HOW TO WRITE AWARD-WINNING PAPERS : MINING AWARD-WINNING PAPER CHARACTERISTICS

Chitchanok Chuengsatiansup[1)]

ABSTRACT: A program based on version space learning has been developed and used together with an already-existing decision tree learning program called CU-Miner to study shared characteristics of conference papers receiving the Best Paper award. Both award-winning and non-award-winning papers from two ACM conferences – SIGIR and SIGSOFT – were examined by using the pre-defined numerically measurable attributes extracted from those papers and then analyzed using machine learning methods. The total numbers of illustrations per page and references were found to be among the distinctions between award-winning and non-award-winning papers and relevant to winning the Best Paper award. Results indicate that to win Best Paper awards, paper should not contain too many figures and/or tables. Furthermore, referring to journals and recently published papers is a characteristic of award-winning papers. This work describes the differences between award-winning and non-award-winning papers and serves as a guideline for technical paper writers to produce better papers for future conferences.

KEY WORDS: *machine learning, version space, decision tree, Best Paper award*

1. INTRODUCTION

At each conference, the most outstanding paper submitted to that conference is honored with the "Best Paper Award." However, criteria for selecting best papers differ from conference to conference and are not known to many technical writers. An assumption that award-winning papers share some common characteristics led to the project to find those features. This study used papers from two ACM conferences – SIGIR and SIGSOFT – as inputs of the machine learning process. I found out that references are important and referring to recent works is crucial for winning Best Paper awards. The results revealed shared characteristics of award-winning papers and serve as guidelines for writing technical papers. These guidelines will help technical writers produce better papers for future conferences.

2. MATERIALS

Both award-winning and non-award-winning papers from two ACM conferences – SIGIR and SIGSOFT – ranged from year 1996 to 2009 were studied. Since there was a large number of non-award-winning papers, only some of those were selected. For ACM SIGIR, 13 award-winning papers and 50 non-award-winning papers were used. For ACM SIGSOFT, 56 award-winning papers and 102 non-award-winning papers were used. These papers were found on [1] [2].

3. METHODOLOGY

The procedure for finding shared common characteristics of award-winning papers was divided into three steps: feature extraction, data pre-processing, and machine learning. I began the study by defining attributes to be examined. Then, those attributes were extracted from both award-winning and non-award-winning papers using pattern matching algorithms. Next, data pre-processing reformatted the results from the previous step to suit the input format of the machine learning process. Finally, decision tree and version space performed learning tasks to find relationships and characteristics of award-winning papers.

4. RESULTS

Results obtained from decision tree learning and version space learning were different in format, but they both described the characteristics of award-winning papers. Decision tree learning stated 'attributes we can use to classify award-winning and non-award-winning papers,' while version space learning expressed 'sets of attributes' that award-winning papers had but non-award-winning ones did not.

4.1 CU-Miner

Among all attributes inputted to CU-Miner program, the program chose only two attributes – total number of illustrations (figures and/or tables) per page and rate of journal references – to classify award-winning and non-award-winning papers. First, the total number of illustrations per page was considered. If that number was greater than 1.25, those papers were non-award-winning. Then, the remaining ones were considered together with the rate of journal references. If the value was greater than 0.0645161, those papers were award-winning.

1) Graduate student, Department of Computer Science, the University of Tokyo, Japan (chitchanok@is.s.u-tokyo.ac.jp)

This scheme correctly classified 93.65% of award-winning and non-award-winning papers. (In the experiment, 59 out of 63 of the ACM SIGIR conference papers were correctly classified.)

4.2 Adapted Version Space

Among all attributes used in the version space learning, a reference year revealed a very interesting point. The percentage of award-winning papers referencing to papers published within previous year was substantially greater than those of non-award-winning papers. More precisely, the percentage of award-winning and non-award-winning papers referencing to papers published within previous year were 24.17% and 12.89% respectively. On the contrary, non-award-winning tended to refer to paper published many years previously. The percentage of award-winning and non-award-winning papers referencing to papers published more than four years previously were 4.15% and 9.24% respectively.

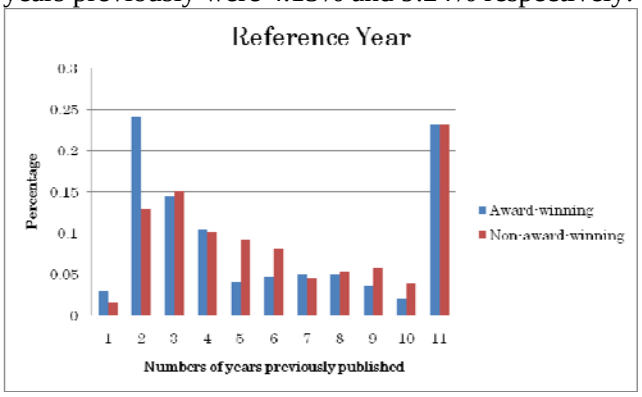

Fig. 1 *Results using reference year*

5. DISCUSSION

5.1 Illustration and Journal Reference

The results from CU-Miner program that used decision tree as a learning method indicate that the total number of illustrations per page and rate of journal references play a crucial role in distinguishing award-winning papers from non-award-winning ones. Non-award-winning papers had more than 1.25 illustrations per page on average, meaning that adding too many figures and/or tables is not recommended. Moreover, all award-winning papers had a rate of journal references greater than 0.0645161, suggesting that referencing journals is a good practice.

5.2 Reference Year

The results obtained from adapted version space learning reveal another important characteristics of award-winning papers, which is the reference year. One of the notable differences between award-winning and non-award-winning papers was award-winning ones referred to papers published within the previous year at a significantly higher rate than non-award-winning ones did. In contrast, non-award-winning papers referred to papers published more than four years previously at a greater rate than award-winning papers. These results show that referring to recently published papers is a shared characteristic of award-winning papers and essential for winning the Best Paper award.

6. CONCLUSION

The idea of papers receiving best paper awards share common characteristics that do not appear in papers that do not receive such awards led to this project of studying papers from two ACM conferences: SIGIR and SIGSOFT. Results suggest that to win the Best Paper award, papers should not contain to many figures and/or tables. Furthermore, referencing to recently published papers and journals is a shared characteristic of award-winning papers.

ACKNOWLEDGMENTS

I would like to extend my appreciation to Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, for providing access to ACM websites. I also thank Boonserm Kijsirikul who gave valuable feedback without which this project would not have been possible.

REFERENCES

[1] www.sigir.org
[2] www.sigsoft.org
[3] Zdravko Markov. *Lecture Notes in Machine Learning*. 2003.