Reviewer's report

Title: Locating previously unknown patterns in data-mining results: a dual data- and knowledge-mining method

Version: 2 Date: 29 August 2005

Reviewer: Howard Hamilton

Reviewer's report:

General

This research could form the basis for a published paper. The approach is somewhat novel. The application including the online database of medical articles seems novel. The thoroughness of the analysis of the results seems inadequate. In order for the results not to appear trivial, more analysis is needed before publication.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

The reference to a “knowledge base” and “ontology” seem misleading here. The actual database used is in the first case a set of titles of journal articles and in the more detailed case the content of selected articles. Perhaps the term "publication database" would be more appropriate.

The term ontology is used in the paper and the algorithm. This term “ontology” is not defined in the paper, but in the algorithm the authors make the assumption that two ontologies can be intersected to yield a list of attributes. This seems to imply that the term ontology as used here refers to a list of attributes. Indeed, as implemented, it seems that an ontology is in fact a list of terms that are then treated as binary attributes. This issue needs to be clarified. Use the terms that are the most descriptive of what was actually done.

p.6, The authors choose the r_i measures based on LOR for their work. The choice of this measure must be justified. At present, it is simply stated that there are many measures and r_i is chosen.


p.7 “one may normalize SS”. Given the previous discussion, it seems odd that normalizing SS is optional. Does SS have any meaning if the data are not normalized? Perhaps a test could be required on the scores and then normalization required depending on the results of the test.

p.7 “one may normalize the two scores”. Given the previous discussion, it seems odd that normalizing the scores is optional. Does SS have any meaning if the scores are not normalized?

Why was 350 chosen? Did it represent all disease entities for some lab results or what? As given, the use of 350 results seems to be arbitrary.
The analysis of the results is inadequate. From 350 potential results, it appears that only three were examined in sufficient detail to determine whether anything of interest was found. It seems that investigating all 350 or at least all of them where the SS score was high should be performed so that the effectiveness of the method could be assessed. I estimate that it would only take a few days to determine the results for 350. Perhaps a table of the top 30 could be presented with the results of analyzing the sentence level KB information. As it is, we have only the anecdotal information that one out of three possibly indicated an unreported relationship. Was it also one out of 30 and one out of 350?

---

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

Algorithm (Table 1): Give the algorithm a descriptive name. The “dual mining method” is inadequately descriptive.

Step 1 of the algorithm. Mention the choice of pattern type. “Choose a type of pattern, a mining method for finding that type of pattern, ...”

---

Discretionary Revisions (which the author can choose to ignore)

p.11 “red points”. Will the figures be printed in color? If not, remove the reference to “red” and use solid black points.

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

I declare that I have no competing interests