Definitions, Formulas, and Simulated Examples for Plagiarism Detection with FAIR Metrics

Adam Craig
Brain Health Alliance
Ladera Ranch CA, USA
acraig@bhavi.us

Adarsh Ambati
Brain Health Alliance
Ladera Ranch CA, USA
aambati@bhavi.us

Shiladitya Dutta
Brain Health Alliance
Ladera Ranch CA, USA
sdutta@bhavi.us

Arush Mehrotra
Brain Health Alliance
Ladera Ranch CA, USA
amerotra@bhavi.us

S. Koby Taswell
Brain Health Alliance
Ladera Ranch CA, USA
ktaswell@bhavi.us

Carl Taswell
Brain Health Alliance
Ladera Ranch CA, USA
ctaswell@BrainHealthAlliance.org

ABSTRACT

In prior work, we proposed a family of metrics as a tool to quantify adherence to or deviation from good citation practices in scholarly research and publishing. We called this family of metrics FAIR as an acronym for *Fair Attribution to Indexed Reports* and *Fair Acknowledgment of Information Records*, and introduced definitions for these metrics with counts of instances of correct or incorrect attribution or non-attribution in primary research articles with citations for previously published references. In the present work, we extend our FAIR family of metrics by introducing a collection of ratio-based metrics to accompany the count-based metrics described previously. We illustrate the mathematical properties of the ratio-based metrics with various simulated examples in order to assess their suitability as a means of identifying papers under peer review as more or less likely to be suspicious for plagiarism. These FAIR metrics would alert peer reviewers to prioritize low-scoring manuscripts for closer scrutiny. Finally, we outline our planned strategy for future validation of the FAIR metrics with an approach using both expert human analysts and automated algorithms for computerized analysis.

KEYWORDS

Scientometrics, bibliometrics, FAIR metrics, citation practices, plagiarism detection.

ASIS&T THESAURUS

Scientometrics (1569), semantic analysis (1435), plagiarism (176).

INTRODUCTION

Plagiarism continues to pose a threat to the integrity of scholarly research and publishing [Steen, 2011]. While lexical comparison tools are useful for detecting exact copying or mildly obfuscated copying, they often fail to detect plagiarism of conceptual knowledge, substantive content and/or novel ideas [Meuschke & Gipp, 2013], defined by [Maurer et al., 2006] as “using similar concept or opinion which is not common knowledge.” In contrast, semantic analysis offers an alternative approach that can even detect plagiarism despite translation of the text into another language [Potthast et al., 2010]. Strategies range from using Latent Semantic Indexing (LSI) for detection of topic boundaries [Rehurek, 2008] to multi-phase processes such as those used by [Sindhu & Idicula, 2016] and [Osman & Salim, 2013]. The FAIR metrics that we present here do not represent a new algorithmic approach to semantic analysis. Instead, the FAIR metrics that we expound here do provide a quantitative tool to measure the relative numbers of pairs of matching and mismatching statements found through semantic analysis. We define and formulate these metrics to quantify FAIR citing behavior for the purpose of promoting established traditional standards that repudiate plagiarism when publishing scholarly literature in general, and the medical scientific research literature in particular.

In [Craig & Taswell, 2018a], we presented core design principles for the FAIR metrics, which we explained could serve as a countervailing influence to the perverse incentive engendered by overreliance on current citation metrics that discourage some authors from citing rival authors, i.e., not adhering to the tradition of *standing on the shoulders of giants*[1] and citing other authors who previously published related work in their same research field that should be relevant to their discussion of the literature. Subsequently, we formulated a set of such metrics suitable for use with primary research articles [Craig & Taswell, 2018b] where we defined metrics based on the possible relationships between whether a statement is actually novel or not and whether the authors claim or discuss it as novel or not. In this current paper, we continue our past work on the FAIR metrics, described previously as simple counts of instances of the

82nd Annual Meeting of the Association for Information Science & Technology, Melbourne, Australia, October 19–23, 2019

[1] This phrase has been attributed to Isaac Newton who wrote it in a letter to Robert Hooke in 1675, but the metaphor of dwarfs standing on the shoulders of giants has a history dating back many centuries earlier [Wikipedia, 2019].

different types of statements, and extend it by formulating a family of four ratio-based FAIR metrics with mathematical notation and definitions for their formulas, demonstrating their numerical properties and evaluating their suitability as plagiarism detection tools via simulated examples of the underlying counts for the calculated values of the metrics.

FORMULATION OF RATIO-BASED FAIR METRICS

Following [Craig & Taswell, 2018a,b], we further refined and extended our definitions of the simple count-based FAIR metrics to clarify how they interrelate with the formal notation summarized in Table 1 and explained in detail here for the ratio-based FAIR metrics.

Let $G(A)$ mean that the function G operates on set A. Let $G(A|B)$ mean that the function G operates on set A conditionally given set B, or that the function G operates on set A in comparison with set B. Let C refer to the set C of statements in a Control paper or in a Comparison Collection of papers. Let T refer to the set T of statements in a Test paper.

Now let $K(C)$ be the number K of Known, observed, original, and previously published statements found in the set C. Let $M(T|C)$ be the number M of Misquoted, Misattributed and/or Mistaken statements found in set T similar to known statements found in set C that have been repeated from C with incorrect citation and thus done without FAIRness. Let $Q(T|C)$ be the number Q of Quoted statements found in set T similar to known statements found in set C that have been repeated from C with correct citation and thus done with FAIRness. Let $P(T|C)$ be the number P of Paraphrased or Plagiarized statements found in set T similar to known statements found in set C that have been repeated from C without any citation and thus done without FAIRness. Let $N(T|C)$ be the number N of Novel or Non-plagiarized statements found in set T not similar to any known statements found in set C, ie, those that have not been repeated from C and that do not require any citation of a reference from C and thus done with FAIRness.

For the combined category counts S and R: Let $S(T|C) = M(T|C) + Q(T|C) + P(T|C) \leq K(C)$ be the total number S of Similar statements (= misquoted + quoted + plagiarized statements) found in set T repeated from and similar to known statements found in set C. Then $S(T|C)$ represents the *intersection* of statements found in both T and C, ie, those that have been repeated in T from C. Let $R(T|C) = M(T|C)+Q(T|C)+P(T|C)+N(T|C) \geq K(C)$ be the total number R of all Reported statements (= misquoted + quoted + plagiarized + novel statements) found in set T when compared to statements found in set C. Then $R(T|C)$ represents the *union* of statements found in both T and C, ie, both those that have and have not been repeated in T from C. Also, require the imposed condition that $0 < S(T|C) \leq K(C) \leq R(T|C)$ in order to prevent division by zero in any of the formulas for the ratio-based FAIR metrics.

We then defined formulas for the FAIR metrics as ratios of the four categories of basic counts M, Q, P and N and the two categories of combined counts S and R so that they could serve as measures of FAIRness quality independent of the size, scope and scale of the test set T in comparison with the control set C. These formulas are summarized in Table 2 and explained in detail here.

Define the first FAIR metric

$$F_1(T|C) = Q(T|C)/S(T|C) \tag{1}$$

as the ratio of the Quoted count to Similar count. When Misquoted and Quoted counts are zero with $M(T|C) = Q(T|C) = 0$ and the Plagiarized and Similar counts are equal with $P(T|C) = S(T|C) > 0$, then $F_1(T|C) = 0$. When Misquoted and Plagiarized counts are zero with $M(T|C) = P(T|C) = 0$ and the Quoted and Similar counts are equal with $Q(T|C) = S(T|C) > 0$, then $F_1(T|C) = 1$. Observe that increasing F_1 between 0 and 1 means increasing FAIRness.

Define the second FAIR metric

$$F_2(T|C) = [Q(T|C) - M(T|C)]/S(T|C) \tag{2}$$

as the ratio of the Quoted-Misquoted difference count to Similar count. When Quoted and Plagiarized counts are zero with $Q(T|C) = P(T|C) = 0$ and the Misquoted and Similar counts are equal with $M(T|C) = S(T|C) > 0$, then $F_2(T|C) = -1$. When Misquoted and Plagiarized counts are zero with $M(T|C) = P(T|C) = 0$ and the Quoted and Similar counts are equal with $Q(T|C) = S(T|C) > 0$, then $F_2(T|C) = 1$. Observe that increasing F_2 between -1 and 1 means increasing FAIRness with a zero boundary and sign change indicating a transition when $Q(T|C) > M(T|C)$.

Define the third FAIR metric

$$F_3(T|C) = [Q(T|C) - P(T|C)]/S(T|C) \tag{3}$$

as the ratio of the Quoted-Plagiarized difference count to Similar count. When Misquoted and Quoted counts are zero with $M(T|C) = Q(T|C) = 0$ and the Plagiarized and Similar counts are equal with $P(T|C) = S(T|C) > 0$, then $F_3(T|C) = -1$. When Misquoted and Plagiarized counts are zero with $M(T|C) = P(T|C) = 0$ and the Quoted and Similar counts are equal with $Q(T|C) = S(T|C) > 0$, then $F_3(T|C) = 1$. Observe that increasing F_3 between -1 and 1 means increasing FAIRness with a zero boundary and sign change indicating a transition when $Q(T|C) > P(T|C)$.

Define the fourth FAIR metric

$$F_4(T|C) = [Q(T|C) - N(T|C)]/R(T|C) \tag{4}$$

as the ratio of the Quoted-Novel difference count to Reported count. When Misquoted and Quoted counts are zero with $M(T|C) = Q(T|C) = 0$ and the Novel and Reported counts are equal with $N(T|C) = R(T|C) > 0$, then $F_4(T|C) = -1$. When Misquoted, Plagiarized and Novel counts are zero with $M(T|C) = P(T|C) = N(T|C) = 0$ and the Quoted and Reported counts are equal with $Q(T|C) = R(T|C) > 0$, then $F_4(T|C) = 1$. Observe that increasing F_4 between -1 and 1 means increasing FAIRness with a zero boundary and sign change indicating a transition when $Q(T|C) > N(T|C)$.

Symbol	Definition
C	set C of statements in a Control paper or in a Comparison Collection of papers
$F_1(T\|C)$	FAIR metric scaled on interval $[0, 1]$ for T compared with C (dependent on Q and S)
$F_2(T\|C)$	FAIR metric scaled on interval $[-1, +1]$ for T compared with C (dependent on Q, M and S)
$F_3(T\|C)$	FAIR metric scaled on interval $[-1, +1]$ for T compared with C (dependent on Q, P and S)
$F_4(T\|C)$	FAIR metric scaled on interval $[-1, +1]$ for T compared with C (dependent on Q, N and R)
$G(A)$	function G operates on set A
$G(A\|B)$	function G operates on set A in comparison with set B (on set A given set B)
$M(T\|C)$	number M of Misquoted (incorrectly cited) statements found in T compared with C
$N(T\|C)$	number N of Novel (uncited) statements found in T compared with C
$K(C)$	number K of Known statements found in C
$P(T\|C)$	number P of Plagiarized (uncited) statements found in T compared with C
$Q(T\|C)$	number Q of Quoted (correctly cited) statements found in T compared with C
$R(T\|C)$	number R of Reported statements found in T compared with C
$S(T\|C)$	number S of Similar statements found in T compared with C
T	set T of statements in a Test paper

Table 1. Notation for FAIR metrics

Symbol		Formula
$F_1(T\|C)$	=	$Q(T\|C)/S(T\|C)$
$F_2(T\|C)$	=	$[Q(T\|C) - M(T\|C)]/S(T\|C)$
$F_3(T\|C)$	=	$[Q(T\|C) - P(T\|C)]/S(T\|C)$
$F_4(T\|C)$	=	$[Q(T\|C) - N(T\|C)]/R(T\|C)$
$S(T\|C)$	=	$M(T\|C) + Q(T\|C) + P(T\|C) \leq K(C)$
$R(T\|C)$	=	$M(T\|C) + Q(T\|C) + P(T\|C) + N(T\|C) \geq K(C)$

Table 2. Formulas for ratio FAIR metrics with required condition $0 < S(T|C) \leq K(C) \leq R(T|C)$

Name	M	Q	P	N	S	R	F_1	F_2	F_3	F_4
None	0	0	0	0	0	0	NaN	NaN	NaN	NaN
MOnly	10	0	0	0	10	10	0	-1	0	0
QOnly	0	10	0	0	10	10	1	1	1	1
POnly	0	0	10	0	10	10	0	0	-1	0
NOnly	0	0	0	10	0	10	NaN	NaN	NaN	-1
MAndQ	10	10	0	0	20	20	0.5	0	0.5	0.5
MAndP	10	0	10	0	20	20	0	-0.5	-0.5	0
MAndN	10	0	0	10	10	20	0	-1	0	-0.5
QAndP	0	10	10	0	20	20	0.5	0.5	0	0.5
QAndN	0	10	0	10	10	20	1	1	1	0
PAndN	0	0	10	10	10	20	0	0	-1	-0.5
NoN	10	10	10	0	30	30	0.33	0	0	0.33
NoP	10	10	0	10	20	30	0.5	0	0.5	0
NoQ	10	0	10	10	20	30	0	-0.5	-0.5	-0.33
NoM	0	10	10	10	20	30	0.5	0.5	0	0
All	10	10	10	10	30	40	0.33	0	0	0
GoodA	0	5	0	5	5	10	1	1	1	0
GoodB	0	10	0	10	10	20	1	1	1	0
GoodC	0	15	0	15	15	30	1	1	1	0
BadA	5	0	5	0	10	10	0	-0.5	-0.5	0
BadB	10	0	10	0	20	20	0	-0.5	-0.5	0
BadC	15	0	15	0	30	30	0	-0.5	-0.5	0
SplitA	5	5	2	2	12	14	0.42	0	0.25	0.21
SplitB	10	10	4	4	24	28	0.42	0	0.25	0.21
SplitC	15	15	6	6	36	42	0.42	0	0.25	0.21

Table 3. Examples of the effects of different combinations of zero and nonzero counts on the FAIR metrics

This ratio-based family of metrics $F_i(T|C)$ for FAIRness has been designed to depend on the Misquoted count $M(T|C)$, Quoted count $Q(T|C)$, Plagiarzied count $P(T|C)$, and Novel count $N(T|C)$ relative to the Similar count $S(T|C)$ or Reported count $R(T|C)$ in different ways for different scenarios that all nevertheless focus on various aspects of the definition of FAIRness for appropriately citing previously published references. In particular, note especially that the presence of novelty should not nullify nor otherwise hide the simultaneous presence of plagiarism and vice versa. All metrics in the family $F_i(T|C)$ for FAIRness have been designed such that increasing values correspond to increasing FAIRness. The interval of values $[0, 1]$ for F_1 is different from the interval of values $[-1, +1]$ for F_2, F_3, and F_4 because F_1 has a numerator with a simple count instead of the differences of counts used in the numerators for the other metrics in the family.

SIMULATED EXAMPLES OF FAIR METRICS

Table 3 lists a variety of different FAIRness metric scenarios with simulated examples of possible situations that might occur. The cases *None* through *All* illustrate the numerical stability of the metrics, which are only undefined in two obviously pathological cases: an empty paper devoid of any statements whatsoever and the *NOnly* paper consisting entirely of novel claims. In this latter hypothetical case, while the authors have not plagiarized any known prior work as indicated by a count of $P(T|C) = 0$, they have also failed to place their novel claims in the context of the existing body of knowledge found within the related field of published literature as indicated by a count of $Q(T|C) = 0$. For another possible explanation of this case, such a scenario might occur

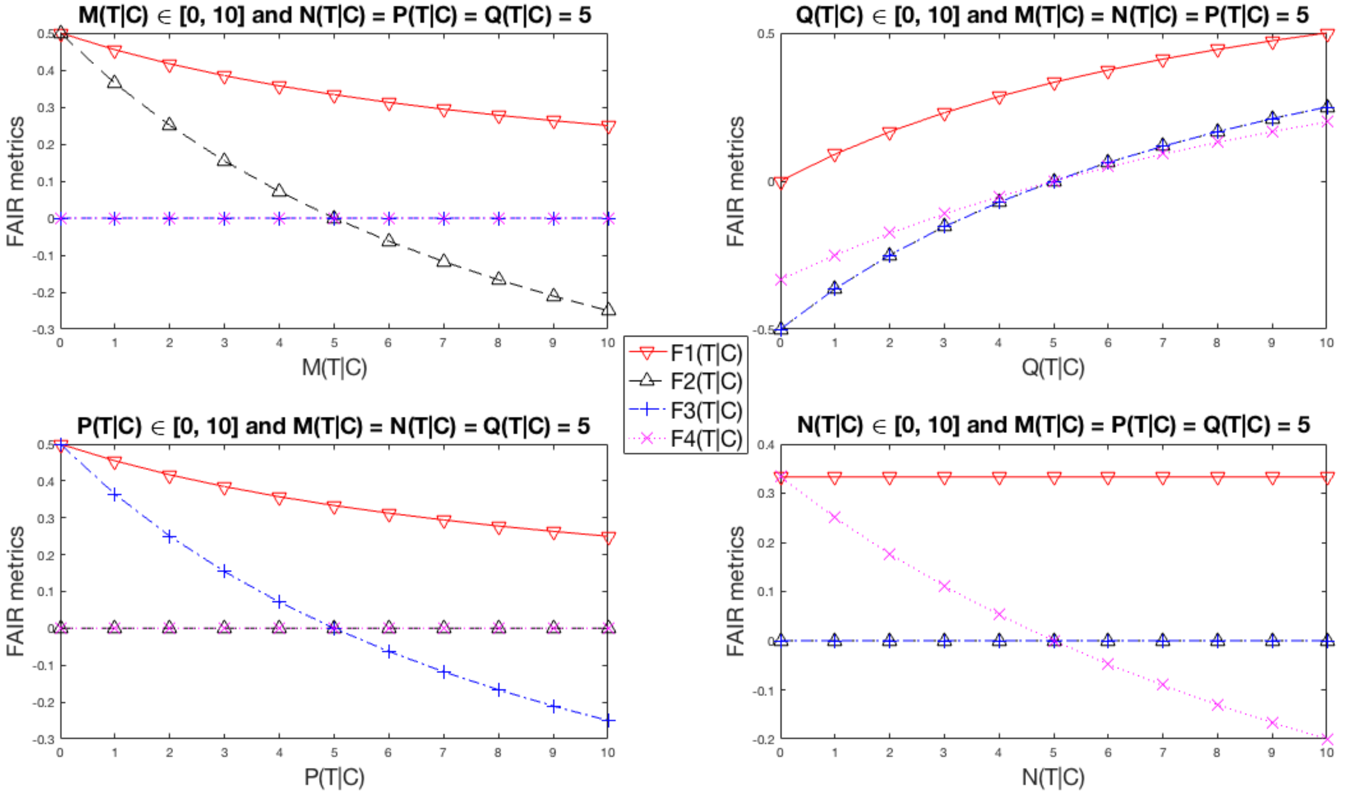

Figure 1. Effect on FAIR metrics of varying each count in turn while holding the others constant

if the evaluation system failed to find sets of statements for any of the papers that the authors cited. This case may arise if the automated system performing the semantic analysis for the counts has incorrectly retrieved the C statements from the wrong source, such as an incorrectly identified service in the Nexus-PORTAL-DOORS-Scribe cyberinfrastructure system [Craig et al., 2016; Taswell, 2008, 2009, 2010], devoted to a problem-oriented domain other than the one to which the T paper relates. The rest of the cases, *GoodA* through *SplitC*, illustrate that each metric remains constant so long as the proportions of the counts remain constant. This stability implies that the metrics serve appropriately as measures of adherence to good citation practices independently of the scope of the work and the number of statements identified in the T paper.

Figure 1 shows how each metric varies as we vary one of the counts and hold the rest constant. $F_2(T|C)$ decreases most rapidly as $M(T|C)$ increases (upper left), reflecting its design as a measure of misattribution. $F_3(T|C)$ decreases most rapidly as $P(T|C)$ increases (lower left), showing its suitability for flagging possible instances of plagiarism. $F_1(T|C)$ decreases less rapidly but in response to increases in $M(T|C)$, $P(T|C)$, or both. As such, it serves as a more general measure of FAIRness. All metrics increase as $Q(T|C)$ increases (upper right), because proper citation of prior work is the desirable content of which we are measuring the prevalence relative to other content. Although the presence of novel statements is desired in primary research publications, $F_1(T|C)$, $F_2(T|C)$, and $F_3(T|C)$ remain constant as $N(T|C)$ increases (lower right), while $F_4(T|C)$ decreases. This behavior of the FAIR metrics reflects the standard expectation that scholars should place novel claims in the proper context by relating them to published knowledge and hypotheses on the topic. Not inappropriately favoring papers with novel claims also makes these FAIR metrics applicable to reproducibility and replication studies as well as review articles, which do not need to report novel findings in order to fulfill their roles in the ecosystem of scholarly literature.

STRATEGY FOR VALIDATION OF FAIR METRICS

Now that we have designed and demonstrated the FAIR metrics with the requisite mathematical properties, our next step will be to test and validate them on real-world examples of papers known in advance to be plagiarizing or novel. Although automated semantic similarity detection has been improving, the capabilities of existing tools remain inferior to those of human minds [Agirre et al., 2016]. Moreover, any methodology to compute the FAIR metrics, regardless of whether human guided or computer automated, will require processing several steps.

The first step extracts the most essential and significant statements from each article in a clear, concise, consistent format. Currently, methods for extracting RDF triples from free-form text still depend on the rough structure of the text in order to derive triples. These methods convert each sentence into a set of triples without regard for the importance or non-importance of the claim that each RDF triple represents within the semantic thematic content, reasoning and discussion communicated by the article analyzed [Klyne & Carroll, 2004]. The second step acquires a learned summary of the statements. An example of a machine learning approach for this task appears in [Leskovec et al., 2004].

The third step discerns common knowledge that does not require citation from knowledge that should require citation. The calculation framework needs to filter out common knowledge, as it is not novel but does not require citation either, meaning that it is not considered for categorization into any of the four basic counts $M(T|C)$, $Q(T|C)$, $P(T|C)$, or $N(T|C)$. Without the experience of the world that humans take for granted, automated agents must rely on compendious semantic networks, such as ConceptNet 3 [Havasi et al., 2007]. Furthermore, each problem domain has its own repository of common knowledge that experts and ontology engineers must encode in a domain ontology [Fensel, 2001]. Therefore, the system would need to search the relevant ontologies for equivalents to each key claim.

The fourth step compares each statement in the test paper T to potential equivalents in the set of statements from prior works C. For a human analyst, deciding whether two natural language statements are equivalent can require in-depth understanding of the concepts and relationships involved and may require careful consideration of nuance and context, leading to variance even in human-generated ratings of semantic similarity, as seen in [Agirre et al., 2016]. Despite these challenges, automated semantic equivalence detection systems continue to advance, as we can see from the improvement in top scores between [Agirre et al., 2012] and [Agirre et al., 2016].

Finally, the FAIR metric calculations must be evaluated for consistency across the family of results with F_1, F_2, F_3, and F_4 for a given test paper T in comparison to the control collection of papers C. While each of these requisite steps has possible computerized algorithmic solutions, none has yet achieved the equivalent of human reasoning ability. With these issues in mind, we have chosen to begin with an approach in which human analysts manually extract and compare the core claims of each text. Future results from this human expert-derived analysis will provide a best-case scenario benchmark against which to compare subsequent automated computerized solutions that calculate FAIR metrics from counts of statements analyzed via natural language processing (NLP) and concept similarity detection (CSD) tools which we are developing [Bae et al., 2017; Craig et al., 2019; Dutta & Taswell, 2018].

DISCUSSION OF PILOT STUDY

In a preliminary pilot study, we tested the feasibility of this manual approach with eight volunteer analysts using confirmed instances of plagiarism in the fields of brain science and computer science. To keep the workload for each analyst manageable with balanced data sets, we created a text corpus consisting of collections of triples of papers with each triple consisting of one paper in each of the following three roles: a plagiarizing test paper T_p (with subscript p for plagiarizing) as a primary research article retracted for plagiarism, found through the Retraction Watch database, a comparison control paper C as the paper that T_p plagiarized according to its retraction notice, and a novel test paper T_n (with subscript n for novel) as another primary research article not known to have been retracted for plagiarism, but also with a close conceptual relationship to C, found through a Google Scholar search with a list of key terms from C.

From our pilot study, it became clear that having a control set C of prior literature consisting only of statements from a single paper meant discarding too many statements found in T_p or T_n attributed to other previously published papers because the analyst might not find them in C and thus be unable to classify them as either M or Q counts. Thus, the pilot study design produced low-biased values for these two counts, M and Q, which we knew were unrealistic and not representative of the real-world situation. When we conduct our full study with human analysts using a more labor intensive but also more realistic approach for the experimental study design, we will need to construct for each T_p and T_n a more comprehensive set C consisting of statements from all papers that we know T_p and T_n either plagiarized and/or cited. Then we should be able to obtain better estimates for all of the four basic categories of counts M, Q, P, and N.

Another important lesson that we learned from the pilot study involved the manner in which plagiarizing papers could escape detection by the FAIR metrics by changing the entities described in the claims in such a way that the statements are no longer similar enough to match as the same statements found in T compared with C. For example, [Yao et al., 2017] is nearly identical in research methodology and wording of text to [Li et al., 2015], but it substitutes ligand-receptor pair CXCL12/CXCR4 for each reference to CCL21/CXCR7 and transcription factor Twist for each instance of Slug in the text. Since these substitutions make the claims of the plagiarizing paper conceptually different from those of the original when the similarity comparison is performed without the use of weighting factors, it can achieve a relatively high FAIR metric score. Such cases fall outside of the scope of the simple FAIR metrics as currently formulated for T compared with C unless we modify them further to allow for additional weighting factors obtained from the outputs of more sophisticated analyses performed with artificial neural networks that have been trained on C.

CONCLUSION

Nevertheless, we have already shown that the simple unweighted versions of the FAIR metrics can serve to complement the lexical plagiarism detection techniques that have been effective at detecting copied text surrounding altered key terms [Meuschke & Gipp, 2013]. The current unweighted FAIR metrics can also serve to alert human peer reviewers and editors who can then better decide how to make appropriate judgments when scrutinizing manuscripts with low FAIRness scores suspected of plagiarism. Our next goal for subsequent studies of the FAIR metrics with primary research articles, using collections with those which are known *a priori* with high confidence to be either plagiarizing or non-plagiarizing, will be to evaluate and establish differences in the distributions of FAIR metric scores for plagiarizing papers compared with those for non-plagiarizing papers. Acquired experience with different problem-oriented research domains will enable members of each of those scientific communities to establish thresholds for each of the FAIR metrics for sufficient FAIRness to avoid triggering alerts for possible plagiarism. While we would expect $P(T|C)$ ideally to be zero for correctly referenced non-plagiarizing articles and nonzero for plagiarizing articles, the presence of both mistakes by authors in their writing of manuscripts as well as flaws in the NLP extraction and semantic comparison of statements could complicate the analysis and results. Even so, we expect that the distributions of the FAIR metric scores will be sufficiently different for plagiarizing versus non-plagiarizing articles that they will serve as a useful tool for expert reviewers to prioritize papers for further examination when the FAIR metrics yield low values suspicious for plagiarism.

REFERENCES

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., & Wiebe, J. (2016). SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Agirre, E., Diab, M., Cer, D., & Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Bae, S. H., Craig, A. G., & Taswell, C. (2017). Expanding Nexus diristries of dementia literature with the NPDS concept-validating search engine agent. In *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3704–3707.

Craig, A., Ambati, A., Dutta, S., Kowshik, P., Nori, S., Taswell, S. K., Wu, Q., & Taswell, C. (2019). DREAM principles and FAIR metrics from the PORTAL-DOORS Project for the semantic web. In *Proceedings 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE. in press.

Craig, A., Bae, S. H., Veeramacheneni, T., Taswell, S. K., & Taswell, C. (2016). Web service APIs for Scribe registrars, Nexus diristries, PORTAL registries and DOORS directories in the NPD system. In *Proceedings of the 9th International SWAT4LS Conference*.

Craig, A. & Taswell, C. (2018a). The FAIR metrics of adherence to citation best practices. In *Proceedings of ASIS&T 81st Annual Meeting SIGMET Workshop*.

Craig, A. & Taswell, C. (2018b). Formulation of FAIR metrics for primary research articles. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Dutta, S. & Taswell, C. (2018). SPARQL-based search engine and agent for finding brain literature and converting references to NPDS metadata records. In *Proceedings of the 11th International Conference on Brain Informatics*, number B277.

Fensel, D. (2001). Ontologies. In *Ontologies*, pages 11–18. Springer.

Havasi, C., Speer, R., & Alonso, J. (2007). ConceptNet 3: A flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer.

Klyne, G. & Carroll, J. J. (2004). Resource description framework (RDF): concepts and abstract syntax.

Leskovec, J., Grobelnik, M., & Milic-Frayling, N. (2004). Learning substructures of document semantic graphs for document summarization. In *LinkKDD*.

Li, G., Yang, Y., Xu, S., Ma, L., He, M., & Zhang, Z. (2015). Slug signaling is up-regulated by ccl21/cxcr7 to induce emt in human chondrosarcoma. *Medical Oncology*, 32(2):2.

Maurer, H. A., Kappe, F., & Zaka, B. (2006). Plagiarism-a survey. *J. UCS*, 12(8):1050–1084.

Meuschke, N. & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, 9(1).

Osman, A. H. & Salim, N. (2013). An improved semantic plagiarism detection scheme based on chi-squared automatic interaction detection. In *International Conference on Computing, Electrical and Electronic Engineering (ICCEEE)*, pages 640–647. IEEE.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., & Rosso, P. (2010). Overview of the 2nd international competition on plagiarism detection. In *Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*, pages 1–14.

Rehurek, R. (2008). *Semantic-based plagiarism detection*. PhD thesis, Masarykova univerzita, Fakulta informatiky.

Sindhu, L. & Idicula, S. M. (2016). A plagiarism detection system for Malayalam text based documents with full and partial copy. *Procedia Technology*, 25:372–377.

Steen, R. G. (2011). Retractions in the scientific literature: do authors deliberately commit research fraud? *Journal of Medical Ethics*, 37(2):113–117.

Taswell, C. (2008). DOORS to the semantic web and grid with a PORTAL for biomedical computing. *IEEE Transactions on Information Technology in Biomedicine*, 12(2):191–204. In the Special Section on Bio-Grid.

Taswell, C. (2009). The hierarchically distributed mobile metadata (HDMM) style of architecture for pervasive metadata networks. In *Proceedings of ISPAN 2009 The 10th International Symposium on Pervasive Systems, Algorithms and Networks*, pages 315–320. IEEE.

Taswell, C. (2010). A distributed infrastructure for metadata about metadata: The HDMM architectural style and PORTAL-DOORS system. *Future Internet*, 2(2):156–189. In Special Issue on Metadata and Markup.

Wikipedia (2019). Standing on the shoulders of giants.

Yao, C., Li, P., Song, H., Song, F., Qu, Y., Ma, X., Shi, R., & Wu, J. (2017). Retraction note to: Cxcl12/cxcr4 axis upregulates twist to induce emt in human glioblastoma. *Molecular Neurobiology*, 54(9):7553–7553.