

Прикладные задачи анализа данных

Пост-троечные последовательности

Дьяконов А.Г.

**Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)**



Разработка рекомендательной системы

Международное соревнование «VideoLectures.Net Recommender System Challenge (ECML/PKDD Discovery Challenge 2011)»

<http://tunedit.org/challenge/VLNetChallenge?m=summary>

Опишем лучший алгоритм из 62

MIT WORLD SERIES
Creativity: The Mind, Machines, and Mathematics: Public Debate

author: Ray Kurzweil, Kurzweil Technologies, Inc.
author: David Gelernter, Yale University
author: Rodney A. Brooks, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, MIT
published: Dec. 16, 2011, recorded: November 2008, views: 165

Categories:
Top» Mathematics

Turn off the lights

See Also:
Streaming Video Help
Windows Media Player Firefox Plugin - Download

Related content:
See Also Personal history More by author

Visitors who watched this lecture also watched...

- Hilbert, Gödel, and Metamathematics today
800 views - Jeremy Avigad, 2011
- Alan Turing: Codebreaker and AI Pioneer
262 views - S. Jack Copeland, 2006
- Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind
651 views - Manish Mittal, 2007
- NLP at Google
2870 views - Katja Filippova, 2010
- What is cognitive science?
728 views - Jost T Siebmann, 2010
- Souramavid, the Movie

Lecture popularity: ☆☆☆☆ You need to login to cast your vote.
Мне нравится Одному пользователю это нравится. BOOKMARK

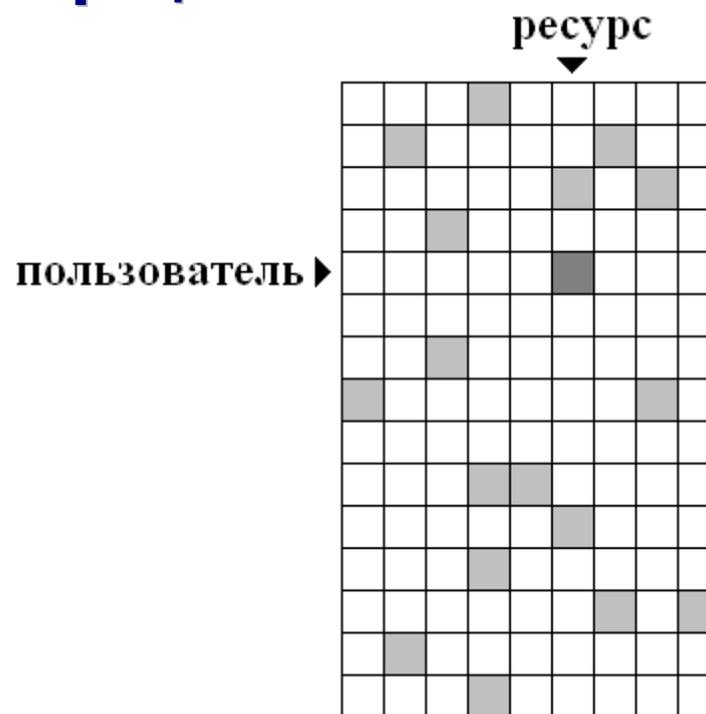
Description

Дано: статистика популярности (+ описания лекций)

Надо: дать рекомендацию пользователю

– предложить лекции для просмотра

Обычно: матрица «пользователи – ресурсы»



Методы коллаборативной фильтрации

~ похожие пользователи – похожие ресурсы

**Новое направление в анализе данных –
правильное обезличивание и усреднение**
Pooled sequences

Формирование пост-троечных последовательностей

$102 \rightarrow 33 \rightarrow 2 \rightarrow 34 \rightarrow 35 \rightarrow 2 \rightarrow 102 \rightarrow 17 \rightarrow 36,$

удаляем из неё повторы:

$102 \rightarrow 33 \rightarrow 2 \rightarrow 34 \rightarrow 35 \rightarrow 17 \rightarrow 36$

после тройки $\{2,33,35\}$ смотрел $\{17,36\}$.

$7 \times \{2, 33, 35\} : \quad 2 \times 9, \quad 5 \times 13, \quad 3 \times 17, \quad 1 \times 30, \quad 1 \times 36$

102	→	33	→	2	→	34	→	35	→	17	→	36
35	→	33	→	100	→	2	→	9	→	13	→	17
2	→	7	→	103								
2	→	35	→	33	→	13	→	9	→	17		
2	→	100	→	35	→	33	→	13	→	30		
100	→	2	→	35	→	7	→	33				
35	→	10	→	33	→	13						
33	→	107	→	2	→	35	→	13				
98	→	2	→	99	→	35	→	33	→	13		

Дано: некоторые пост-троечные последовательности (109044 шт.)

Найти: другие пост-троечные последовательности
(точнее: 10 первых членов в нужном порядке)

7x	{2, 33, 35} :	5×13, 3×17, 2×9, 1×30, 1×36
5x	{2, 20, 21} :	3×1, 2×13, 2×30, 2×33, 2×40
8x	{33, 20, 35} :	4×9, 4×13, 4×30, 2×7, 2×8
2x	{1, 3, 35} :	2×7, 1×8, 1×13
...		

?x {3, 20, 8} ?, ?, ?, ?, ?, ...

рекомендации!

Качество:

$$\frac{1}{|Z|} \sum_{z \in Z} \frac{|\{r_1, \dots, r_{\min(S, R, z)}\} \cap \{s_1, \dots, s_{\min(S, R, z)}\}|}{\min(S, R, z)}$$

r_1, \dots, r_R – рекомендации

s_1, \dots, s_S – правильные ответы

$$Z = \{5, 10\}$$

Обозначения

Пост-троечная последовательность – вектор

$$v(\{a,b,c\}) = (v_1(\{a,b,c\}), \dots, v_L(\{a,b,c\})),$$

L – число лекций,

$v_j(\{a,b,c\})$ – сколько раз была просмотрена j -я лекция после тройки.

Как решать?

Объединение и пересечение множеств

(отдельная лекция по Fuzzy Sets)

мультимножества и нечёткие множества

$$\{1,2,2,3\} \cup \{2,3,4,4\} = \{1,2,2,2,3,3,4,4\}$$

$$\begin{array}{r} (1,2,1,0, \dots) \\ + (0,1,2,2, \dots) \\ = (1,3,3,2, \dots) \end{array} \quad \begin{array}{l} \text{сложение характеристических} \\ \text{векторов} \end{array}$$

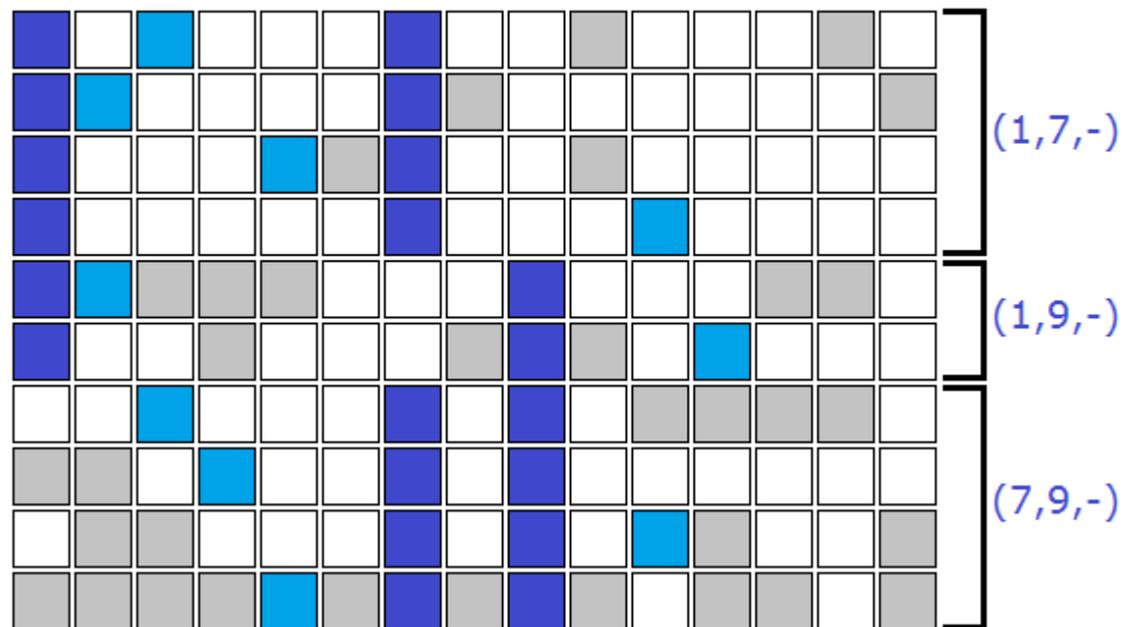
или

$$\{1,2,\mathbf{2},3\} \cup \{2,\mathbf{3},4,4\} = \{1,2,\mathbf{2},3,\mathbf{3},4,4\}$$

$$\begin{array}{r} (1,2,1,0, \dots) \\ \max (0,1,2,2, \dots) \\ = (1,2,2,2, \dots) \end{array} \quad \begin{array}{l} \text{если у элементов есть цвета, то} \\ \text{можем гарантировать, что в} \\ \text{объединение войдёт максимум} \\ \text{элементов...} \end{array}$$

«Объединение информации»

$3 \times \{1, 7, 9\} : 3 \times 5, 2 \times 3, 1 \times 10, 1 \times 12$



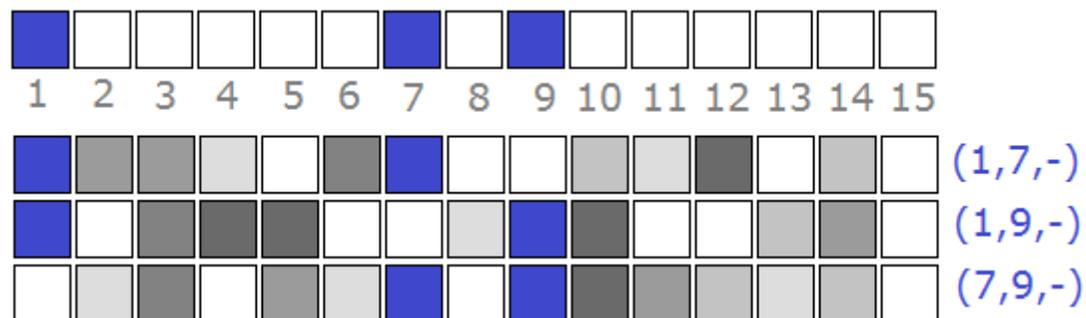
«Объединение информации»

Объединяем с помощью суммирования:

$$s(\{a,b\}) = \sum_d v(\{a,b,d\}),$$
$$s(\{a,c\}) = \sum_d v(\{a,c,d\}),$$
$$s(\{b,c\}) = \sum_d v(\{b,c,d\}).$$

Получили информацию по парам

«Пересечение информации»



$$s(\{a,b\}) \cdot s(\{b,c\}) \cdot s(\{a,c\})$$

$$(s(\{a,b\}) + \varepsilon) \cdot (s(\{b,c\}) + \varepsilon) \cdot (s(\{a,c\}) + \varepsilon)$$

но предварительно использовались нормировки...

Пример нормировки

Аналог IDF

$$v'(\{a,b,c\}) = \left(\frac{v_1(\{a,b,c\})}{\log(|\{\tilde{t} \in T \mid v_1(\tilde{t}) > 0\}| + 2)} \cdots \frac{v_L(\{a,b,c\})}{\log(|\{\tilde{t} \in T \mid v_L(\tilde{t}) > 0\}| + 2)} \right)$$

$|\{\tilde{t} \in T \mid v_j(\tilde{t}) > 0\}|$ – число троек из обучения,

в пост-троечные последовательности которых входит j -я лекция.

Как формировались итоговые оценки

$$\gamma = \log(s(\{a,b\}) \cdot s(\{b,c\}) \cdot s(\{a,c\}))$$

$$\downarrow$$

$$\gamma = \log(s(\{a,b\})) + \log(s(\{b,c\})) + \log(s(\{a,c\}))$$

$$\downarrow$$

$$(s(\{a,b\}) + \varepsilon) \cdot (s(\{b,c\}) + \varepsilon) \cdot (s(\{a,c\}) + \varepsilon)$$

не происходит зануления большинства элементов вектора (и потери информации)

$$\downarrow$$

$$\gamma = \frac{\log(s(\{a,b\}) + 0.02)}{\text{std}(\omega(s(\{a,b\}))) + 0.5} + \frac{\log(s(\{b,c\}) + 0.02)}{\text{std}(\omega(s(\{b,c\}))) + 0.5} + \frac{\log(s(\{a,c\}) + 0.02)}{\text{std}(\omega(s(\{a,c\}))) + 0.5}$$

$\omega \sim$ множество ненулевых элементов вектора

Как формировались итоговые оценки

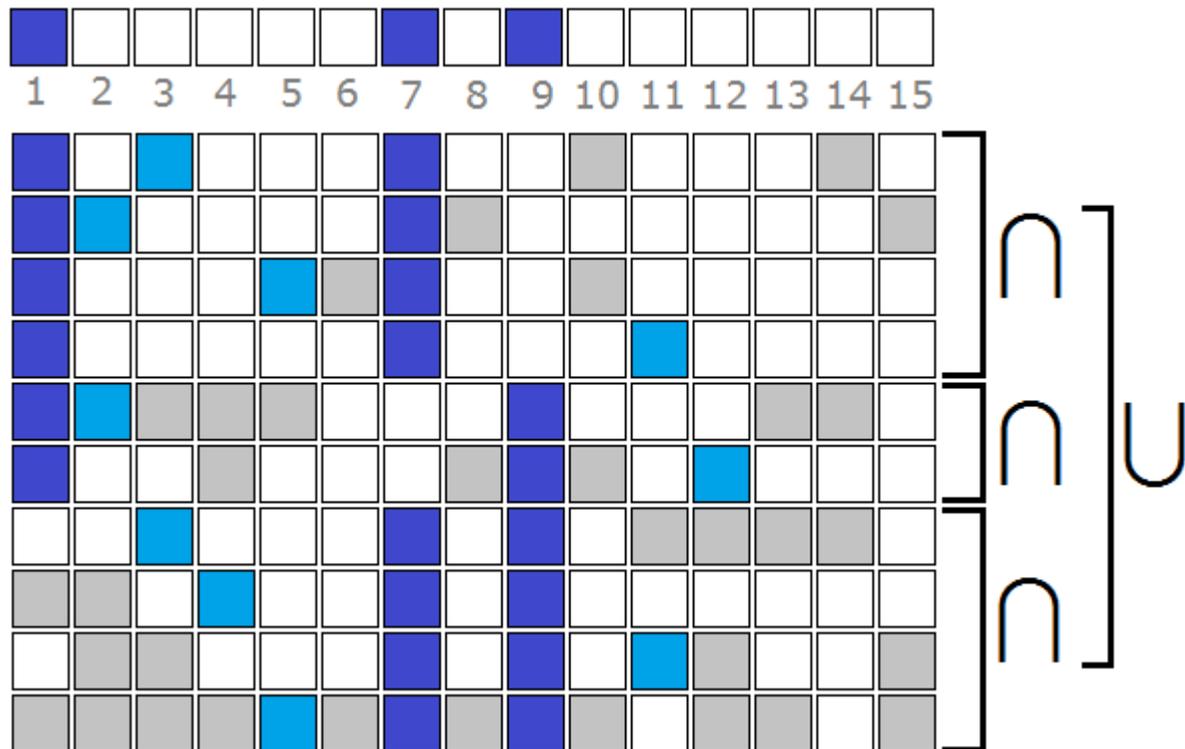
γ (вид выражения)	качество
$(s(\{a,b\}) + \varepsilon) \cdot (s(\{b,c\}) + \varepsilon) \cdot (s(\{a,c\}) + \varepsilon), \varepsilon = 0$	57.27%
$(s(\{a,b\}) + \varepsilon) \cdot (s(\{b,c\}) + \varepsilon) \cdot (s(\{a,c\}) + \varepsilon), \varepsilon = 0.01$	62.11%
$(s(\{a,b\}) + \varepsilon) \cdot (s(\{b,c\}) + \varepsilon) \cdot (s(\{a,c\}) + \varepsilon), \varepsilon = 0.1$	61.60%
$(s(\{a,b\}) + \varepsilon) \cdot (s(\{b,c\}) + \varepsilon) \cdot (s(\{a,c\}) + \varepsilon), \varepsilon = 1$	58.84%
$(s(\{a,b\}) + s(\{b,c\}) + \varepsilon) \cdot (s(\{b,c\}) + s(\{a,c\}) + \varepsilon) \cdot (s(\{a,c\}) + s(\{a,b\}) + \varepsilon), \varepsilon = 0$	58.63%
$(s(\{a,b\}) + s(\{b,c\}) + \varepsilon) \cdot (s(\{b,c\}) + s(\{a,c\}) + \varepsilon) \cdot (s(\{a,c\}) + s(\{a,b\}) + \varepsilon), \varepsilon = 0.001$	59.87%

РП второго места:

$$\gamma = s(\{a,b\}) \log(s(\{a,b\})) + s(\{b,c\}) \log(s(\{b,c\})) + s(\{a,c\}) \log(s(\{a,c\}))$$

Идея алгоритма

А так... всё просто



Rank	Team	Preliminary Result	Final Result
1	+ D'yakonov Alexander	0.62102	0.62415
2	meridion	0.60791	0.61172
3	UniQ	0.58727	0.59063
4	+ Haibin Liu	0.47384	0.47507
5	+ barney	0.47060	0.47243
6	vyatka	0.45149	0.45553
7	+ Saul Delabrida	0.44343	0.44571
8	+ Inner Peace	0.28096	0.28282
9	+ DMIR	0.27137	0.27439
10	dddnnn	0.25921	0.26185

Если известно хорошее статистическое описание объекта, то признаковое описание бесполезно

Автор, область, кратность (в п-т посл-ти)	Название
Anastasia Krithara Text Mining	Active, Semi-Supervised Learning for Textual Information Access
Isabelle Guyon Machine Learning	Introduction to Machine Learning
Mikaela Keller Statistics	Basics of probability and statistics
Ulrike von Luxburg, Clustering, 5x	Lectures on Clustering
William Cohen, Text Mining, 4x	Text Classification
John Shawe-Taylor, Statistical Learning, 3x	Statistical Learning Theory
Cynthia Rudin, Boosting, 3x	The Dynamics of AdaBoost