

大規模データ処理システム Big Data Processing System

協調フィルタリング (1)

Collaborative Filtering (1)

- 「Aを買った人はBを買っています。」
- 消費者の消費行動をログとして蓄積している。
 - 慶應太郎, 学問のすすめ, 2014/06/09, ...
 - 慶應太郎, 福翁自伝, 2014/04/30, ...
 - 慶應花子, 福翁自伝, 2014/03/31, ...
 - 湘南次郎, 尊王論, 2013/03/01, ...
 - 慶應花子, 尊王論, 2012/12/01, ...
- A people, who buy A, buy B too.
- A system collects purchasing log.
 - Taro KEIO, Gakumon no susume, 2014/06/09, ...
 - Taro KEIO, Fukuou jiden, 2014/04/30, ...
 - Hanako KEIO, Fukuou jiden, 2014/03/31, ...
 - Jiro SHONAN, Sonnouron, 2013/03/01, ...
 - Hanako KEIO, Sonnouron, 2012/12/01, ...

協調フィルタリング (2)

Collaborative Filtering (2)

- アイテム毎に集計する（ユーザのベクトルを作る）。
 - 学問のすすめ: 慶應太郎
 - 福翁自伝: 慶應太郎, 慶應花子
 - 尊王論: 湘南次郎, 慶應花子
- Totaling by each item
 - Gakumon no susume: Taro KEIO
 - Fukuou jiden: Taro KEIO, Hanako KEIO
 - Sonnou ron: Jiro SHONAN, Hanako KEIO

協調フィルタリング (3)

Collaborative Filtering (3)

- アイテムごとのベクトルの類似度を計算する。
 - コサイン類似度:
$$\cos(a,b) = a \cdot b / (|a| \cdot |b|)$$
$$= \text{両方アクセスしたユーザ数} / a, b \text{それぞれにアクセスしたユーザ数}$$
 - 1に近ければ類似している。0に近ければ類似していない。
 - $\cos(\text{学問のすすめ}, \text{福翁自伝}) = 1 / (\sqrt{1} * \sqrt{2}) = 0.707$
- Make similarity of vectors of each item
 - Cosine similarity:
$$\cos(a,b) = a \cdot b / (|a| \cdot |b|)$$
$$= \text{num of users who access both} / \text{num of users who access a or b}$$
 - If similarity is close to 1, it might be similar. If similarity is close to 0, it might be not similar.
 - $\cos(\text{“Gakumon no susume”}, \text{“Fukuou jiden”}) = 1 / (\sqrt{1} * \sqrt{2}) = 0.707$

協調フィルタリングを考える (4)

Collaborative Filtering (4)

大きく2段階の処理

- アイテム毎に集計する
 - mapによって、<アイテム, ユーザ> というkey-valueペアを作る。
 - reduceによってアイテム毎にユーザのベクトル<アイテム, (ユーザ1, ユーザ2...)>を作成する。
- 類似度を計算する
 - $\cos(\text{学問のすすめ}, \text{慶應太郎})$ を計算する。
 - アイテム数²の類似度が生成される。

Two phases

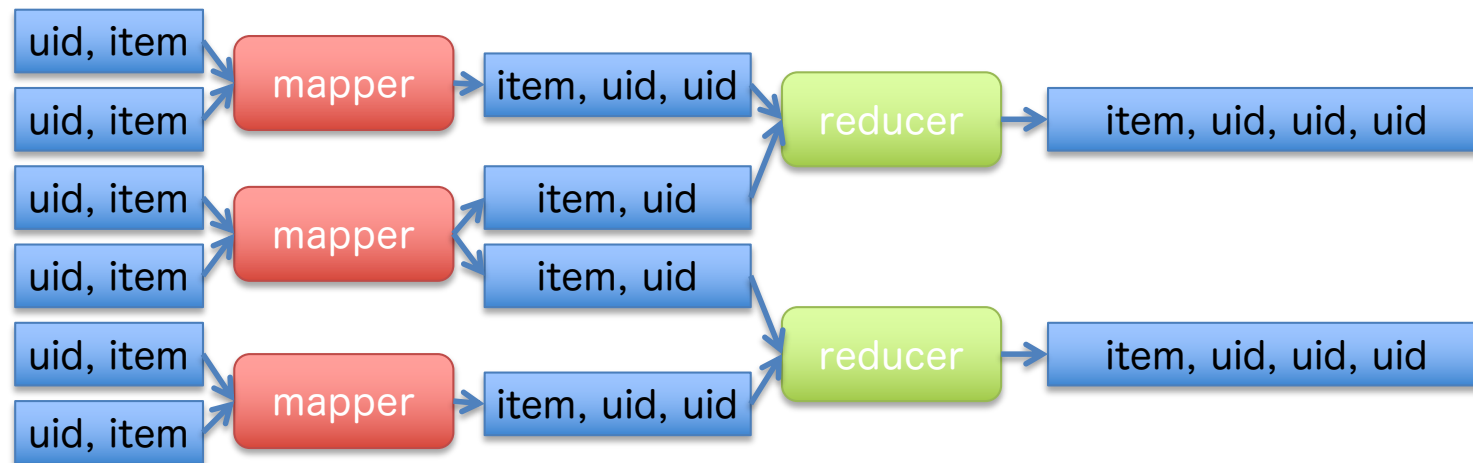
- Totaling by items
 - Make <item, user> key-value pair using map procedure
 - Make <item, (user1, user2, ...)> vector using reduce procedure
- Make similarity
 - Make $\cos(\text{item1}, \text{item2})$
 - The process produces (num of items)² similarity

協調フィルタリングを考える (5)

Collaborative Filtering (5)

Two phases

- Totaling by items
 - Make <item, user> key-value pair using map procedure
 - Make <item, (user1, user2, ...)> vector using reduce procedure



- Make similarity
 - Make $\cos(\text{item1}, \text{item2})$
 - The process produces $(\text{num of items})^2$ similarity



全てのアイテム間の関係を計算しなければならず、ローカリティが保てない
All combination of items must be calculated.
It means there's no locality.

協調フィルタリングを考える (6)

Collaborative Filtering (6)

- コサイン類似度を求めるためには(a, b)のアイテムペアに対する、下記の情報がわかればよい。
 - 両方を購入した人数
 - aを購入した人数
 - bを購入した人数
- アイテム毎の購入者数(aを購入した人数、bを購入した人数)はアイテムリストを作った段階で自明。
- 故に、両方を購入した人数を求めることができれば、コサイン類似度は求められる。

※コサイン類似度:

$$\begin{aligned}\cos(a,b) &= a \cdot b / (|a| \cdot |b|) \\ &= \text{両方アクセスしたユーザ数} / a, b \text{それぞれにアクセスしたユーザ数}\end{aligned}$$

協調フィルタリングを考える (6)

Collaborative Filtering (6)

- When you make Cosine similarity, you need following information for a pair of a and b.
 - the number of peoples who bought both a and b
 - the number of peoples who bought a
 - the number of peoples who bought b
- It is easy to count the number of peoples who bought a or b. Because we already have the item list.
- We need to count the number of peoples who bought both a and b.

※ Cosine similarity :

$$\cos(a,b) = a \cdot b / (|a| \cdot |b|)$$

= num of users who access both / num of users who access a or b

協調フィルタリングを考える (7) Collaborative Filtering (7)

- 両方を購入した人数を計算するためには、次のようにすれば良い。
To count num of peoples who buy both...

- ユーザが買ったもののリストを作成する。 / Make a list peoples buy.

- α (a, b, c, d, e).
- β (a, b, c, f, g).
- γ (b, c, g, h, i).

$$\alpha, \beta, \gamma: \text{People}$$

$$a, b, c, \dots: \text{item}$$

その上で、ある品を買った人が他に何を買っているのかを調べる。

$\langle a, (b, c, d, e) \rangle$	$\langle b, (a, c, d, e) \rangle$	$\langle c, (a, b, d, e) \rangle$	$\langle d, (a, b, c, e) \rangle$	$\langle e, (a, b, c, d) \rangle$
$\langle a, (b, c, f, g) \rangle$	$\langle b, (a, c, f, g) \rangle$	$\langle c, (a, b, f, g) \rangle$	$\langle f, (a, b, e, g) \rangle$	$\langle g, (a, b, e, f) \rangle$
$\langle b, (c, g, h, i) \rangle$	$\langle c, (b, g, h, i) \rangle$	$\langle g, (b, c, h, i) \rangle$	$\langle h, (b, c, g, i) \rangle$	$\langle i, (b, c, g, h) \rangle$

map

$\langle (a, b), (2, 2, 0) \rangle$ $\langle (b, a), (2, 0, 2) \rangle$ $\langle (a, c), (2, 2, 0) \rangle$ $\langle (c, a), (2, 0, 2) \rangle$ $\langle (a, d), (1, 2, 0) \rangle$...	$\langle (b, a), (2, 3, 0) \rangle$ $\langle (a, b), (2, 0, 3) \rangle$ $\langle (b, c), (3, 3, 0) \rangle$ $\langle (c, b), (3, 0, 3) \rangle$ $\langle (b, d), (1, 3, 0) \rangle$...	$\langle (c, a), (2, 3, 0) \rangle$ $\langle (a, c), (2, 0, 3) \rangle$ $\langle (c, b), (3, 3, 0) \rangle$ $\langle (b, c), (3, 0, 3) \rangle$ $\langle (c, d), (1, 3, 0) \rangle$...	$\langle (d, a), (1, 1, 0) \rangle$ $\langle (a, d), (1, 0, 1) \rangle$ $\langle (d, b), (1, 1, 0) \rangle$ $\langle (b, d), (1, 0, 2) \rangle$ $\langle (d, c), (1, 1, 0) \rangle$...
--	--	--	--

reduce

aとbが同時に
現れる回数

aが現れる回数
(行列)模二

協調フィルタリングを考える (8)

Collaborative Filtering (8)

- 品物のペア毎に集計し、類似度を計算する。
Totaling by item pair. Make similarity.

$\langle (a, b), (2, 2, 0) \rangle$	$\langle (b, a), (2, 3, 0) \rangle$	$\langle (c, a), (2, 3, 0) \rangle$	$\langle (d, a), (1, 1, 0) \rangle$	map
$\langle (b, a), (2, 0, 2) \rangle$	$\langle (a, b), (2, 0, 3) \rangle$	$\langle (a, c), (2, 0, 3) \rangle$	$\langle (a, d), (1, 0, 1) \rangle$	
$\langle (a, c), (2, 2, 0) \rangle$	$\langle (b, c), (3, 3, 0) \rangle$	$\langle (c, b), (3, 3, 0) \rangle$	$\langle (d, b), (1, 1, 0) \rangle$	
$\langle (c, a), (2, 0, 2) \rangle$	$\langle (c, b), (3, 0, 3) \rangle$	$\langle (b, c), (3, 0, 3) \rangle$	$\langle (b, d), (1, 0, 1) \rangle$	
$\langle (a, d), (1, 2, 0) \rangle$	$\langle (b, d), (1, 3, 0) \rangle$	$\langle (c, d), (1, 3, 0) \rangle$	$\langle (d, c), (1, 1, 0) \rangle$	
...	

- $\langle (a,b), (2,2,0) \rangle \langle (a,b), (2,0,3) \rangle \rightarrow \langle (a,b), (2,2,3) \rangle \rightarrow 0.816$
- $\langle (b,a), (2,0,2) \rangle \langle (b,a), (2,3,0) \rangle \rightarrow \langle (b,a), (2,3,2) \rangle \rightarrow 0.816$
- ...
- $\langle (b,c), (3,3,0) \rangle \langle (c,b), (3,3,0) \rangle \rightarrow \langle (b,c), (3,3,3) \rangle \rightarrow 1.000$
- ...
- $\langle (b,d), (1,3,0) \rangle \langle (b,d), (1,0,1) \rangle \rightarrow \langle (b,d), (1,3,1) \rangle \rightarrow 0.577$
- ...

reduce

should match