Software/services to detect plagiarism

How it works—software vs. service

Briefly, there are several effective algorithms for the comparison of text, which can quickly and accurately compare a submitted document to a large library of published documents, be they peer-reviewed journal publications or web content. These algorithms compare significant keywords (including synonyms, acronyms, lexical variants), statistically improbably phrases (including paraphrased content), and/or align sentences to compute a measure of similarity, and then provide those results to the user, including control over thresholds that trigger users to inspect 'suspiciously similar' text. Then, these sections of similar text in both the query and that found by the search algorithms are usually displayed as a list or side-by-side to the user to make the final judgment as to acceptability.

Selecting a plagiarism detection service

There are many things to be considered before selecting a plagiarism (or document similarity) detection service. These include compatibility with one's document management system, completeness (what database do they compare a query to), security, and of course cost. More such considerations are provided in Table 1. Although there are many that offer a plagiarism detection service, and they all claim to have certain advantages over the competition, there has been no head to head competitive analysis by an independent entity to determine the relative performance of each. In Table 2 is a sampling of the available companies and organizations. However, as representative examples of certain types of services/organizations, 3 will be discussed in more detail—CrossCheck, IThenitcate, and eTBLAST—a membership-based plagiarism service for the publication industry, the leading commercial plagiarism detection service for the publication industry, and a free service, respectively.

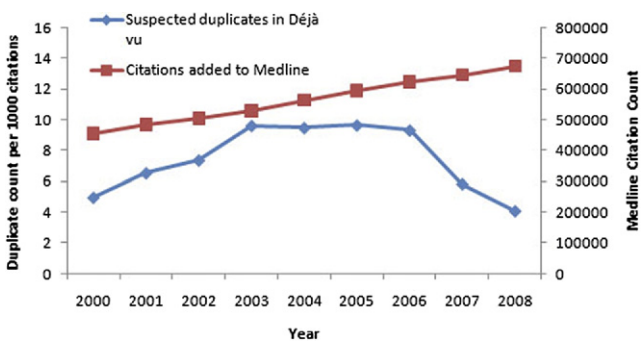

Fig. 1. In 2008, there were 4.1 new highly similar pairs of manuscripts per 1,000 published papers in Medline and deposited in the Déjà vu database. This is a major decline that has taken place in the last 2 years. One could speculate on a number of reasons, including fear of detection by would-be perpetrators, but whatever the cause, the problem is getting better but it is still significant in size. (Color version of figure is available online.)

Table 1
Considerations when selecting a plagiarism detection system

Databases searched, completeness, appropriateness
Which databases are searched and are they appropriate for my needs?
What is my search missing?
How often are the search databases updated?
Sensitivity and specificity of search algorithm
How well does the similarity search work? Or is that known or proprietary?
What is the false positive and false negative rate? What is this for my typical queries?
How do I handle a false positive? Are there so many that sorting though them is exhausting?
Compatibility with journal manuscript submission system
How do I automate the checking process?
Is there an API available that is compatible with my system?
Security
How is my data transmitted to and from the service?
How long does my query stay in the system?
User interface
When the results come back, are they presented in a meaningfully and easily assimilated way?
Control over threshold and other parameter settings
Can I control the settings to minimize false positives and false negatives?
Can I give priority to certain manuscript sections (Abstract, Results, Introduction, Methods) where different levels of similarity may be tolerated?
Ease of use
How easy is it to get started?
Can I do a test run?
Is the automation really working well?
Is this helping me? Is it worth it?
Cost and contract terms
What is the cost? How is the cost computed, unlimited use, or other?
Do I have an annual fee?
What about free services?
Stability, history, and reputation of the supplier
How long has the company or service been in business?
Can they provide a customer reference list?
Use and persistence of your query data
What happens to my query after I submit it?
Is my query deleted or become a permanent part of the search provider's database?
Who owns the results?

CrossCheck

CrossCheck is the service provided by the not-for-profit membership based organization, CrossRef, who originally developed the Digital Object Identifier (DOI), which is a reference linking service that provides persistence and linkage for citations. This organization has become a re-seller of the iParadigm's tool, IThenticate, offering it through a membership plus a fee per use financial model. This organization, experienced and knowledgeable of the publication industry, did not develop its own system, but does offer an alternative cost model for the user for the IThenticate services.

IThenticate

IThenticate is a service offered by IParadigms, the same company that has produced the very successful Turnitin

Table 2
Sampling of free and paid plagiarism detection services

Company/organization	Product	Cost
CrossRef.org	Crosscheck (powered by iThenticate)	Annual membership plus a per document fee
eTBLAST.org	eTBLAST, déjà vu	Free
iParadigms	iThenticate	Various, per document fee
Applied Linguistics	Grarmmarly	Membership fee (although advertized as free
Plagiarism-Checkers	CheckForPlagiarism.ne	Annual subscription fee
Indigo Stream Technologies	Copyscape	Free searches against web, Premium service has a fee per submission

plagiarism detection software for use by teachers and professors. The IThenticate product (presumably) has the same proprietary similarity and search engine as Turnitin, but has different (or more) target databases of literature against which they compare a query. Search and detection services offered to publication stakeholders are available, as mentioned above, from CrossCheck, but other purchase models are available directly from iParadigms.

eTBLAST

eTBLAST is a free service offered now by the Virginia Bioinformatics Institute and supports several databases, including Medline and arXiv citations, and publically available full text. This software service was originally designed as a text analytics software package for reference finding, but it has added benefits offered to the publication stake-

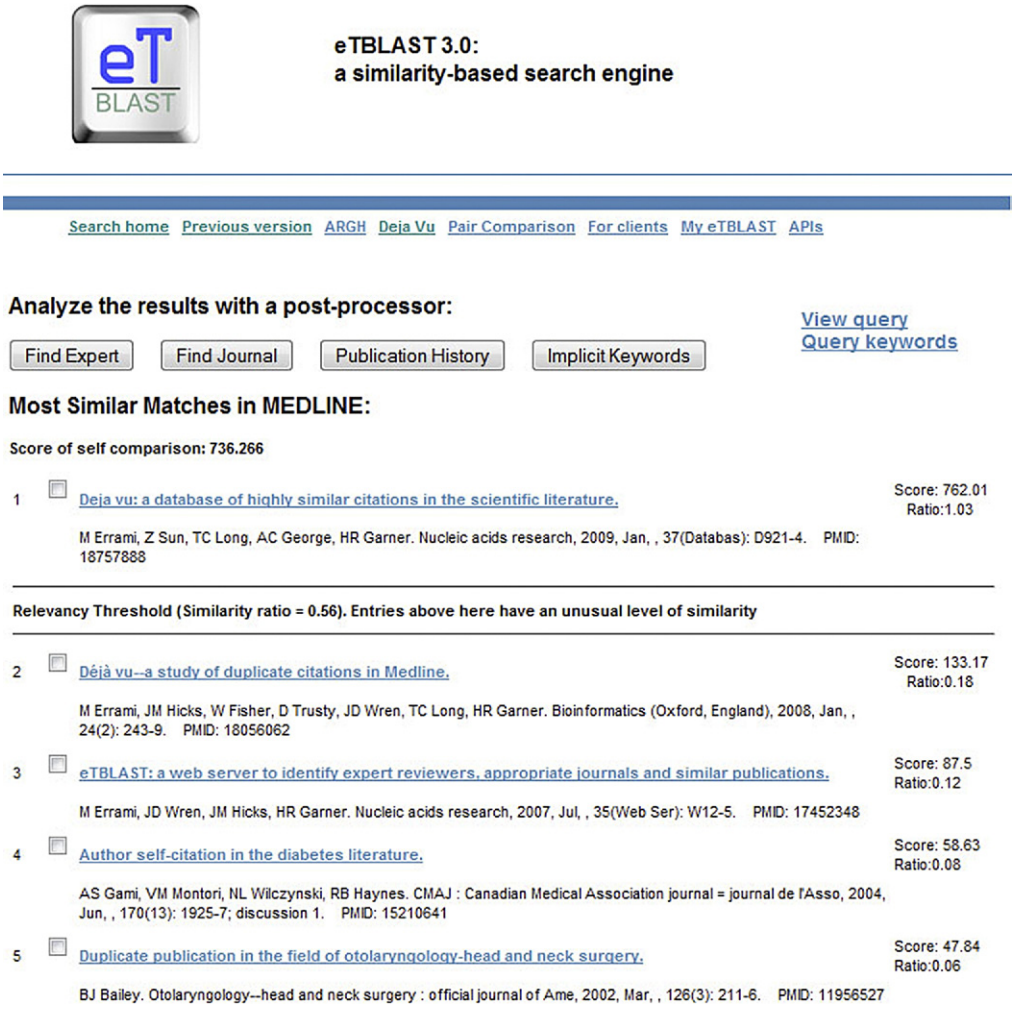

Fig. 2. Sample output from eTBLAST. In this example, an abstract was retrieved from Medline for a paper that was previously published and submitted to eTBLAST. That abstract had 180 total words, 96 of which were keywords, and it took 16 seconds to 18,941,414 other similar citations in Medline. This example was used to illustrate the output from this engine, which provides a list of citations ranked by level of similarity. Because this query was identical to an existing entry in Medline, it ranked first. In addition, eTBLAST delineated it from the rest because the similarity was greater than 56%, a threshold that was calibrated and reported as suspiciously similar. (Color version of figure is available online.)

Fig. 3. Clicking on the link of the highest ranked entry in the output presented in Fig. 2 opens up a page where the words that are similar to the query are shaded. This enables the user to quickly determine if further checking of the full text of the manuscript should be done. A link to the original entry in PubMed is provided, and if this paper was an Open Access publication, as it is, the full text for the paper is available to compare with the query. Please note that across the top are a series of other links, and in particular, the Pair Comparison link provides the ability for a user to put in text from 2 suspiciously similar sources and then view a comparison, demarked as was done in this figure. (Color version of figure is available online.)

holders, including the ability to suggest experts as possible reviews and alternative journals for publication. As illustrative examples of the types of output provided by plagiarism detection services, output from the eTBLAST service are shown in Figs. 2 and 3. It is also the engine used to identify highly similar pairs of citations in Medline that have been deposited into the on-line database, Déjà vu, which has become a resource for ethics and sociological studies as well as a teaching-by-example tool.

On a final note when selecting a plagiarism detector, there are some features or limitations that potential users may want to consider. Some examples include, when using eTBLAST, it has the advantages of being free, but it is a service provided by a university, and although care has been taken to make sure user data are as secure as possible, including the destruction of user queries after the analysis is complete, the user assumes full responsibility for its use. On the other hand, the model for Turnitin (and presumably, IThenticate, although it is not clear in their documentation) is to keep all queries and add them to their database, so even submissions rejected for reasons other than plagiarism are still kept, and may show up in future queries. There have been lawsuits over this filed on copyright infringement grounds.

Comparing pairs of documents, regardless of the original method used to 'detect' them

Independent of the method used to identify 2 documents that may be similar, the comparison of those documents can be done by eye or the comparison can be aided by software. This can greatly speed the process and make the results more accurate and quantitative. There are at least 2 approaches that can be used by publication stakeholders. The first is the "Pair Comparison" feature of eTBLAST. This simple comparison system is used by pasting in 2 sets of text into the web (select "Pair Comparison" link at http://etblast.org). A quantitative measure of the similarity and a graphic similar to the presentation in Fig. 4 is presented as output. The second approach is to use a feature in Microsoft Office Word 2007 to compare documents. This simple approach is exploited through the "Compare two versions of a document" tab under the "Review" tab. After opening two documents, several panes or used to show the user the overlap between the 2 documents.

The last word—cleaning up the corpus

The business model of the commercial and not-for-profit companies is to provide plagiarism detection services, and stay away from identifying existing highly similar or plagiarized documents within the scientific corpus. There have been some attempts to identify such documents; however, it is clear that there remain many unidentified documents that may have ethical issues. An even bigger issue is that those documents continue to be unwittingly used by professionals to make scientific, even clinical decisions. Even after questionable documents have been identified, judged, and retracted, that retraction notice may never propagate back to the indexing