如何理解论文查重报告的相似率

作者：writer 来源：LetPub

本文原地址：https://www.iikx.com/news/article/340.html

当一篇文章经查重软件处理后，相似率是第一个出现的结果。因为我们很容易把注意力放在这个表示有问题的数字上，所以新用户通常会问的问题是什么样的相似率说明有问题?

这个问题的答案是，没有一个神奇的数字能够告诉你一篇文章是否包含有问题的内容。相似率只是为你提供一个粗略的标题，以确保你能够直接注意到那些有大量重复的文章，而快速忽略掉几乎没有重复的文章。除此之外，相似率本身不会给你确切的答案，也绝对不能告诉你这篇文章是否有抄袭的情况。

为什么会这样呢?

其实，当评估一篇文章的整体相似率时要考虑到若干因素。

首先，需要注意的是相似率告诉你的是一篇文章中和其他文章相同 (即所谓的匹配)的文字的总量。这个总量可能是由许多较小的匹配组成的。相似率30%有可能是指30%匹配同一篇文章，但更有可能的情况是，这30%是由许多较小的匹配相加而成，这些小的匹配最大都不超过4或5%。这只有在看详细的查重报告时才能看出来。

当然，一篇有6个5%匹配的文章可能和一篇30%都抄自同一篇文章的文章一样有抄袭的问题。不过不看查重报告就没法确定了。

其次，匹配出现在文章的哪一部分有时比到底有多少文字匹配更重要。例如，某些学科领域的编辑可能不太在意方法部分的重复，因为要描述一个过程也只有那么多的方式。而另一方面，在讨论或结论部分的匹配，尽管它可能只占手稿的一小部分，如果没有适当的引用，也会引起编辑的怀疑。

同样的，一类文章的可接受的阈值未必适合另一类型的文章：综述文章相似率通常会比研究文章高一些。

同样需要记住的是在未编辑的手稿中可能存在一些简单地错误而导致查重软件错误地标出存在匹配的部分。查重软件的排除书目功能是依赖于在文章的参考文献部分有一行是reference这个标题。如果这个标题在手稿中被省略，参考文献部分将不被排除在外。

同样，排除引文功能是通过查找引号。如果作者没有使用引号或是在开头或结尾漏掉一个引号时，系统不会识别出引用的文字，即使编辑们可以通过文章布局和参考文献一眼看出是引用的文字

。

基于以上所有的原因，比起单单只看查重的相似率而言，更重要的是查看查重报告。

Understanding the Similarity Score

The similarity score is the ?rst thing you see when a document is processed and, because it ’ s easy to focus on this number as signifying a problem, a common question new users of the system ask is ‘ what level of similarity score indicates a problem? ’

The answer to this question is there is no such thing as a ‘ magic number ’ that will tell you whether a document contains problematic content.The similarity score gives you a rough ‘ headline ’ that ensures heavily duplicated papers are brought straight to your attention and allows you to quickly disregard papers with hardly any matches. Beyond that, the score itself doesn ’ t give you de?nitive answers and de?nitely cannot tell you whether you have a case of plagiarism.

Why is this?

Well, there are a number of factors that need to be taken into account when assessing a paper ’ s overall similarity score.

Firstly, it ’ s important to note the similarity score is telling you the total amount of matching text. This is probably going to be made up of a number of smaller matches. It is possible a 30% score will turn out to be a 30% match to one source, but it ’ s much more likely that when you look at the reports you ’ ll ?nd the 30% is made up of a number of smaller matches, the largest of which might be just 4 or 5%.

Of course, a paper with six separate matches of 5% could well be as problematic as one that has copied 30% of its content from a single source, but it ’ s impossible to tell whether this is the case without looking at the reports.

Secondly, where the match appears can sometimes be more important than how big the match is. For example, editors in certain subject areas may be less concerned about sizable matches in methods sections, where there are only so many ways to describe a certain process. A match in the discussion or conclusions with no appropriate citation, on the other hand, could set alarm bells ringing even though it only accounts for a small percentage of the manuscript.

Similarly, acceptable thresholds for one type of article may not be appropriate for another: Review articles could be expected to have a higher overall similarity score than original research articles.

It is also important to bear in mind there could be simple errors in the unedited manuscript that mean matches are picked up incorrectly. The exclude bibliography feature of sofewares relies on the reference section having a title on its own line within the document. If this is omitted from the manuscript, the references will not be excluded.

Similarly, the exclude quotes feature looks for quotation marks. If the author has not used quotation marks or missed one at the start or end of the passage, the system will not recognize it as a quote, even though it

might be apparent to the editor due to its layout and reference.

For all of these reasons it ’ s important to look at the reports rather than rely on the similarity score alone.