Distanzberechnung

Für die Berechnung der Distanzen zwischen Filmen wurden im Rahmen der Aufgabenstellung insgesamt drei Funktionen definiert und deren Ergebnisse im Programm *Cytoscape* grafisch dargestellt, um die Ergebnisse zu evaluieren. Als Datenbasis diente das erweiterte MovieLens-Dataset[1] von der Konferenz *HetRec 2011*, da dieses eine bessere Datenstruktur besitzt und mehr Metadaten zur Verfügung stellt als das Standard-Dataset[2]. Umgesetzt wurden die Distanzberechnungen in der Programmiersprache *PHP*. Nachfolgend werden die einzelnen Distanzfunktionen beschrieben und anschließend miteinander verglichen.

Als erste Distanzfunktion wurde die euklidische Distanz verwendet, die über die einzelnen Genres von Filmen berechnet wird. Die Genres werden in einem 20-dimensionalen Vektor repräsentiert, welcher für jedes Genre den Wert 100 enthält, wenn der Film diesem Genre zugeordnet wird und den Wert 0, wenn der Film dieses Genre nicht besitzt.

Funktion Nr. 2 (siehe Distanzfunktion Rating) verwendet die soeben beschriebene euklidische Distanz als Basis und lässt die Durchschnittsbewertungsdifferenz der Filme als Faktor mit einfließen. Die euklidische Distanz muss hierbei normalisiert und mit 1 addiert werden, da sonst die Distanz 0 beträgt, wenn nur die Genres übereinstimmen. Dasselbe geschieht für die Bewertungsdifferenz in der Formel, da auch hier 0 herauskommen würden, wenn zwei Filme dieselben Bewertungen besitzen.

Die letzte Distanzfunktion (siehe Distanzfunktion Custom) bezieht die euklidische Distanz, die Durchschnittsbewertungsdifferenz, den Regisseur, den Erscheinungsjahrunterschied, sowie das Produktionsland in die Berechnung mit ein. Der Jahresunterschied wurde durch 200 dividiert, damit er weniger Einfluss auf die Distanz hat. Die Faktoren für Regisseur- und Landübereinstimmung wurden ebenfalls frei gewählt und können angepasst werden.

Vergleich

Der erste Ansatz mit der reinen euklidischen Distanz lieferte einen sehr großen Cluster mit sehr wenigen Nebenclustern und erscheint somit sehr ungeeignet, da er den Graphen schlecht partitioniert. Die Bewertungs-Distanzfunktion lieferte bessere Ergebnisse, allerdings nur mit einem maximalen Kantengewicht von 200. Des Weiteren bildeten sich hier ausschließlich isolierte Cluster, die keine Verbindung zu den anderen Clustern aufwiesen. Die beste Distanzfunktion im gegenwärtigen Test war die angepasste Distanzfunktion, die mehr Metadaten miteinbezog. Mit einem maximalen Kantengewicht von 150 wurde der Graph in kleinere Cluster unterteilt, die aber auch Verbindungen untereinander aufwiesen. Es waren aber auch wenige kleine, isolierte Cluster vorhanden. Durch weitere Verfeinerung der Distanzfunktion könnten in weiterer Folge eine bessere Partitionierung des Graphen und bessere Vorschläge erzielt werden.

[1] http://www.grouplens.org/node/462#attachments
[2] http://www.grouplens.org/node/73#attachments

Die Ergebnisse der einzelnen Distanzfunktionen mit jeweils 1000 Filmen liegen im hochgeladenen ZIP-File bei, welches *Cytoscape*-Dateien und deren exportierte Graphen als Grafiken enthält.

Formelverzeichnis

Allgemeine Faktoren für die Distanzberechnung

$$ratingFactor(m_1, m_2) = 1 + \left|\frac{rating(m_1) - rating(m_2)}{5}\right|$$

$$yearFactor(m_1, m_2) = 1 + \left|\frac{year(m_1) - year(m_2)}{200}\right|$$

$$directorMatch(m_1, m_2) = \begin{cases} 0{,}75 \; if \; same \; director \\ 0 \end{cases}$$

$$countryMatch(m_1, m_2) = \begin{cases} 0{,}25 \; if \; same \; country \\ 0 \end{cases}$$

Distanzfunktion Rating

$$ratingDistance(m_1, m_2) = 100 * (1 + \frac{euclid(m_1, m_2)}{100}) * ratingFactor(m_1, m_2)$$

Distanzfunktion Custom

$$distanceCustom(m_1, m_2)$$

$$= \frac{100 * \left(1 + \frac{euclid(m_1, m_2)}{100}\right) * \; yearFactor(m_1, m_2)}{\left(1 + directorMatch(m_1, m_2) + countryMatch(m_1, m_2)\right) * \frac{1}{ratingFactor(m_1, m_2)}}$$

Legende

m_1……Movie 1

m_2……Movie 2