ANALYSING INFORMATION QUALITY OF WIKIPEDIA ARTICLES

W.Chinthani Sugandhika Sirisoma

208106X

Degree of Master of Science

Department of Information Technology

University of Moratuwa
Sri Lanka

May 2022

ANALYSING INFORMATION QUALITY OF WIKIPEDIA ARTICLES

W.Chinthani Sugandhika Sirisoma

208106X

Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Information Technology

Department of Information Technology

University of Moratuwa
Sri Lanka

May 2022

DECLARATION OF THE CANDIDATE AND SUPERVISORS

I declare that this is my own work, and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: UOM Verified Signature Date: 30/05/2022

The above candidate has carried out research for the Masters Dissertation under my supervision.

Name of the principal supervisor: Dr. (Mrs.) Supunmali Ahangama

Signatu UOM Verified Signature Date: 02/06/2022

Name of the co-supervisor Dr. Sapumal Ahangama

UOM Verified Signature

Signature Date: 07.06.2022

DEDICATION

I dedicate this dissertation to my parents and my loving husband for always being there with me, cheering me up and standing by my side through all the good times and bad.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to all those who helped me to make this research successful with all their guidance and advice. Firstly, I greatly thank my Principal Supervisor, Dr. (Mrs.) Supunmali Ahangama and Co-supervisor Dr. Sapumal Ahangama for their enormous guidance, advice and encouragement gave for me to conduct this research. Next, I would like to give my heartfelt gratitude to my panel of reviewers Dr. Amal Shehan Perera and Dr. C. R. J. Amalraj for their valuable review comments, suggestions and advice during the progress reviews. I would also thank University of Moratuwa for offering me the grant SRC/LT/2021/24 for carrying out my research successfully. Further, I would like to thank the three participants who participated in conducting the inter-rater reliability test. Lastly, I would like to thank all the academic and non-academic staff of the Department of Information Technology who helped me in diverse ways to make this research a success and a reality.

ABSTRACT

User Generated Content (UGC) is growing in significance for information sharing along with the introduction of Web 2.0. Being one of the largest UGC databases in the world, Wikipedia also stands as the largest community-based collaborative encyclopedia ever created. However, Wikipedia's open-source and collaborative structure presents a serious information quality (IQ) concern. Malicious users take advantage of Wikipedia's popularity on the World Wide Web (WWW) when conducting malicious activities such as link spamming. Wikipedia is therefore often discouraged for use in academic-related activities and research. However. there are some high-quality articles that are both rich in information and quality. Statistical models and machine learning algorithms have been used in existing methods for determining Wikipedia's IQ. However, the outcomes of these models are not satisfactory. Therefore, in this study a novel theoretical model for evaluating IQ is presented, based on Google's E-A-T framework. The model comprises three IQ constructs Expertise, Authority and Trustworthiness. A collection of IQ dimensions that affect the aforementioned three IQ constructs as well as 45 IQ attributes to assess the IQ dimensions were identified and presented based on empirical findings and study results. A Selenium 3.14 web automation script was used to automatically and inexpensively extract the IQ attributes from Wikipedia articles' content and metadata statistics. The data study employed a sample of 2000 articles from six WikiProjects, including 1000 Featured Articles (FA) and 1000 non-FA articles. The suggested model's classification and clustering accuracies were compared to those of three previously published models. The proposed model was compared with three previously published models in terms of classification and clustering accuracy. It received classification and clustering accuracies of 95% and 93% respectively, which is a drastic improvement over the existing models. Furthermore, an average inter-rater agreement of 84% was observed. Accordingly, this comprehensive experiment fairly validates the effectiveness of the suggested model. This study contributes to the related knowledge area by introducing a novel framework to assess Wikipedia articles' IQ.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
IQ	Information Quality
UGC	User Generated Content
WWW	World Wide Web
FA	Featured Article
SERP	Search Engine Result Pages
SEO	Search Engine Optimization
ORES	Online Objective Revision Evaluation Service
VIF	Variance Inflation Factor
EFA	Exploratory Factor Analysis
SVM	Support Vector Machine
KNN	K-Nearest Neighbour