DOES YOUR RESEARCH HAVE AN IMPACT? HAS YOUR WORK BEEN CITED? WHY?

1 Why are paper citations important?

Citations are used to reference prior work in a field, support claims and acknowledge people for their contributions. Moreover, citations in research papers are also used to evaluate the level of research activity and contribution of a researcher; this is usually done by counting the number of times he/she has been cited. The count is usually incorporated into an evaluation model such as the H-index (see Figure 1) and the Impact Factor (IF).

Examples of verbs in citation sentences

- We **expand** on the work of John et al. (2011) …
- David and Li (2003) **introduced** a new technique for …

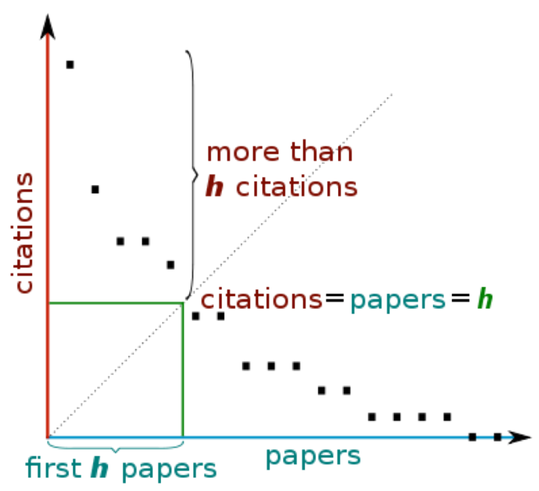
Figure 2: Verbs in citation sentences (highlighted in bold)

2 What is the problem?

One study examining citations found that 40% of citations are perfunctory (not essential to understanding the material presented in the paper) [1]. Another points out that "many references are cited out of politeness, policy or piety" [2]. These types of citations distort the real value of citation counts thus making models such as the IF and H-index not very accurate measures of research contributions.

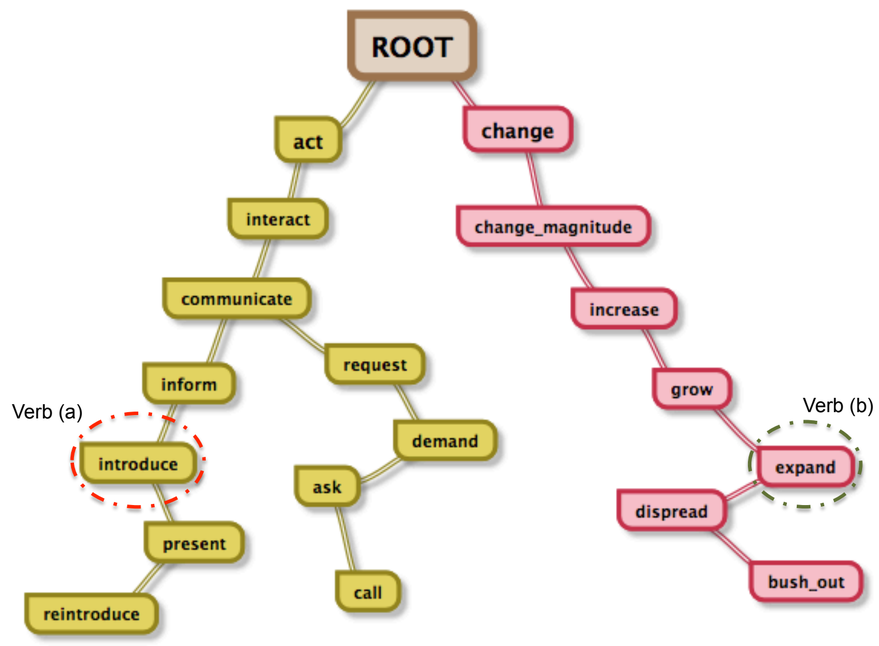
3 What are we going to do about it?

We differentiate between citations and categorise them into different categories depending on how the author of the paper used them. This is known as citation classification.
For example, citations referring to important work being built upon are grouped differently to citations being criticised.



A scientist has an index *h* if *h* of his or her *N* papers have at least *h* citations each and the other (*N* – *h*) papers have less than or equal to *h* citations each [3].

Figure 1: H-index graph



The Path similarity measure between two verbs a and b is calculated as follows:

$$\mathrm{PATH}(a, b) = \frac{1}{L(a,b) + 1}$$

Where L(a,b) is the shortest path (the least number of nodes) between a and b in the WordNet hierarchy.

For example, the Path similarity between the verbs "introduce" and "expand" is as follows: L(a,b) is 10. Therefore,

$$\mathrm{PATH}(a, b) = \frac{1}{10 + 1} = 0.090909$$

Figure 3: A WordNet hierarchy containing verbs "expand" and "introduce"

4 What have others done?

Previous work on citation classification involved classifying citations into predefined categories and it was mainly done in one of two ways:

- **Manual rules:** Experts analysed a number of citation sentences and created rules that define in which category will a citation fall. This method has not resulted in accurate classification. The rules are also domain-dependent, if the rules were created for biology citations, they will not work for computer science citations for example.
- **Supervised learning:** A computer algorithm is taught by example how to classify. This involves an expert classifying a set of citation sentences (called a training dataset) to the appropriate categories. The training dataset is then fed to the algorithm which will learn from the examples how to classify citations. This method is dependent on the accuracy of the training examples, and so the accuracy will only be as good as the training examples. The algorithm will not work well for citations it did not learn about or encounter previously.

5 What do we do?

We use an unsupervised learning technique called clustering to categorise the citation sentences into categories. Clustering is performed based on the similarity between the verbs inside the citation sentences. The overall result of the clustering procedure is a number of categories each containing citation sentences that are similar to each other.

We calculate the similarity between the verbs (see Figure 2) using the WordNet English lexical database [6] via three similarity measures (Path, Wu-Palmer [4] and Leacock-Chodorow [5]). Figure 3 shows an example for calculating the Path similarity between the verbs "introduce" and "expand".

The advantage of this technique is that we do not need to manually create rules or training samples to teach an algorithm and it is also domain independent.

6 Evaluation and Future Work

We compared the three similarity measures and the best one in terms of average inter/intra cluster distance was the Leacock-Chodorow measure. Overall, 12 valid categories emerged from the experiment. Each category contains verbs (each representing a citation sentence) that are similar to each other.
In the future, we will look into automatically labeling the resulting categories to indicate the type of citation sentences grouped within it.

7 Conclusion

Measures that evaluate the impact of research which rely on pure citation counts have drawbacks. Citation classification can address the drawbacks by categorising citations into categories based on the purpose or function of the citation. We used an unsupervised machine learning technique to categorise the citation sentences and compared 3 measures that were used during the categorisation. The Leacock-Chodorow measure was found to be the best with 12 valid categories resulting. Our technique overcomes some of the drawbacks of other techniques used to perform citation classification.

By: Mohammad Abdullatif
Supervisors: Gillian Dobbie, Yun Sing Koh
Department of Computer Science
The University of Auckland

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

References:
1. Moravcsik, M.J. & Murugesan, P. (1975). "Some results on the function and quality of citations. Social studies of science", 5, 86-92.
2. Ziman, J.M. (1968). "Public knowledge: an essay concerning the social dimension of science". Cambridge U.P.
3. Hirsch, J. E. (2005). "An index to quantify an individual's scientific research output". Proceedings of the National Academy of Sciences, 102(46), 16569-16572.
4. Wu, Z. & Palmer, M. ,"Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ser. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 133–138.
5. Leacock, C. & Chodorow, M. , "Combining local context and WordNet similarity for word sense identification". In C. Fellbaum (Ed.), MIT Press, 1998, pp. 305–332.
6. Fellbaum, C. , "WordNet: An Electronic Lexical Database, ser. Language, Speech, and Communication". MIT Press, 1998.