Integrated access to disease information: the PubMed disease sensor

Aurélie Névéol, PhD, George C. Jiang, PhD, Zhiyong Lu, PhD
U.S. National Library of Medicine, NIH Bethesda, MD 20894
(neveola,jianggc)@ncbi.nlm.nih.gov; Zhiyong.lu@nih.gov

Abstract
A recent usage log analysis showed that disease information is frequently sought by PubMed users. While many trustworthy resources besides PubMed provide valuable disease-related information, integrated access to information on the same disease in multiple sources is limited. In response, we developed automated methods for mapping disease mentions in free text to UMLS disease concepts. We apply this research in PubMed's disease sensor, which is currently triggered for 5% of PubMed users.

Introduction and objective
For common query topics such as disease [1], there is a variety of trustworthy sources that currently need to be accessed independently. To better assist users searching disease-related information in one database it would be helpful to automatically identify other sources relevant to their need and to provide direct access to them. To this end, disease-related information needs must be identified (disease recognition in queries) and matched to relevant sources (disease recognition in documents).

Methods and Results
The main challenges of disease recognition include a proper definition of the diseases of interest, and the ability to address the high language variation in natural language used to refer to a given disease. In a preliminary study [2], we defined “diseases” as UMLS® (Unified Medical Language System®) concepts in the Semantic Group disorder and explored statistical and natural language processing methods for disease recognition. Based on this work, we developed a large dictionary containing 448,708 terms linked to 157,641 UMLS concepts and used it to map disease relevant documents such as Gene Reviews to UMLS concepts. Figure 1 presents an application of this work in the PubMed® disease sensor, where this dictionary is used to identify disease terms in PubMed queries in real-time. As shown in Figure 1, the use of concept level disease recognition enables the linking of queries to relevant documents even when the specific terms used in the query differ from those used in the document. This implementation of the disease sensor is currently shown to 5% of PubMed users and we expect it to be soon released to 100%. Future work will also address extending the pool of target documents to include sources such as A.D.A.M. encyclopaedia articles in addition to Gene Reviews.

Figure 1. Sample PubMed query with disease sensor results shown above default PubMed results

References