HOCHSCHULE HEILBRONN

Hochschule für Technik Wirtschaft Informatik

Studiengang Electronic Business (EB)

Diplomarbeit (280000)

Intelligente Suche in der Pharmaforschung

vorgelegt bei

Professor Dr. Gröschel

von

Jochen Löhl

Matr.-Nr. 156745

im

WINTERSEMESTER 2005/2006

Inhaltsverzeichnis

Management Summary

Der Prozess der Entwicklung eines Arzneimittels beinhaltet alle Schritte, die notwendig sind, um einen Wirkstoff für ein bestimmtes Anwendungsgebiet zu identifizieren, dessen Wirkungen durch Tests an Zellkulturen, Tieren und Menschen zu untersuchen und eine Zulassung als Arzneimittel für ihn zu erhalten. Dieser Prozess lässt sich in zwei Hauptabschnitte unterteilen: Die präklinischen und die klinischen Studien.

Im Rahmen der präklinischen Entwicklung werden grundlegende Studien durchgeführt, bei denen durch die Kombination verschiedener chemischer Bausteine viel versprechende Substanzen für einen bestimmten Anwendungszweck gesucht werden. Die Wirksamkeit dieser Substanzen wird anschließend durch Tests an Tieren und Zellkulturen untersucht. Zum Abschluss der präklinischen Studien werden pharmakologische und toxikologische Untersuchungen durchgeführt, deren Ergebnisse darüber entscheiden, ob der entwickelte Wirkstoff Menschen verabreicht werden kann.

Waren die präklinischen Studien erfolgreich, folgen die klinischen Studien. Diese werden ihrerseits in vier Phasen unterteilt. In den ersten drei Phasen werden Wirkungen und Nebenwirkungen eines Wirkstoffs durch seine Verabreichung an menschliche Probanden überprüft. Treten dabei die erwünschten Wirkungen ein und halten sich unerwünschte in Grenzen, kann der Wirkstoff als Arzneimittel zugelassen werden. Nach erfolgter Zulassung als Arzneimittel werden in der vierten Phase der klinischen Studien Langzeittests durchgeführt, um beispielsweise langfristig auftretende Nebenwirkungen zu identifizieren.

Die Entwicklung eines Arzneimittels ist für ein Unternehmen der Pharmabranche mit großen finanziellen Risiken behaftet. Insgesamt kann die Entwicklungsdauer bis zu 15 Jahre betragen, wobei Kosten zwischen 250 und 500 Millionen Euro anfallen können. Hinzu kommt, dass nur ein geringer Prozentsatz der zu Beginn der präklinischen Studien betrachteten Substanzen später als Arzneimittel zugelassen wird. Selbst bei 80 Prozent der Wirkstoffe, für die klinische Studien angesetzt werden, wird die Entwicklung später abgebrochen. Deshalb muss ein erfolgreich zugelassener Wirkstoff durch den durch ihn erzielten Umsatz sowohl die Kosten für seine eigene Entwicklung als auch die Kosten fehlgeschlagener Entwicklungen decken. Zudem bleibt dem Pionierunternehmen nur ein kurzer Zeitraum, in dem es den Wirkstoff alleine vertreiben darf, da der Patentschutz bereits zu einem frühen Zeitpunkt beantragt werden muss und erst nach der Zulassung als Arzneimittel Umsätze erzielt werden können.

Nach Ablauf des Patentschutzes drängen meist Generikahersteller auf den Markt, die einen identischen Wirkstoff mit einem lediglich geringen Entwicklungsaufwand anbieten.
Somit ist es für Pharmaunternehmen wichtig, Arzneimittel zu entwickeln, die einerseits über einen lückenlosen Patentschutz, andererseits auch über ein ausreichendes Marktpotential verfügen, um zu einem wirtschaftlichen Erfolg werden zu können. Dazu werden die Forschungs- und Entwicklungsaktivitäten von Wettbewerbern und allgemeine technologische Trends betrachtet. Dies geschieht durch die Analyse von angemeldeten Patenten und anderen Quellen, z.B. Fachartikeln. Solche Texte sind aufgrund des Verbreitungsgrades elektronischer Medien, insbesondere des Internets, in großer Zahl erhältlich. Die Schwierigkeit für ein Pharmaunternehmen besteht dabei jedoch weniger in der Verfügbarkeit der Analysequellen als vielmehr in deren Auswertung. Da es sich hierbei meist um komplexe Sachverhalte handelt, die zudem in Form unstrukturierter Texte vorliegen, müssen alle Quellen durch menschliche Mitarbeiter gelesen und analysiert werden, was aufgrund der Vielzahl der Dokumente zu einem erheblichen zeitlichen Aufwand führt.
Durch Methoden des Textminings könnte die Analyse von Dokumenten in der Pharmabranche einfacher und effizienter gestaltet werden. Textmining ermöglicht das computergestützte Entdecken neuen Wissens in unstrukturierten Texten, wobei grundsätzliche drei Verfahren zum Einsatz kommen:

- Erschließung des Dokumenteninhalts
- Klassifikation von Dokumenteninhalten
- Strukturermittlung in Dokumentensammlungen

Um die Einsatzmöglichkeiten dieser Verfahren in der Pharmabranche zu überprüfen, wurde für diese Arbeit Software der TEMIS S.A. (**Te**xt**mi**ning **S**olutions), einem führenden Anbieter von Textmining-Software in Europa, betrachtet. Dabei wurde insbesondere der Insight Discoverer™ Extractor (IDE), ein Server für die Erschließung von Dokumenteninhalten, der Insight Discoverer™ Categorizer (IDK), ein Server für die Klassifikation von Dokumenten und der Insight Discoverer™ Clusterer (IDC), der eine automatisierte Strukturermittlung in Dokumenten ermöglicht, eingesetzt. Die Extraktion durch den IDE lässt sich dabei durch Regeln beeinflussen, die in so genannten Skill Cartridges™ gespeichert werden.
Die praktische Umsetzbarkeit wurde anhand zweier Prototypen untersucht, die im Rahmen dieser Diplomarbeit erstellt wurden. Der erste ist eine Applikation zur Extraktion von Informationen, die für Pharmaunternehmen relevant sind, aus unstrukturierten Textdokumenten. Diese besteht aus einer auf dem Framework Struts basierenden Web-Applikation, in die der Insight Discoverer™ Extractor über dessen Java-API eingebunden wurde. Durch Regeln, die

durch die in den IDE integrierte Competitive Intelligence Skill Cartridge™ für die Pharmabranche bereitgestellt werden, ist der IDE in der Lage, Informationen über Pharmaprodukte, Firmenzusammenschlüsse und –übernahmen, technologische und wirtschaftliche Partnerschaften, Aktivitäten von Regulierungsbehörden, klinische Studien, Gerichtsverfahren und andere wichtige Sachverhalte zu extrahieren. Diese Ergebnisse stehen anschließend in der Web-Applikation zur Verfügung und können dort weiterverarbeitet werden. Durch eine solche Verwendung des IDE kann der Rechercheaufwand in Pharmaunternehmen erheblich reduziert werden, da nun vor der Analyse neuer Dokumente durch Mitarbeiter automatisiert festgestellt werden kann, welche Art von Information in einem Dokument enthalten ist.

Bei der Einbindung zeigte sich, dass sich der IDE mit einem angemessenen Aufwand in eine bestehende Anwendung, beispielsweise ein Recherchesystem, integrieren lässt und sich die von ihm erzeugten Ergebnisse flexibel weiterverarbeiten lassen. Auch die Qualität der durch den Prototyp generierten Extraktionsergebnisse war zufrieden stellend. Zwar gab es teilweise Aussetzer oder Fehlinterpretationen, in der Mehrzahl der Testdokumente konnte der IDE jedoch die wichtigsten Informationen identifizieren.

Der zweite Prototyp dient zur Klassifikation von Dokumenten. Zu diesem Zweck wurden der Insight Discoverer™ Extractor und der Insight Discoverer™ Categorizer in eine auf Struts basierende Web-Applikation, ähnlich der des ersten Prototyps, integriert. Durch diesen Prototyp können Dokumente durch den IDK einer Kategorie, z.B. einer bestimmten Technologie oder einem Forschungsprojekt, zugeordnet werden. Dies reduziert den Aufwand der Analyse von Dokumenten durch Mitarbeiter, da nun nur diejenigen Dokumente betrachtet werden müssen, die einer Kategorie zugewiesen sind, die für die jeweilige Aufgabenstellung relevant ist. Hierzu werden zuerst die Merkmale von Trainingsdokumenten durch den IDE extrahiert. Anschließend werden diese Merkmale an den IDK übergeben, der daraus ein Modell generiert. In einem zweiten Schritt können neue Dokumente einer in diesem Modell definierten Kategorie zugeordnet werden. Dazu werden deren Merkmale wiederum durch den IDE extrahiert und anschließend an den IDK übergeben. Dieser kann sie nun anhand der Merkmale einer Kategorie zuordnen.

Auch bei der Erstellung dieses Prototyps zeigte sich, dass eine Einbindung der TEMIS-Software in eine bestehende Anwendung problemlos möglich ist und diese dadurch für verschiedene Anwendungszwecke genutzt werden kann. Durch den Prototyp war es möglich, eine Sammlung von Pharmapatentbeschreibungen im Rahmen einer Testreihe in die entsprechenden Kategorien einzuordnen.

Abschließend ließ sich feststellen, dass die untersuchten TEMIS-Produkte hervorragend für den Einsatz in Unternehmen der Pharmabranche geeignet sind und dazu beitragen können, den dort anfallenden Rechercheaufwand zu reduzieren.