Citation Clustering for Identifying Research Contribution

Madhumita Satish Jadhav*, Jyoti Dhas, Deepali Joshi, Namrata Jadhavrao, Sayali Kadam
Vishwakarma Institute of Technology, Pune, Maharashtra, India.

* Corresponding author. Tel.: 9561501309; email: mitujadhav@gmail.com
Manuscript submitted April 20, 2015; accepted August 1, 2015.
doi: 10.17706/jcp.10.6.406-411

Abstract: The h-index is an index that measures productivity and citation impact of the published work but it has been criticized because it does not consider context of citation and reason behind citation. This indicates that there is a need for an improved h-index by a new approach which includes important citations received by a paper instead of the whole list of citations. Citation classification is an emerging area of research that categorizes citations based on the purpose behind the citation.

To perform citation classification there is need of a standard set of classes called as classification scheme. Such standard scheme is not available so we aim to generate a citation classification scheme automatically i.e. by using hierarchical clustering. The clustering is performed by using similarity vectors. The main contribution of this research is to generate similarity distance matrix of keywords and verbs extracted from the citation sentences with the help of WordNet.

Key words: Leacock-Chodorow similarity vector, WordNet, hierarchical clustering.

1. Introduction

The measurement of research impact is useful when a researcher is applying for promotion, requesting funding or being interviewed for a new position. Up till now this measurement has been performed with the help of h-index [1] but it has some limitations such as it relies on pure citation counts treating all citations as equal and ignoring the reasons behind the citations [2], [3]. As given in [4] 40% of the citations were found to be unconcerned to understand the paper. Thus, to overcome limitations of h-index there is need to introduce citation classification scheme. To perform citation classification, a set of classes is required, into which classification of citations is performed. Along with the classes, a classifier is needed to perform the classification.

This classification of clusters of citation sentences is based on following features:

1) In the citing paper position of the citation sentence.
2) Count of the citation sentence.
3) The syntactic and semantic structure of the citation sentence.
4) Words representing hint or indication in the sentence.

All of the above these features are described in literature for citation classification [5]. In this paper, clustering is mainly based on the similarity between the verbs of citation sentences and key-terms in paper. The purpose behind using verbs and key-terms is to denote the relationship between the citing work and cited work. The similarity between verbs and key-term is measured using the Leacock-Chodorow [6] similarity measure which is based on the graph structure of an WordNet which English lexical database [7].

Our main goal is to adapt clustering technique to generate set of classes from a dataset of citation sentences automatically. Classes generated automatically from the clustering process represent the purpose, reason behind the citation and actual manner in which papers are citing each other.

2. Citation Extraction

Citation scenario is depends on parent child relationship and it is specified as follows:

In Fig 1, paper *P* is the citing paper and *P*1 to *Pn* are cited papers. Set of all citations *S*1 to *Sn* are extracted from the paper *P*1 to *Pn* and then classify each citation *Si* to the appropriate class *Ci* from the set of classes *C*1 to *Cm* known as a classification scheme. Each class *Ci* from the classification scheme represents the purpose of the citation. We generate classification scheme from a large dataset containing citation sentences.

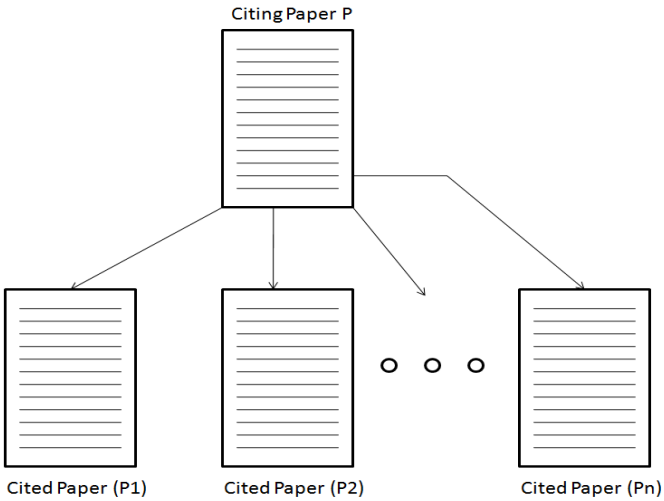

Fig. 1. Citation scenario.

3. Dataset Compilation

To have a valid set of classes, it is necessary to cluster a large dataset of citation sentence. We have considered citation sentence from 1000 papers of different fields. Below is a sample of 3 sentences from the dataset.

S1: Kwok and Yeung [a] introduced the important issue of retraining a neural network.

S2: Our paper expands the past work of NLP presented in [b].

S3: This compares data in two time windows [c], [d].

3.1. Verbs Extraction

In the context of citation classification, the purpose of the citation is indicated by verbs in citation sentences. Therefore, we cluster on verbs as the feature. Built in functions of WordNet is used to extract the verb from sentence. Verb is extracted in its base form. From the above sentences, the following verbs are extracted: V1: introduce V2: expand V3: compare. The next step is to calculate the similarity between the verbs.

3.2. Keywords Extraction

The reason of the citation is not accurately cleared only using verbs because same verbs are used in different contexts. Therefore, it is beneficiary to extract key-terms from the paper to understand citation purpose more precisely. For example, in sentence S1: Kwok and Yeung [a] introduced the important issue of retraining a neural network. The term neural network is extracted as key-term which clarifies the reason of

citation more accurately as compared to verb of the sentence which is introduce.

3.3. Calculation of Verb Similarity Vector

In this work, we generate a feature vector for each extracted verb and key-terms. A feature vector is an n-dimensional vector of numeric values. This vector represents the similarity between extracted verbs, key-terms and every other verbs, key-terms extracted from the citation sentences dataset. Similarity vector is calculated by Leacock-Chodorow similarity measure which is based on WordNet hierarchy which is lexical database of English.

WordNet groups words into synonym sets called synsets. Synsets are linked together with many semantic links. Synsets are organized in hierarchies where words become more abstract as you go up to the root node.

In this paper we used hierarchy of WordNet to calculate similarity between key-terms and verbs. Similarity value for any two key-terms or verbs can be generated using Leacock Chodorow similarity measure.

In Fig 2. similarity between verb (a) i.e. introduce and verb (b) i.e. expand is calculated by using Leacock Chodorow similarity measure by using following equation :

$$LCH_{(a,b)} = \max[-\log(\mathrm{Len}_{(a,b)} / 2 * \mathrm{Depth}_{(\max)})] \quad (1)$$

where, Len(a, b) is the shortest path of WordNet hierarchy of verbs "introduce" and "expand". $\mathrm{Depth}_{(\max)}$ is the maximum depth from the root to the deepest leaf verb. Similarity score between verbs—introduced and —expands is as follows:

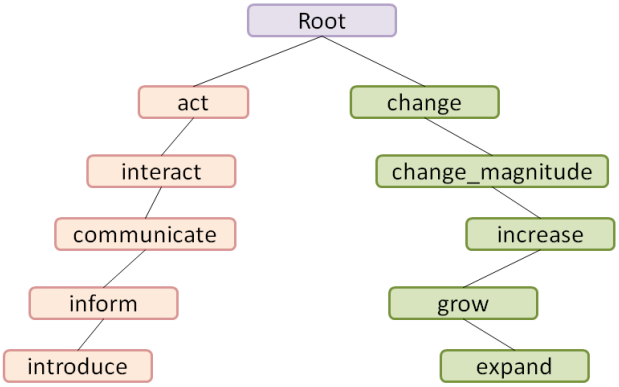

Fig 2. WordNet hierarchy for verbs “expand” and “introduce”.

Len(a, b) is 10 and here the maximum depth is 14 (the depth is depends on version of WordNet we have used WordNet version 3.0 whose depth is 14) Therefore, similarity score is - log((10 + 1) / 2 × 14) = 1.9459.

The Similarity scores of the three verbs i.e. introduce, expands, compare yield the values shown in Table 1.

Table 1. Similarity Measure Matrix

Verbs/Verbs	Expand	introduce	compare	Present
expand	0.9343	1.9459	2.2335	1.9459
introduce	1.9459	0.9343	2.2335	1.2527
compare	2.2335	2.2335	0.9343	Infinity
present	1.9459	1.2527	Infinity	0.9343

The domain of paper is understood by the title of the paper that is in which sense it is referred. In our experiment, once the similarity vectors are generated for all the extracted verbs and key-terms then according to that similarity vector clustering is performed. For example, if similarity score of verbs or keywords is close to each other then data belongs to same cluster. Similarly if there is a large difference

between similarity score of verbs and keywords then data belongs to different cluster otherwise new cluster is formed for that new data.

4. Hierarchical Clustering

Data clustering is an important technique in the data mining field. Its aim is to group unlabeled data into different groups based on the similarities and dissimilarities between the data elements [8]. A cluster is a collection of objects which are ―similar between them and are ―dissimilar to the objects belonging to other clusters. Typically, a clustering process involves feature selection, selection of a similarity measure, grouping of data and an examination of the resulting output [9]. So we have used verbs as a feature vector. Steps for clustering are as follows:

1) Start by assigning each verb to a cluster, so that if you have *N* verbs, you now have *N* clusters, each containing just one verb. Let the distances (similarities) between the clusters the same as the distances (similarities) between the verbs they contain.
2) Find the closest (most similar) pair of verbs and merge them into a single cluster, so that now you have one cluster less.
3) Compute distances (similarities) between the new cluster and each of the old clusters.

Repeat steps 2 and 3 until all verbs are clustered into a single cluster of size *N* or repeat it until we get required number of clusters

4.1. Algorithm

Input: Similarity *N***N* Matrix of verbs

Output: *m* number of clusters

1) Begin with the disjoint clustering having level $L(0) = 0$ and sequence number $m = 0$.
2) Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to $d[(r),(s)]=\min(d[(i),(j)])$ where the minimum is over all pairs of clusters in the current clustering.
3) Increment the sequence number: $m = m + 1$. Merge clusters (r) and (s) into a single cluster to form the next clustering *m*. Set the level of this clustering to $L(m) = d[(r),(s)]$
4) Update the proximity matrix, *D*, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r, s) and old cluster (k) is defined in this way: $d[(k), (r, s)] = \min d[(k), (r)], d[(k),(s)]$.

If all objects are in one cluster or m clusters (depends on how many clusters we want), stop. Else, go to step 2.

4.2. Results

In Table 2, verb *present* and *introduce* have least distance so they are merged into one cluster and corresponding rows and column of *present* and *introduce* is removed. Here three clusters are formed, cluster 1 contains verb expand, cluster 2 has verbs present and introduce and whereas cluster 3 have verb compare, either we can stop further clustering or continue as per our requirement of number of clusters.

Table 2. Result after Applying 1st Pass of Clustering Algorithm

Verb/Verb	expand	present/introduce	compare
expand	0	1.9459	2.2335
present/introduce	1.9459	0	2.2335
compare	2.2335	2.2335	0

5. Conclusion

Measures that evaluate the impact of research output such as the h-index rely on pure citation counts for producing citation scores. Pure citation counts have been criticized in the literature because they assume all citations are equal and also they are susceptible to manipulation. Citation classification can address both issues by categorizing citations into groups based on the purpose or function of the citation. Our aim in this work is to use an unsupervised machine learning technique to generate a citation classification scheme which can be used for the classification of citations.

For our future work, we will use more features when calculating the similarity between the citation sentences. In our work citation classification is performed on CPU. If dataset is large then it takes more time for execution. Thus in future we can perform citation classification in parallel using GPU which will result into increased speedup.

References

[1] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America: Vol. 102* (pp. 16569–16572).

[2] Lindsey, D. (1989). Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid. *Scientometrics*, *15(3), 189–200.*

[3] Lawrence, P. (2007). The mismeasurement of science. *Curr Biol., 17(15),* R583.

[4] Moravcsik, M., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5(1),* 86.

[5] Dong, C., & Schafer, U. (2011, November). Ensemble-style self-training on citation classification. *Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: Asian Federation of Natural Language Processing* (pp. 623–631).

[6] Leacock, C., & Chodorow, M. (1998). *Combining Local Context and WordNet Similarity for Word Sense Identification* (pp. 305–332). MIT Press.

[7] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database, ser. Language, Speech, and Communication*. MIT Press.

[8] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999, September). Data clustering: A review. *ACM Comput. Surv., 31*, 264–323.

[9] Jain, A., & Dubes, R. (1988). Algorithms for clustering data. *Prentice Hall Advanced Reference Series*. Prentice Hall.

Madhumita S. Jadhav was born on Sept. 20, 1993 at Jalna, Maharashtra, India. Mahumita is a final year student of the Computer Department at Vishwakarma Institute of Technology, Pune, Maharashtra, India. Madhumita has completed her diploma in 2012 in computer Science from Government Polytechnic, Pune.

She has published one national level paper in NCMOC in the field of data mining. Her field of research interest is data mining and big data. She was also an intern at "Innobytes Technology" in Pune for 3 months as a software developer.

Jyoti B. Dhas was born on Jan. 28, 1994 at Jamgaon, Maharashtra, India. Jyoti is a final year student of the Computer Department at Vishwakarma Institute of Technology, Pune, Maharashtra, India. Jyoti has completed her diploma in 2012 in computer science from Government Polytechnic, Pune.

Deepali J. Joshi was an assistant professor since 2008 of the Computer Department at Vishwakarma Institute of Technology, Pune, Maharashtra, India. Deepali has completed her bachelor degree from Bharati Vidyapeeth's College of Engineering Pune. She has one year industrial work experience in honeywell process solution.

Her field of research interest is data mining and software engineering. She has published 2 International papers.

Namrata K. Jadhavrao was born on Oct. 15, 1993 at Pune, Maharashtra, India. Namrata is a final year student of the Computer Department at Vishwakarma Institute of Technology, Pune, Maharashtra, India. Namrata has completed her diploma in 2012 in computer science from Government Polytechnic College Pune.

Sayali K. Kadam was born on June 10, 1994 at Pune, Maharashtra, India. Sayali is a final year student of the Computer Department at Vishwakarma Institute of Technology, Pune, Maharashtra, India. Sayali has completed her diploma in 2012 in computer science from Marathwada Mitra Mandal Polytechnic College Pune.