i Title

Doris Hoogeveen , Karin M. Verspoor , Timothy Baldwin, CQADupStack: A Benchmark Data Set for Community Question-Answering Research, Proceedings of the 20th Australasian Document Computing Symposium, p.1-8, December 08-09, 2015, Parramatta, NSW, Australia

ii Keywords

ii1 CQA: Community question-answering (cQA) websites such as WikiAnswers1 and Yahoo! Answers.

ii2 CQADupStack: A benchmark dataset presented in the paper for cQA websites.

ii3 StackExchange: Stack Exchange is a network of question-and-answer websites on topics in varied fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process.

ii4 Retrieval Models: Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Models that help in information retrieval.

iii Breif summaries

iii1 Motivation:

While many different methods have been proposed for identifying duplicate questions in cQA data it is difficult to compare them due to the lack of a publicly available benchmark dataset. Many researchers use their own sets, obtained in various ways. In this paper they aim to solve this problem with the release of a newly constructed data set of anonymized community question answering data that is publicly available for research purposes.

iii2 Related Work:

As the data is dynamic, it is hard to reproduce the exact set used in the research. Many different sets that have been used in cQA work. Some of them are :

1.) 480,190 questions with answers from WikiAnswers 5 used for answer finding in Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding by D. Bernhard and I. Gurevych.

2.) And also like in Finding Question-Answer Pairs from Online Forums by G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun.

iii3 Informative Visualizations:

Figure 1: taken from paper: shows the overview of resolved and unresolved questions

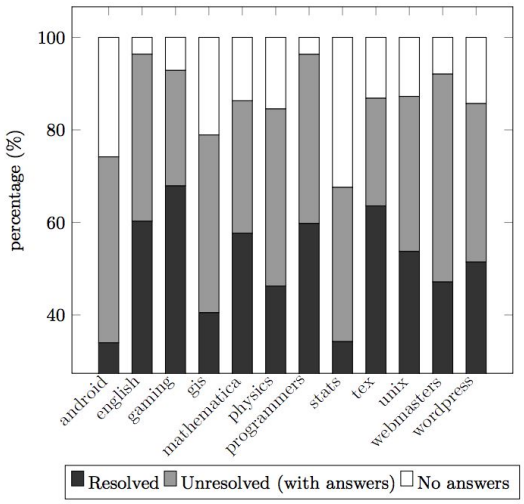

Figure 1: An overview of the percentage of questions in the subforums that are resolved (i.e. they have an answer that has been marked as the right one), the percentage of questions that has not been resolved but that does have answer posts, and the percentage of questions that does not have any answers.

Figure 2: Taken from paper. On duplicate and non duplicate question pairs.

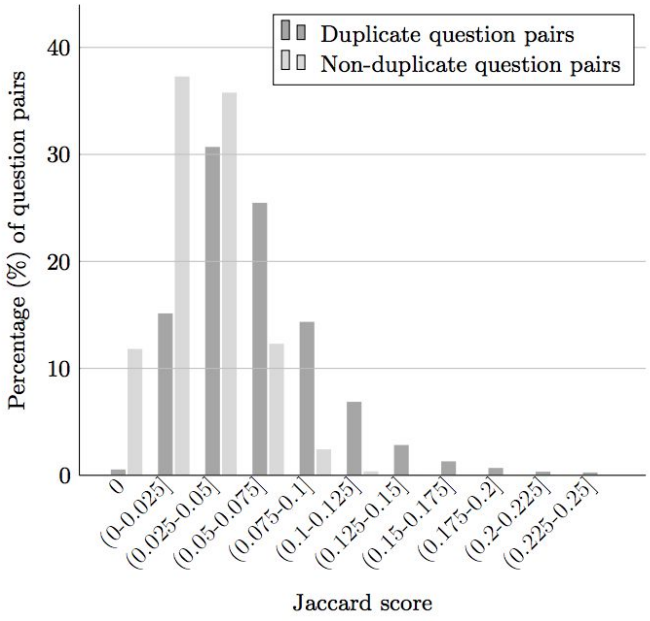

Figure 2: A histogram showing the Jaccard coefficient of duplicate and non-duplicate question pairs. Punctuation and stopwords have been removed. The non-duplicate pairs have been randomly sampled up to a number equal to the duplicate pairs.

iii4 Future Work:

Current evaluation metrics do not handle queries for which the correct result is the empty set. For this reason they only report the scores on the queries for which there are relevant results in the indexed set. What to do with the other queries remains an area for future work. Also the limitation of current retrieval evaluation metrics when it comes to evaluating queries for which there are no relevant results in the set. Strategies for handling this problem can also be considered future work.

iv Scope of Improvement:

iv1 Different researchers might need different data sets for their study. All the this one seems to be comprehensive, but they need to consider the variety of research approaches possible.

iv2 The description on how they chose this dataset could have been more detailed. Also info on how this dataset covers various research types.