



System zur Unterstützung der Kuration der Toxikogenomik in der medizinischen Literatur

Dina Vishnyakova^{a, b}, Emilie Pasche^{a, b}, Christian Lovis^b und Patrick Ruch^{a, c}

^a Bibliomics and Text Mining (BiTeM) Group: <http://bitem.hesge.ch>

^b Medizinische Informatik Abteilung, Universität und Uniklinik Genf

^c Informationswissenschaft Abteilung, HES-SO/Fachhochschule Westschweiz

Diese Arbeit stellt das neuartige Textklassifizierungssystem für Toxikogenomik «ToxiCat» (Toxicogenomic Categorizer) vor.

Das Ziel des Systems ist es, die Biokuratoren bei der Priorisierung der Dokumente mit toxikogenomischen Informationen zu unterstützen. Hierbei wird durch die Integration von Chemikalien, Gen-, Protein- und Krankheitsdaten in der biomedizinischen Literatur eine verbesserte Identifikation von Expositions- und Krankheitsbiomarkern ermöglicht. Des Weiteren wird eine Verbesserung durch Mechanismen in chemischen Reaktionen und von komplexen Ätiologien in chronischen Krankheiten erreicht.

ToxiCat beschleunigt die Kuration der komparativen Toxikogenomik-Datenbank (Comparative Toxicogenomics Database, CTD) durch die Klassifikation und Priorisierung von biomedizinischen Dokumenten. Vereinfacht kann das Problem als eine binäre Klassifikation beschrieben werden, deren resultierende Relevanzwerte zur Sortierung der ausgewählten Artikel verwendet werden. Der entwickelte SVM-Klassifikator besteht aus drei Hauptkomponenten: einer Informationssuchmaschine für MEDLINE (EAGLi), einer biomedizinischen Eigennamenerkennung

basierend auf terminologischen Quellen und einem Genormalisierungsservice (NormaGene), der für eine frühere BioCreative-Kampagne entwickelt wurde und einem Ad-hoc-Schlüsselwortdetektor für Krankheiten und Chemikalien. Der Schlüsselwortdetektor basiert auf den kontrollierten Vokabeln von Chemikalien- und Krankheitsbezeichnungen unterstützt durch CTD.

Zusätzlich zur Hauptaufgabe bietet unser System dem Nutzer die Möglichkeit, durch automatische Generierung von Fragen für das Question-answering-System EAGLi mehr über detektierte Objekte zu erfahren.

ToxiCat, dessen Evaluierung mittels dem Task-I des BioCreative 2012 sehr gute Ergebnisse (MAP = 0.83) erzielte, ist im Internet durch ein benutzerfreundliches Interface zugänglich: <http://pingu.unige.ch:8080/Toxicat>.

Dina Vishnyakova
SIMED
Hôpitaux Universitaires de Genève
4, rue Gabrielle-Perret-Gentil
CH-1211 Geneva 14
[dina.vishnyakova\[at\]hcuge.ch](mailto:dina.vishnyakova[at]hcuge.ch)