Sluban and Lavrač *BMC Bioinformatics* 2010, **11**(Suppl 5):P2
http://www.biomedcentral.com/1471-2105/11/S5/P2

BMC Bioinformatics

POSTER PRESENTATION Open Access

Supporting the search for cross-context links by outlier detection methods

Borut Sluban[1*], Nada Lavrač[2,3]

From Workshop on Advances in Bio Text Mining
Ghent, Belgium. 10-11 May 2010

Background and relation to previous work

Outliers in data can either present noise in the data, which has harmful effects on knowledge discovery (and should therefore best be eliminated), or correct data instances that belong to a specific subconcept of the main domain concept (and can potentially carry new interesting insights). Several outlier detection methods have been developed in data and text mining, mainly used for noise filtering and error detection purposes. Except for [1], outlier detection in text mining has not yet been used for exploratory purposes. Our work focuses on using noise/outlier detection methods for a novel task of cross-context link discovery.

Outlier detection through class noise filtering methods

This work uses a class noise detection approach for finding outlier documents which include bridging terms, linking different contexts/domains. It has been shown in [1] that detecting interesting outliers that appear in the literature on a given phenomenon can help the expert to find implicit relationships among concepts of different domains. In our approach we searched for a set of outlier documents using a class noise filtering approach [2] implemented with three different learning algorithms: *Naïve Bayes* (abbreviated: Bayes), *Support Vector Machine* (SVM) and *Random Forest* (RF). These outlier detection methods work in a 10-fold cross-validation manner, where repeatedly nine folds are used for training the classifier and on the complementary fold the misclassified instances are denoted as noise/outliers (of the domain they belong to).

Testing of the methods

To evaluate the relevance of the detected outlier documents for containing context bridging terms, we used the Swanson's Migraine-Magnesium dataset [3] obtained by searching PubMed for documents including these two keywords (after preprocessing resulting in 7,930 articles). We inspected 20 bridging terms appearing in the given preprocessed Migraine-Magnesium domain pair (i.e., 20 out of 42 known bridging terms identified in [3]). We compared their relative frequencies in the detected outlier document sets to their relative frequencies in the whole dataset. The three class noise filters implemented with different classifiers, Bayes, SVM and RF, found 765, 416, and 763 outliers, respectively. In these three sets of outlying documents, 17 (Bayes), 13 (SVM) and 17 (RF) of the 20 bridging terms are present. The relative frequencies of these terms in the sets of detected outlying documents are presented in Figure 1. For instance, the frequency of the term "vasospasm" (t6) in the set of outlier documents detected by the SVM-based class noise filter is 0.007, compared to its frequency 0.001 in the whole dataset. These results show that nearly all the bridging terms present in the sets of outlying documents have higher relative frequencies in these sets compared to the whole dataset.

Conclusion

This work shows that outlier detection methods can shorten the time needed for searching for bridging terms in cross-context link discovery, since the bridging terms are more frequent in sets of outlier documents.

* Correspondence: borut.sluban@ijs.si
[1]Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
Full list of author information is available at the end of the article

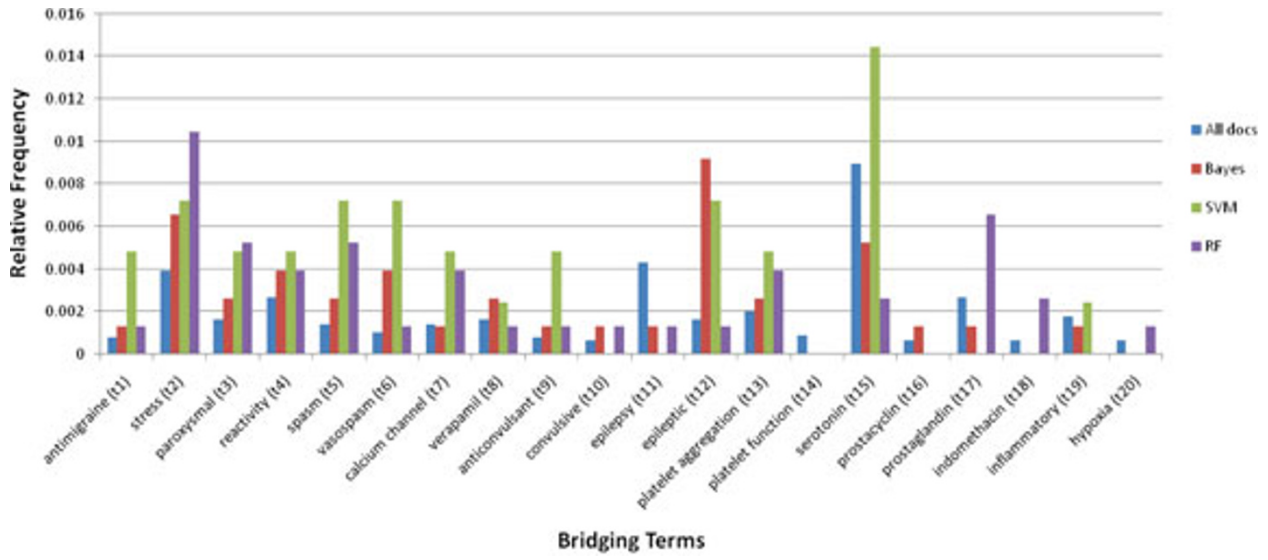

Figure 1 A comparison of relative frequencies of bridging terms (t1, ... , t20) in the following document sets: the entire dataset and in the sets of outlier documents detected by three different outlier detection methods.

Acknowledgement
This work was partially supported by the national project Knowledge Technologies and by the EU project FP7-211898 BISON.

Author details
[1]Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia. [2]Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. [3]University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia.

Published: 6 October 2010

References

1. Petrič I, Urbančič T, Cestnik B, Macedoni-Lukšič M: **Literature mining method RaJoLink for uncovering relations between biomedical concepts.** *J. Biomed. Inform.* 2009, **42(2)**:219-227.
2. Brodley CE, Friedl MA: **Identifying mislabeled training data.** *Journal of Artificial Intelligence Research* 1999, **11**:131-167.
3. Swanson DR: **Migraine and magnesium: eleven neglected connections.** *Perspectives in Biology and Medicine* 1988, **31(4)**:526-557.

doi:10.1186/1471-2105-11-S5-P2
Cite this article as: Sluban and Lavrač: **Supporting the search for cross-context links by outlier detection methods.** *BMC Bioinformatics* 2010 **11** (Suppl 5):P2.