Advanced AI

学习课程

罗平

1-3

从度量相似度到协同过滤

相似度（Similarity）度量的问题

- 比较两张图片是否相似
- 比较两个文档是否相似
- 比较两段音乐是否相似
- 比较两个人的爱好品味是否相似
- 比较两句话是否语义相似
- …… ……

问题转化

- 将图片、文档、音乐、个人品味、短句转化为
  - 实数的向量
- 例子：表示语句“我爱你”的向量为

我	天空	…	爱	…	腾讯	…	你	…
1	0	…	1	…	0	…	1	…

  - 向量的维度为词汇集的大小

Cosine相似度

Given two vectors of attributes, A and B, the cosine similarity, $cos(\theta)$, is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

, where A_i and B_i are components of vector A and B respectively.

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality (decorrelation), and in-between values indicating intermediate similarity or dissimilarity.

Pearson Correlation

$$Pearson\_Sim(\mathbf{A},\mathbf{B}) = \frac{(\mathbf{A}-\overline{\mathbf{A}})(\mathbf{B}-\overline{\mathbf{B}})}{\|\mathbf{A}-\overline{\mathbf{A}}\|\|\mathbf{B}-\overline{\mathbf{B}}\|}$$

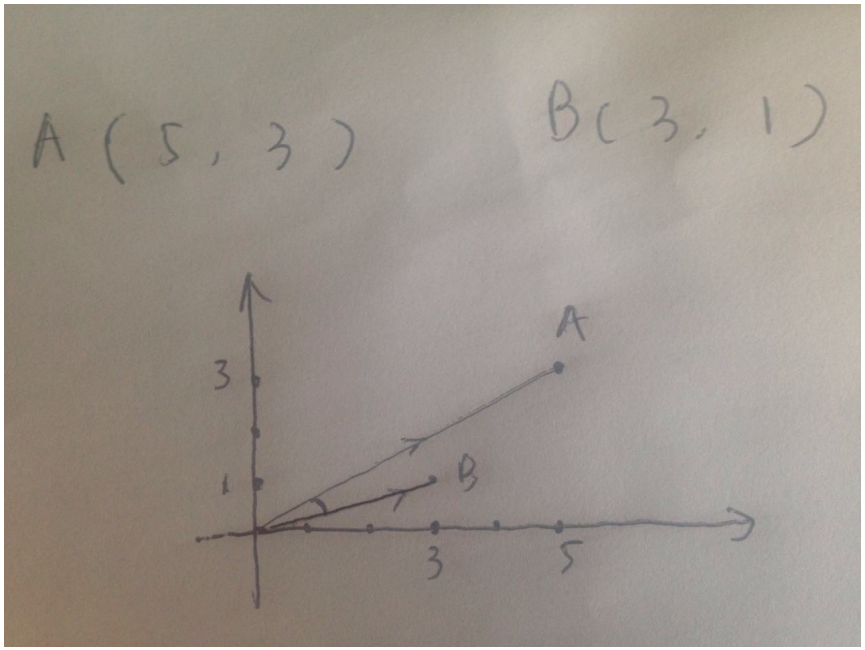
$$= \frac{\sum_i(\mathbf{A}_i-\overline{A})(\mathbf{B}_i-\overline{B})}{\sqrt{\sum_i(\mathbf{A}_i-\overline{A})^2}\sqrt{\sum_i(\mathbf{B}_i-\overline{B})^2}}$$

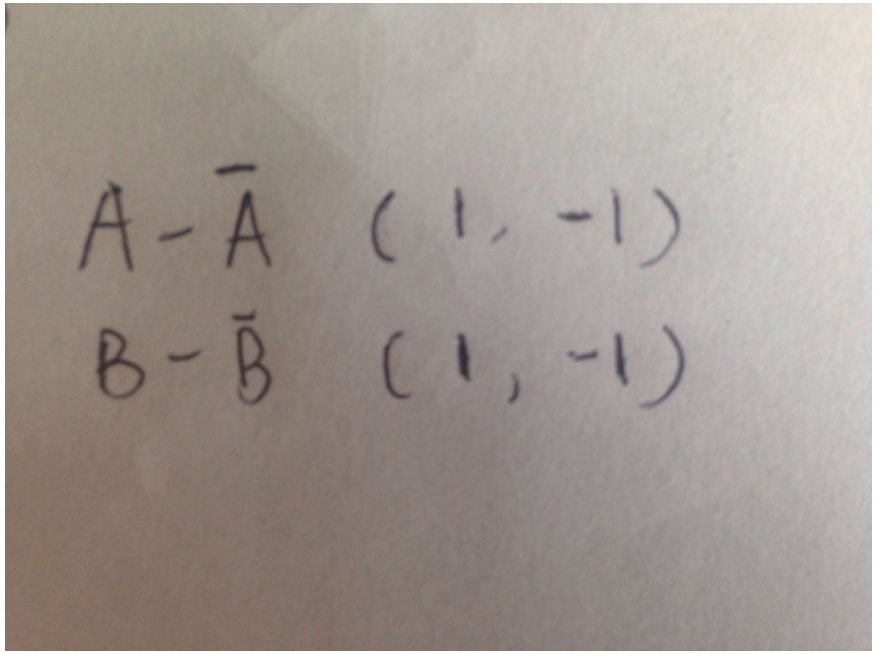
$$Cosine\_Sim(\mathbf{A},\mathbf{B}) = \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

$$Pearson\_Sim(\mathbf{A},\mathbf{B}) = Cosine\_Sim(\mathbf{A}-\overline{\mathbf{A}},\mathbf{B}-\overline{\mathbf{B}})$$

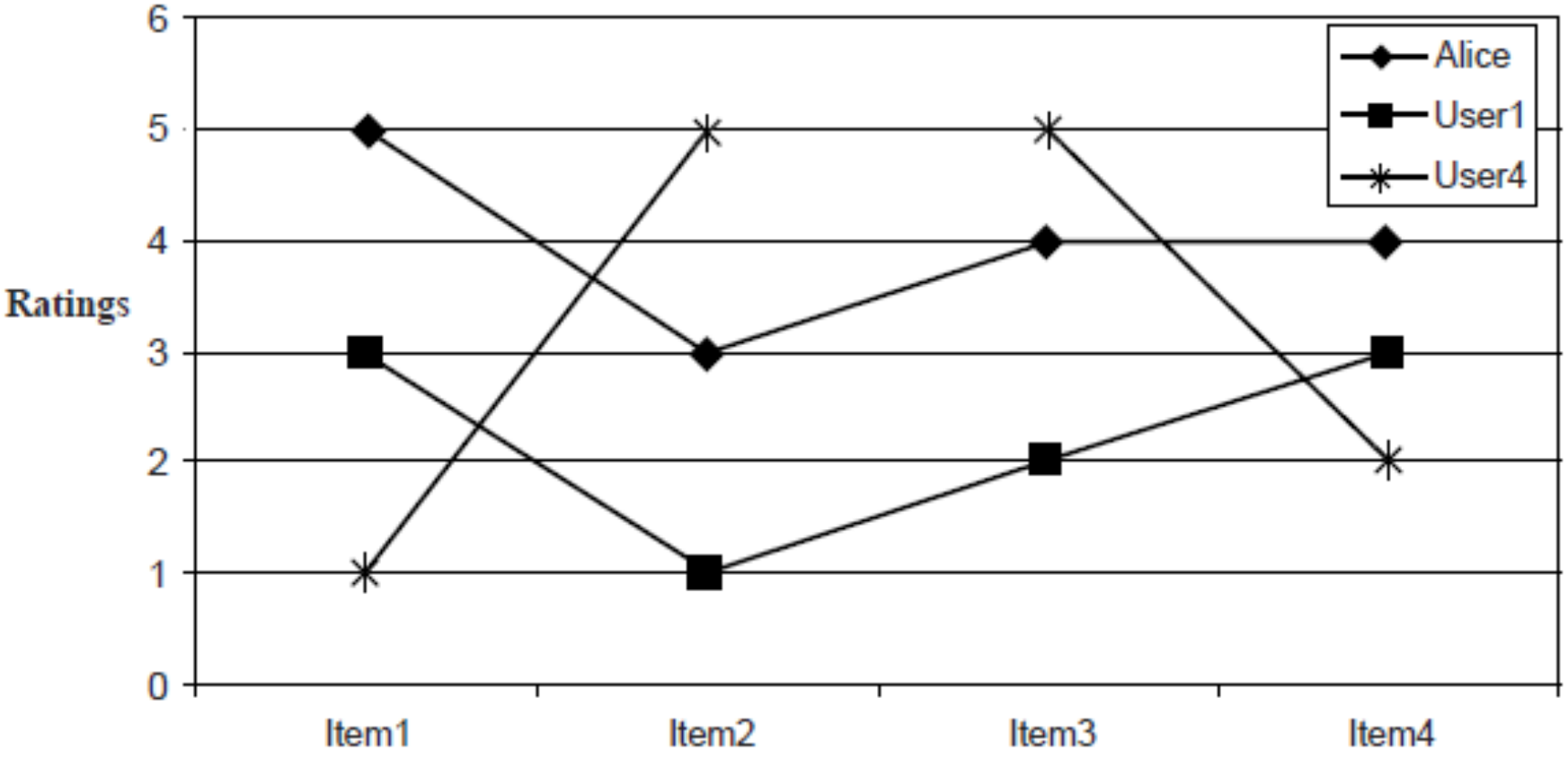
比较两种相似度

- A:(5,3)
- B:(3,1)



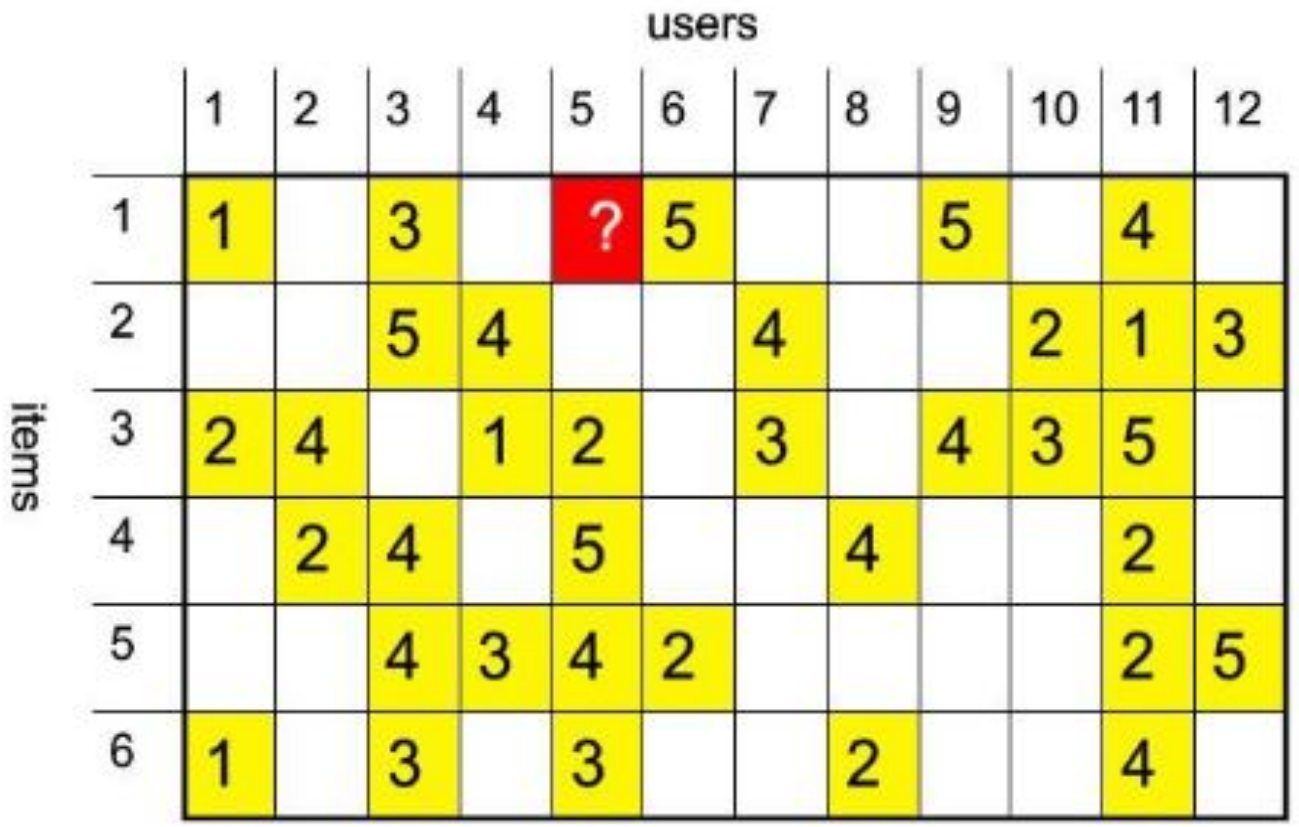


比较两种相似度



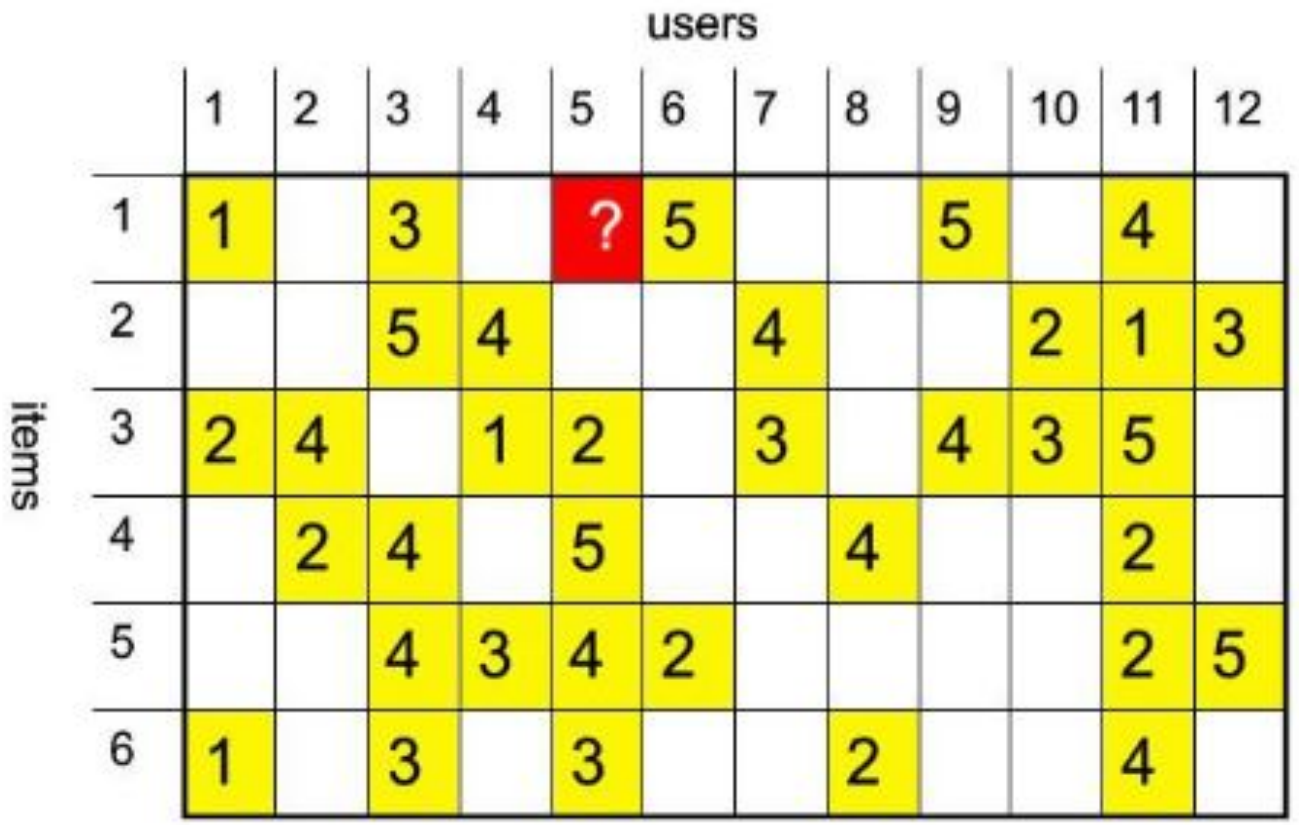
协同过滤（Collaborative Filtering）

- 推荐系统的最常用算法
- 评分矩阵



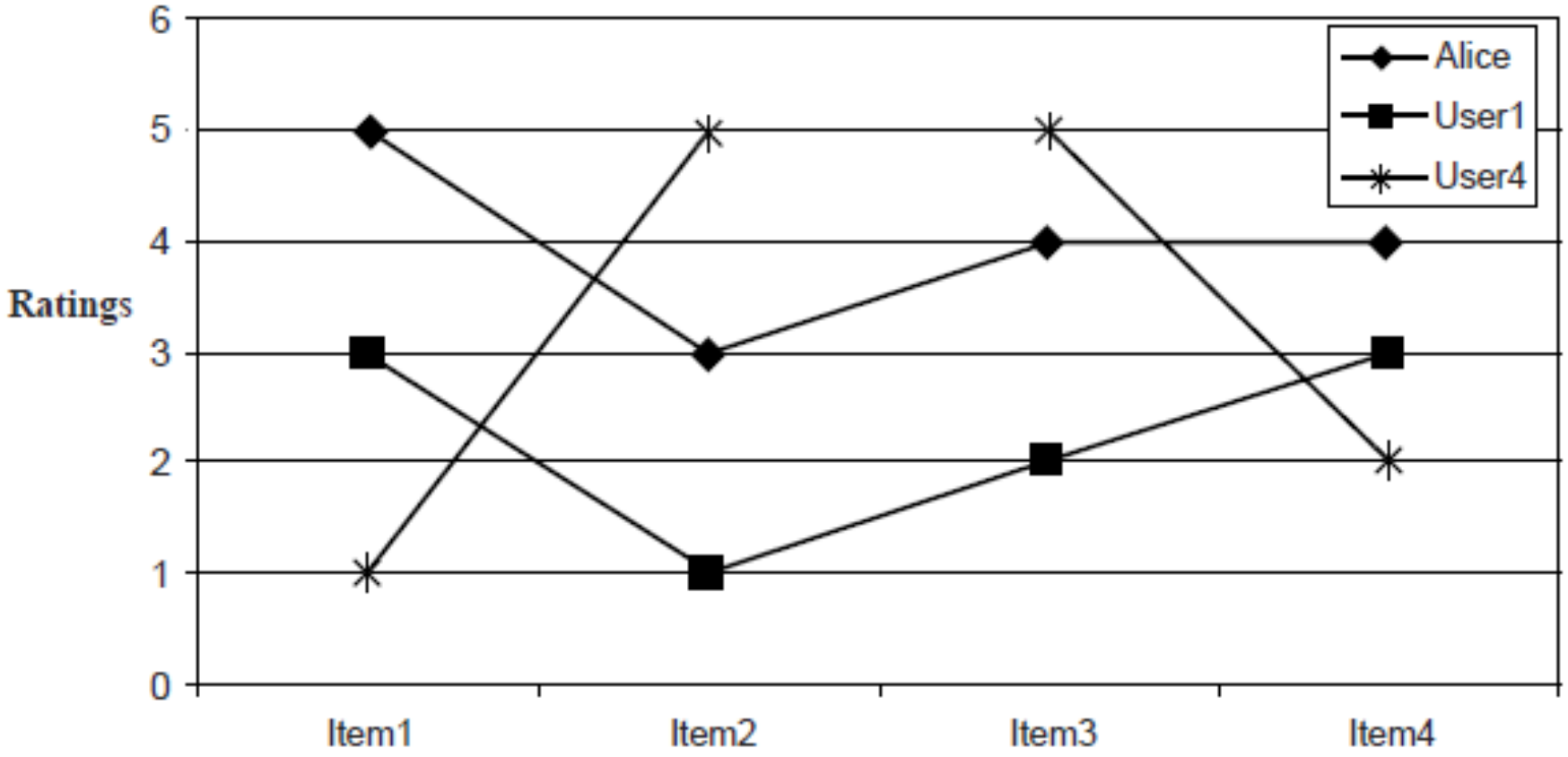
协同过滤（Collaborative Filtering）

- 核心思想
  - 找寻“臭味相投”的用户
  - 用这些相似用户的评分，进行预测



协同过滤（Collaborative Filtering）

- 用户品味的相似度计算
  - 保守的用户 vs. 乐施的用户 （评分的平均值不同）
- 使用Pearson Correlation



协同过滤：细节

- 问题：预测用户a在商品i上的评分

$$r_{a,i} = \frac{\sum_{u \in U_a} r_{u,i}}{n}$$

- 其中，U_a是跟用户a品味“最相似”的n个用户的集合
  - 使用**Pearson Correlation**

协同过滤：细节

- 问题：预测用户a在商品i上的评分

$$r_{a,i} = \frac{\sum_{u \in U_a} r_{u,i}}{n}$$

- 评分的归一化

$$r_{a,i} = \overline{r}_a + \frac{\sum_{u \in U_a} (r_{u,i} - \overline{r}_u)}{n}$$

协同过滤：细节

- 评分的归一化

$$r_{a,i} = \bar{r}_a + \frac{\sum_{u \in U_a} (r_{u,i} - \bar{r}_u)}{n}$$

- 加权平均

$$r_{a,i} = \bar{r}_a + \frac{\sum_{u \in U_a} w_{ua} (r_{u,i} - \bar{r}_u)}{\sum_{u \in U_a} w_{ua}}$$

  - w_{ua}是用户u和a之间的Pearson Correlation

协同过滤进级

- 已讲：用户-用户的协同过滤
- 其它方法
  - 商品-商品协同过滤
  - 基于矩阵分解的协同过滤

协同过滤方法的历史

- 1992: Information Tapestry, Doug Terry, Xerox Parc
  - 发明了“Collaborative Filtering”这个词汇
- Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. CSCW, 1994
- 2010, ACM Software System Award

In Memory of John Riedl

UNIVERSITY OF MINNESOTA
Driven to Discover℠

One Stop MyU : For Students, Faculty, and

COLLEGE OF Science & Engineering

MENU

In memoriam: John Riedl

John Riedl, a professor in the University of Minnesota's Department of Computer Science and Engineering and world-renowned expert in the field of recommender systems, died on July 15, 2013 after a three-year battle with cancer. He was 51.

A faculty member at the University of Minnesota since 1989, Riedl is known worldwide as a pioneer in the field of recommender systems—a field he was instrumental in creating and nurturing. Recommender systems are information filtering systems that seek to predict the "rating" or "preference" that users would give to an item (such as music, books, or movies) or social element (such as people or groups) based on previous choices.

The impact of Riedl's work is extensive, both in industry and among the research community. Software derived from his research is used by tens of thousands of

小结

- 概念: Cosine Similarity, Pearson Correlation
- 用户-用户的协同过滤方法