Review of: "A method to reduce false positives in a patent query"

Patrick Juola[1]

1 Duquesne University

Potential competing interests: No potential competing interests to declare.

This paper discusses the issue of finding patents relevant to a given technology. This, of course, is a mandatory step if you are getting a patent, but (as the authors point out) is also useful for various tasks in the study of technology. Most nations provide the ability to search patents electronically (e.g. uspto.gov), but they're usually limited to keyword search. Since there are lots of technologies and patents that could POSSIBLY be associated with a given domain (the authors provide several examples), the accuracy of simple inclusion/exclusion criteria is likely to be low. By using classification networks, the authors propose to identify entire patent categories that "can be excluded without risking removing false positives."

In section 2, the authors explore the reasons for false positives, including synonyms (which generally create false negatives, instead), homonyms, bilingual "homophones" [sic], paraphrase, and acronyms. This section contains several errors in linguistics and is largely unnecessary; that false positives exist is unquestionable and the reasons are, for this paper, not important.

The central idea of this patent is that many patents include several different classes in their metadata (figure 1 presents a patent header with nine categories). This can be used to create a network where patent classes are nodes, and a patent that encompasses two classes creates a link between the corresponding nodes (see fig 2). Prior work has shown that the resulting network contains a "technological core" and may contain clusters that represent aspects of the technology.

The authors propose, firstly, that only the component containing the core is relevant and that none of the other (unconnected) components will contain relevant patents. This is an extremely strong claim.

Using standard network cluster analysis technique, they identify unlabelled "communities" that can be labelled by humans. If the human judges that the community label is not relevant to the domain of interest, then the entire community is not relevant and will not contain relevant patents. Q.e.d.

The basic idea is sound, but the authors make too strong a case. There are no experimental results presented to confirm their claims, and no measurements of the actual false positive rate achieved. Furthermore, just as the initial query may return false positives, it is possible (even likely) that the various components and clusters may themselves have their own versions of misclassified items. A more realistic analysis would take into account the clustering accuracy and the likelihood that the clusters and communities are not themselves homogenous. As the paper stands, it's a good idea but

badly in need of empirical validation and measurement (especially on realistic-scale problems).