Enhancing the Accuracy of Knowledge Discovery: A Supervised Learning Method

Liangxi Cheng, Hongfei Lin, Feng Zhou, Zhihao Yang
Information Retrieval Laboratory, School of Computer Science and Technology
Dalian University of Technology, Dalian, China
liangxicheng@mail.dlut.edu.cn, hflin@dlut.edu.cn, irlab@dlut.edu.cn, yangzh@dlut.edu.cn

I. INTRODUCTION

Given that the amount of biomedical literature available is growing at an explosive speed and people’s ability to read literature is limited, the development of a computer-assisted approach for mining central but hidden knowledge from extensive literature is of utmost importance.

II. METHODS

We use the open discovery defined by Weeber et al. [1] to replicate Swanson’s discoveries on Raynaud Disease-Fish Oils and Migraine-Magnesium with starting concepts of Raynaud Disease and Migraine, respectively.

For a given starting concept C, we collect all MeSH terms co-occurring with C in MEDLINE records within the specified time range, and select linking MeSH terms using both statistical and textual features. Mutual Information Measure (MIM) [2] is employed to quantify dependency between the linking MeSH term and C: I. MIM score of them calculated for the MeSH field; II. MIM score of them calculated for the title and abstract fields. In addition, we select five boolean features to evaluate the degree of connection between the MeSH term and C: III. Whether or not they co-occur in one sentence; IV. Whether or not there is an interaction word in the sentence; V. Whether or not they co-occur in two neighboring sentences; VI. Whether or not there is an interaction word in the two neighboring sentences; VII. Whether or not they only co-occur in the MEDLINE abstract field.

Candidate linking concept is represented by the 7 features, and an SVM classifier is used to classify it as positive or negative. Positive ones will be used further to discover target concepts. We search the MeSH field of MEDLINE records for target concepts that co-occur with positive linking concepts. Target concepts appearing in the set of linking concepts are removed and only those belonging to the semantic types Element, Ion, or Isotope, Vitamin and Lipid are retained.

III. EXPERIMENTAL RESULTS AND DISCUSSION

We adopted the linking term count (LTC) of Pratt and Yetisgen-Yildiz [3] to rank target concepts, and the linking terms that count include: 1. linking concepts co-occurring with the target concept in the title and abstract fields (Rule 1); 2. linking concepts co-occurring with the target concept and an interaction word in one sentence (Rule 2).

A. Performances of Features and Combinations

Linking concepts leading to the target concepts in the MeSH field are considered useful linking terms for ranking target concepts. The percentage of useful LTC (puLTC) among all the linking concepts is used to evaluate the performances of these features. From the experimental result we see that: 1. Feature II can better reflect the dependency of two concepts than Feature I; 2. Textual features improve the result compared with merely statistical features; 3. Interaction words contribute to the performance of the puLTC.

B. Raynaud Disease-Fish Oils

The result shows that the ranking of Fish Oils and the puLTC obtained using positive concepts are much higher than those using the whole linking concepts. Further filtering using the semantic types pushes the ranking of Fish Oils even higher. The Rule 2 is stricter so that in the result the ranking of Fish Oils increases and the puLTC decreases with it. The target concept Fish Oils ranked 19 under the semantic type Lipid in Srinivasan’s method [4]; in our experiment, it ranked 16.

C. Migraine-Magnesium

In the Migraine-Magnesium experiment, the ranking of Magnesium is relatively high. With the filtering of positive linking concepts and semantic type, both the ranking of Magnesium and the puLTC improve. Among the target concepts discovered at the sentence level, Magnesium obtains a rank of number one with an LTC of 39 in our experiment, compared to the best rank of 11 with an LTC of 29 in Pratt and Yetisgen-Yildiz’s method [3].

IV. CONCLUSION AND FUTURE WORK

In this paper we propose a new method for choosing useful and promising linking concepts. The results show that the ranking of potentially relevant target concepts is promoted by using relevant linking concepts and the puLTC is significantly improved. In future work, we aim to find more useful linking concept features to gain better results.

REFERENCES

[1] M. Weeber, H. Klein, and de Jong-van den Berg LTW, “Using concepts in literature based discovery: simulating Swanson’s Raynaud-fish oil and migraine-magnesium examples,” J Am Soc Inf Sci Tec 2001;52(7):548-557.
[2] J.D. Wren, “Extending the mutual information measure to rank inferred literature relationships,” BMC Informatics 2004;5(1):145-158.
[3] W. Pratt and M. Yetisgen-Yildiz, “LitLinker: capturing connections across the biomedical literature,” In: K-Cap’03, Sanibel Island, FL; Oct. 23-25, 2003. pp. 105-112.
[4] P. Srinivasan, “Text mining: Generating hypotheses from MEDLINE,” J Am Soc Inf Sci Tec 2004;55(4):396-413.