

Daten in den Lebenswissenschaften

Vom Paper über Datenbanken zur integrierten Informationsquelle

Ralf Hofestädt, Klaus Kuhn, Wolfgang Müller, Can Türker

Dieser Workshop befasst sich mit Problemen, die bei der Datenhaltung biologischer und medizinischer Daten auftreten. Er hat zum Ziel, Vertreter der Datenbanken, des Information Retrieval und Vertreter der Bio- und Medizinischen Informatik zusammenzubringen. Die Organisatoren entstammen verschiedenen Fachgremien. Ralf Hofestädt ist Sprecher des FB Informatik in den Lebenswissenschaften in der GI. Klaus Kuhn ist Präsident der GMDS, Wolfgang Müller ist im Leitungsgremien der Fachgruppe Information Retrieval der GI, Can Türker im Leitungsgremium der Fachgruppe Mobilität und Mobile Informationssysteme. Alle Workshop-Organisatoren arbeiten im Gebiet der Lebenswissenschaften.

Innerhalb der Informatik in den Lebenswissenschaften gibt es Communities, die aus verschiedenen Motivationen an ähnlichen Fragestellungen arbeiten. Auf der einen Seite stehen die Informatiker, die sich von der Informatik ausgehend mit den spezifischen Problemen von Anwendungen in der Biologie und Medizin befassen. Auf der anderen Seite stehen Biologen und Mediziner, die aus ihrem Fachgebiet heraus sich auch mit Informatik - Fragestellungen befassen. Diese Gruppen miteinander besser zu vernetzen ist ein Ziel, das von den beteiligten Organisationen und auch mit dem Workshop verfolgt wird.

Biologie und Medizin teilen das gleiche, sehr schwierige Problem: Die Fülle der zusammengetragenen Information über das Verhalten der untersuchten biologischen Prozesse übersteigt das von Menschen fassbare Maß. Beispielsweise listet die Datenbank medizinischer Publikationen PubMed z.B. fast 800.000 Publikationen allein für das Jahr 2008. Eine Einschränkung auf Publikationen, die das Wort hepatocyte (Leberzelle) enthalten, ergibt immer noch ca. 3.500 Publikationen allein für das Jahr 2008, also weit mehr als ein Mensch, der noch eigene Forschung machen möchte, sinnvoll in einem Jahr lesen oder auch nur querlesen kann. Gleichzeitig bleiben die hohen Anforderungen an Wissenschaftler und Ingenieure erhalten: Ein Wissenschaftler bzw. Ingenieur soll auf vorherigen - insbesondere auch von anderen erzielten - Resultaten aufbauend neue Resultate erzeugen. Hierzu braucht er eine möglichst vollständige Kenntnis über bisher erzielte Forschungsergebnisse.

Ähnlich liegt es bei der Verwaltung von Daten, die in medizinischen Einrichtungen im Experiment oder in der Praxis auftreten: Eine für einzelne Menschen unfassbare Fülle von Daten soll so verwaltet werden, dass nicht nur Daten gespeichert, sondern das in ihnen vorliegende Wissen verfügbar wird.

Information-Retrieval-Systeme und Datenbanken können hier Abhilfe schaffen: Sie helfen bei der Suche nach interessanten Informationseinheiten. Die Probleme bei der Erstellung dieser Systeme sind vielfältig. Die hieraus resultierende Vielfalt von Werkzeugen bringt ein neues Problem mit sich: Das der Datenintegration, also der integrierten Bereitstellung von Daten aus mehreren Systemen zur weiteren Nutzung. Datenintegration lag auch den Einreichenden dieses Workshops besonders am Herzen.

Fast alle Einreichungen lassen sich zumindest auch aus dem Winkel der Datenintegration betrachten. Betrachten wir die Einreichungen aus einem anderen Blickwinkel, spannen sie jedoch einen weiten Bogen von der computerlinguistischen Arbeit über das Peer-to-Peer-System für experimentelle Daten bis zur Datenintegration für Patientendaten in der Medizin.

Schon die Lektüre der Arbeiten haben uns Organisatoren und Gutachtern viel Freude bereitet. Wir freuen uns auf eine lebhaftige Diskussion im Workshop.

Die Organisatoren dieses Workshops (alphabetisch)

Ralf Hofestädt
Klaus Kuhn
Wolfgang Müller (Kontakt)
Can Türker