure.

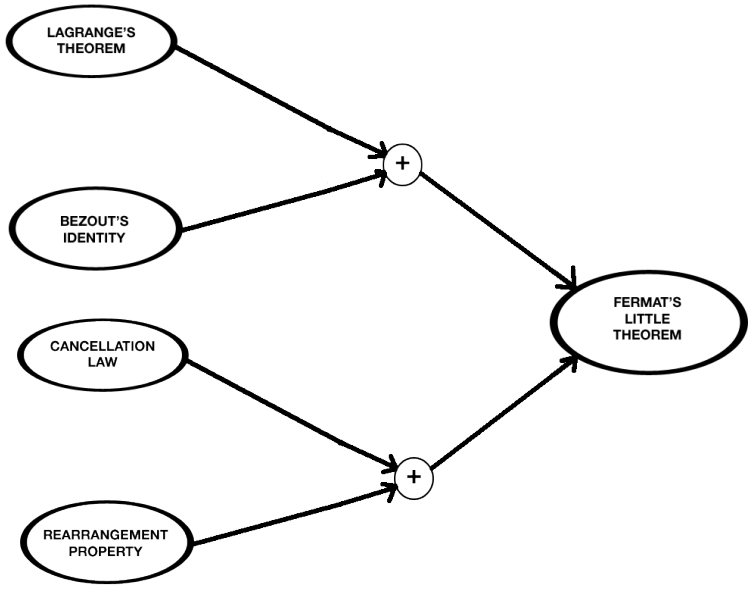
Now what if we want to add other proofs of Fermat's Little Theorem. For example "Proof using modular arithmetic" on the same wikipedia article. Either within the same KMNL file or another, we could add the following.

<d-knowledge>
<d-title>Fermat's Little Theorem</d-title>
<d-premise>Cancellation law</d-premise>
<d-premise>Rearrangement property</d-premise>
<d-reasoning>The proof, discovered by James Ivory, etc. (text of the article goes here).
</d-reasoning>
</d-knowledge>

The two d-knowledge tags together produce this:

(F.2)

Figure 2:



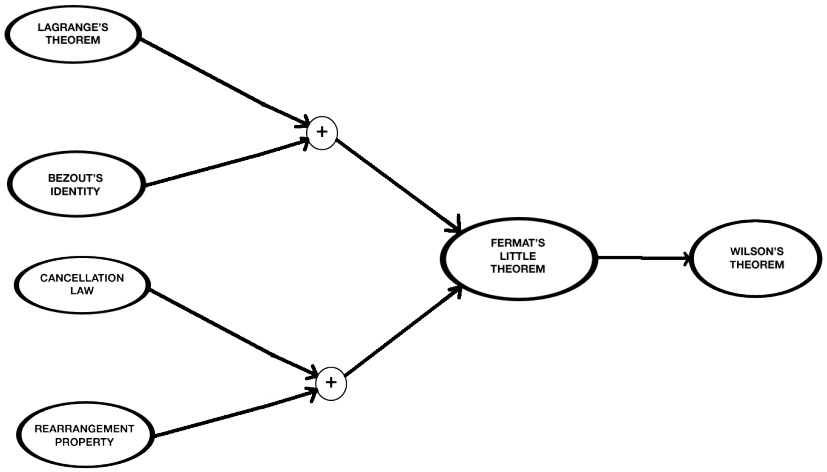
Fermat's Little Theorem can also be the premise for other theorems. For example, Wilson's Theorem. [2] is a wikipedia article that includes a proof of Wilson's Theorem using Fermat's Little Theorem. We can add this to our file:

<d-knowledge>
<d-title>Wilson's Theorem</d-title>
<d-premise>Fermat's Little Theorem</d-premise>
<d-reasoning>Again, the result is trivial for... (text of the article goes here).
</d-reasoning>
</d-knowledge>

This gives us the following image:

(F.3)

Figure 3:



Now "Cancellation Law" and "Rearrangement Property" are likely too common titles to be used as unique identifiers of knowledge. So we may have to use some type of numeric-id tag along with the title. This will depend on how large our DN is.

3 Level of granularity

The DNs we used in our example (Fermat's Little Theorem) are not "complete" in the sense of a pure logical deduction. We don't use logical laws as premises, or algebraic laws as premises. That level of detail would have been impractical. Furthermore we could probably have included more intermediate stages from our premises to our conclusion. So how do we decide on the level of granularity we want? There's no hard and fast answer to this. It depends on the application, the topic etc. We also want to keep it easy for the author.

4 Applications

4.1 Education

For someone engaged in interdisciplinary research, a DN can give an overview of a topic he's not familiar with, without reading entire articles. This is especially true if the researcher's interest is more theoretical than practical. Understanding theory means understanding the deductive structure of arguments. The DN perspective is different from a simple list of topics.

The self-learning student often looks to fill gaps in his/her knowledge. Tracing these gaps if often difficult when dealing with conventional text. A DN makes this much easier.

4.2 Research

As [5] describes, the interfaces between physics, chemistry and biology are ripe for research. And it's at these interfaces where DNs can be immensely helpful.

Take as an example photosynthesis. [4] describes evidence that quantum coherence plays a role in capture sunlight. It suggests that this process can explain the high-efficiency of converting sunlight to chemical energy.

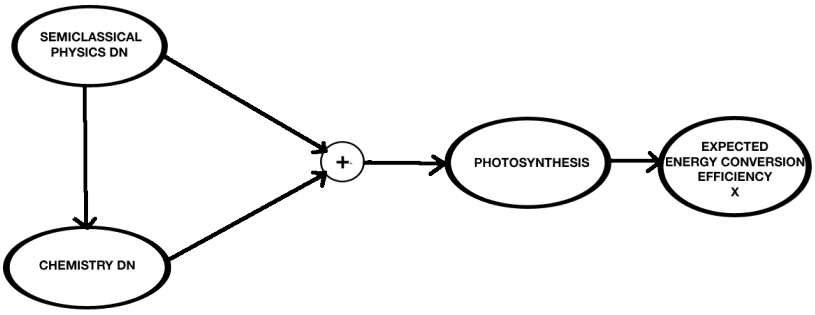
We'll use this example to show how hypothetically a DN might have assisted

in the discovery.

Here's how a related DN might have looked before the article:

(F.4)

Figure 4:

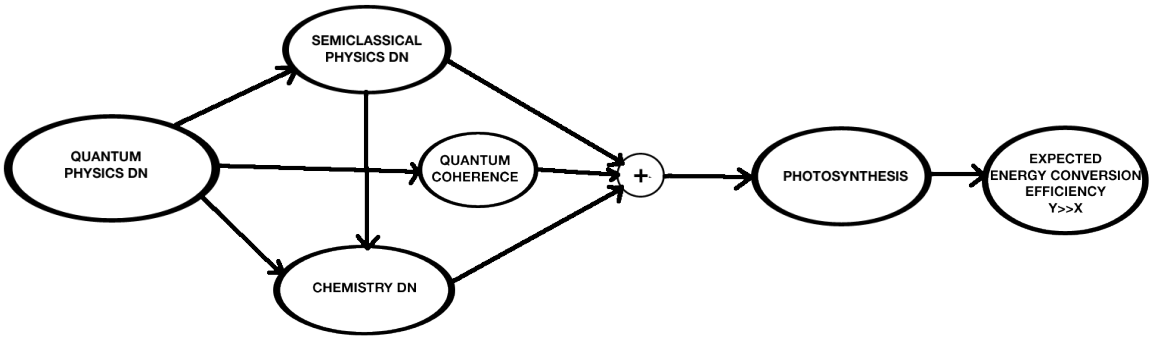


Now the real energy conversion efficiency (found by experiment) is much greater than the expected efficiency according to this network. So it's likely the DN is incorrect or incomplete.

So hypothetically the authors of [4] might have looked at the DN above and made a guess as to where changes would need to be made... possibly leading to the DN below:

(F.5)

Figure 5:



with the new energy conversion efficiency being closer to experiment.

Any consequences of this DN have to be checked against experiment. If the consequences of the DN don't match experiment, then the DN is incorrect or incomplete and changes have to be made again.

This is just a hypothetical situation to demonstrate how DNs could help. Having a visual representation makes finding gaps easier.

5 DNs as an intermediate stage for computerized knowledge

5.1 Mathematics - QED Manifesto

The QED Manifesto [3], was a proposal to formalize all mathematical knowledge in a computer system, with all the proofs checked by the computer. One of the reasons for the lack of success of this project is the difficulty of writing formal mathematics.

The XML system, along with deductive networks, provides an intermediate

stage between traditional mathematics and formal mathematics. The more granular the DN, the closer we get to formal mathematics. Going from the DN stage to formal mathematics will still be a challenge but less so.

5.2 Non-mathematical knowledge - bayesian networks

It wouldn't be difficult to go from our DN to a true bayesian network with probabilities. Once you've extracted the DN, it can be fine-tuned. Perhaps there could be probability tags that designate the bayesian probabilities. A bayesian network of this sort can let a computer calculate the probability the conclusion of some paper is true.

6 Will authors be willing to add these tags?

There might be resistance to this idea at first. But once authors start the process, other authors will see the benefits and be encouraged to participate. Putting in the XML tags would force the author to clarify his own reasoning. And a visual representation of the argument would make a journal reviewer's job much easier, likely reducing review times. Ultimately all scientists want their work to be understood as quickly as possible.

A DN will allow the author to see his/her own contribution to the field, and how it fits with existing literature (assuming it's all the related literature has been put into DN format). This is a a much more powerful measure of the influence a paper than "number of citations".

7 Public DNs

The power of this system becomes most apparent when DNs are available publicly. This allows us to interconnect DNs from different papers. But journals generally do not make articles public. The actual text of the paper goes in the <d-reasoning> tag of the DNML file. And these tags aren't used to form the DN. So perhaps the <d-reasoning> tags can be removed from the file, and the shortened file can be made public. Or perhaps the text in the d-reasoning tag can be replaced by something like this:

<d-reasoning> See Journal X, Article Y etc. (details or links to purchase the full article) </d-reasoning>

So this would allow the public to form the DN without seeing the textual material of the paper. Journals already make abstracts public so they may be open to this idea.

But in order to interconnect all these DNs, the numeric-id tag becomes necessary. We may require some type of central authority to assign unique ids for each knowledge tag.

8 Conclusion

This paper was intended to demonstrate the fruitfulness of inserting XML tags in papers for the purpose of generating deductive networks. The specific examples and tags used in this paper were only given as hypothetical possibilities. Developing the appropriate standardized tags will require further discussion.

References

[1] Wikipedia Contributors. Proofs of fermat's little theorem, 2016. URL `https://en.wikipedia.org/wiki/Proofs_of_Fermat%27s_little_theorem`.

[2] Wikipedia Contributors. Wilson's theorem. *Wikipedia, The Free Encyclopedia*, 2016. URL `https://en.wikipedia.org/wiki/Wilson's_theorem`.

[3] Wikipedia Contributors. Qed manifesto. *Wikipedia, The Free Encyclopedia*, 2016. URL `https://en.wikipedia.org/wiki/QED_manifesto`.

[4] Gregory S. Engel, Tessa R. Calhoun, Elizabeth L. Read, Tae-Kyu Ahn, Tomas Mancal, Yuan-Chung Cheng, Robert E. Blankenship, and Graham R. Fleming. Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature*, 446(7137):782–786, 04 2007. URL `http://dx.doi.org/10.1038/nature05678`.

[5] S. Succi P.V. Coveney, J.P. Boon. Bridging the gaps at the physics-chemistry-biology interface. *Philosophical Transactions of the Royal Society A - Mathematical, Physical and Engineering Sciences*, 2016. URL http://rsta.royalsocietypublishing.org/content/374/2080/20160335.