

Unsicherheit - Vagheit

Probabilistic Information Retrieval

Die Wahrscheinlichkeitsrechnung wird in der Künstlichen Intelligenz zur Realisierung vieler Systeme benötigt: Sei es der Agent, der aufgrund von unsicherem oder ungenügendem Wissen rationale Entscheidungen treffen soll, oder das Diagnose-System, welches bestimmen soll, welche Krankheit der Patient hat.

Grundsätzlich wird zwischen zwei Arten von unvollständigem Wissen unterschieden:

Unsicherheit und Vagheit.

Im Falle von Unsicherheit (probabilistischer Ansatz) hat man eine bestimmte Aussage, die nur wahr oder falsch sein kann. Da die Information über diese Aussage unbekannt ist, ordnet man dieser Aussage - der Vermutung entsprechend - eine Wahrscheinlichkeit zu, mit der man sie für wahr hält (degree of belief). Für Agenten sind hierbei besonders die bedingten Wahrscheinlichkeiten interessant, die theoretisch als joint probability distribution dargestellt und praktisch mit dem Satz von Bayes berechnet werden.

Im Falle von Vagheit hat man eine bestimmte, sichere Information, die jedoch in Bezug auf eine unscharfe Aussage interpretiert werden soll. Man kann z.B. keine "Ja - Nein"- Entscheidung darüber treffen, ob ein Buch mit 300 Seiten dick ist. Darum ordnet man dieser Aussage ("Das Buch ist dick") einen Wahrheitsgrad zu, z.B. 0,8: Das Buch ist "ziemlich" dick (degree of truth). Dieser Ansatz ist unter dem Namen "Fuzzy Logic" bekannt und findet vor allem in der Steuerungstechnik Anwendung.

Eine weitere Anwendung für die Wahrscheinlichkeitsrechnung in der Künstlichen Intelligenz ist das Probabilistic Information Retrieval: Information Retrieval-Verfahren sind Verfahren zum Durchsuchen großer Informationsmengen. Zum Beispiel zum Durchsuchen des Internet, oder großer Wissensdatenbanken, werden sehr schnelle Verfahren benötigt, um Dokumente zu einem bestimmtem Thema zu finden.

Beim Probabilistic IR stellt man sich die zu durchsuchenden Dokumente als eine Menge vor: Die Grundidee ist es, dass eine bestimmte Teilmenge von Dokumenten jeweils eine bestimmte Abfrage ideal beantwortet. Genau diese Menge soll bestimmt werden.

Dazu berechnet man die Ähnlichkeit der Dokumente zu der Abfrage.

Die Ähnlichkeit ist ein Maß, wie gut ein Dokument zur Abfrage passt, und wird nur aus Wahrscheinlichkeiten berechnet.

Es konnte eine Formel gefunden werden, welche dieses Maß sehr schnell für alle Dokumente berechnen kann, wenn schon eine Näherung für die ideale Antwort auf die Abfrage vorhanden ist. So kann man durch mehrere Wiederholungen des Algorithmus die Lösung bestimmen.

Die Vorteile des Algorithmus sind das Erstellen eines Rankings, so dass das am besten passende Dokument zuerst angezeigt werden kann, sowie die Möglichkeit des Benutzereingriffs, um das Ergebnis weiter zu verbessern.

Außerdem liefert der Algorithmus auch ohne Benutzereingriff bessere Ergebnisse als klassische Methoden, und er basiert (im Gegensatz zum Vector-Model) auf nachvollziehbarer Mathematik.

Nachteile sind, dass nicht berücksichtigt wird wie oft ein Suchwort in einem Dokument vorkommt, und dass man das Ergebnis erst abschätzen muss.

Quellen:

- Stuart Russel, Peter Norvig. Artificial Intelligence : Modern Approach. Prentice Hall, 1st edition (January 15, 1995)
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Publishing Company (1999)