

Book displays at Flourish and Blotts

Leonardo Ji

2017-04-11

Questions

We did this work to find a way to arrange the book displays in order to increase sales at Flourish and Blotts bookstore. The data set given to us contains 90,000+ historical sales transactions and we want to answer the following questions:

1. What are the best-selling titles?
2. If the manager has to create a book display to appeal to readers who belong to book clubs, what books should be included? He said that the typical book club audience would be someone who is reading titles featured by Oprah's Book Club (<https://static.oprah.com/images/o2/201608/201608-obc-complete-list-01a.pdf>).
3. Can you recommend other books that he should include in display cases? The manager is adamant that your recommendations do not include the following:
 - a. Books in a series (i.e. Girl with the Dragon Tattoo series would be an example). The manager already knows series books should be displayed together.
 - b. The title "Wild Animus" You were really surprised by this request and pressed the manager for an explanation. He replied that you should read this blog entry: <https://litreactor.com/columns/what-the-hell-is-wild-animus>

Data

Load 'Matrix', 'arules', 'arulesViz' packages.

First we Read zip file separated by tabs, specify the column of transaction IDs and of item IDs, remove duplicate entries.

```
bookbaskets <- read.transactions("C:/Users/lj015625/Desktop/DataMining Class/HW4/data/bookdata.tsv.gz",
                                format="single",
                                sep="\t",
                                cols=c("userid", "title"),
                                rm.duplicates=T)
```

```
## distribution of transactions with duplicates:
```

```
## items
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
## 701 222 106 68 43 39 23 24 18 18 16 10 7 7 13 7 8 5
## 19 20 21 22 23 25 26 27 28 29 30 31 33 34 35 38 39 42
##  3  9  4  4  3  2  2  5  4  5  4  4  1  2  1  1  1  2
## 44 45 47 48 49 52 56 57 59 61 63 71 73 80 84 86 91 93
##  1  1  1  1  1  1  2  1  2  1  2  1  1  1  1  1  1  1
## 95 96 99 103 158 206 260 891
##  1  1  1  1  1  2  1  1
```

We can see there are 92,108 transactions and 220,447 customers. This creates a sparse matrix of book titles by transactions.

```

# 92,108 book purchases.
#220,447 user IDs.
inspect(bookbaskets[1:5]) #Examine the first five transactions

##      items
## [1] {New Vegetarian: Bold and Beautiful Recipes for Every Occasion}
## [2] {Il Dio Delle Piccole Cose}
## [3] {Cybernation}
## [4] {Lasher: Lives of the Mayfair Witches}
## [5] {Chicken Soup for the Teenage Soul on Tough Stuff : Stories of Tough Times and Lessons Learned,
##      Dragon Ball Z, Vol. 1,
##      Harry Potter and the Chamber of Secrets,
##      Harry Potter and the Sorcerer's Stone,
##      Holes,
##      Prey,
##      Primary Colors: A Novel of Politics,
##      Rising Sun,
##      The Cat Who Smelled a Rat,
##      The Fellowship of the Ring,
##      The Hunt for Red October,
##      The Return of the King,
##      The Two Towers}

dim(bookbaskets)

## [1] 92108 220447

summary(bookbaskets)

## transactions as itemMatrix in sparse format with
## 92108 rows (elements/itemsets/transactions) and
## 220447 columns (items) and a density of 0.00005034811
##
## most frequent items:
##
##              Wild Animus
##              2502
##      The Lovely Bones: A Novel
##              1295
##              She's Come Undone
##              934
##              The Da Vinci Code
##              905
## Harry Potter and the Sorcerer's Stone
##              832
##              (Other)
##              1015847
##
## element (itemset/transaction) length distribution:
## sizes
##      1      2      3      4      5      6      7      8      9     10     11     12
## 51286 10804 5760 3850 2700 2044 1609 1241 1075 901 755 643
##    13    14    15    16    17    18    19    20    21    22    23    24
##   555   460   464   393   342   332   268   258   237   222   195   179
##    25    26    27    28    29    30    31    32    33    34    35    36
##   182   170   156   154   129   114   120   103   128   101    88    94

```

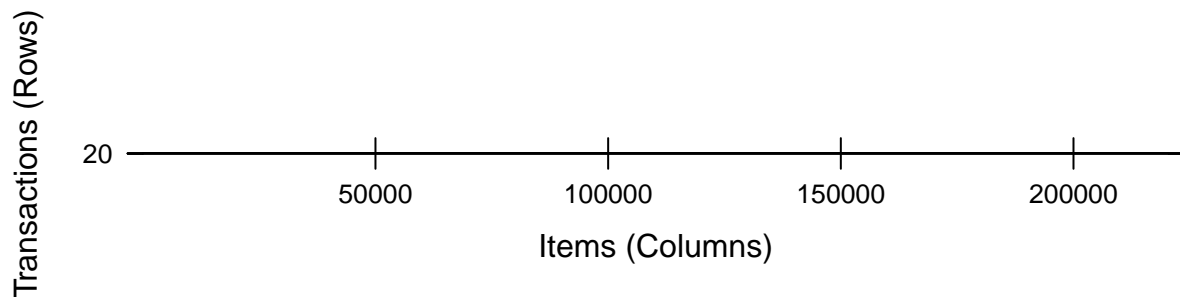
##	37	38	39	40	41	42	43	44	45	46	47	48
##	98	82	71	80	60	81	70	65	73	77	79	54
##	49	50	51	52	53	54	55	56	57	58	59	60
##	54	47	43	50	52	43	38	47	43	39	46	44
##	61	62	63	64	65	66	67	68	69	70	71	72
##	36	30	31	24	46	25	27	35	29	28	32	22
##	73	74	75	76	77	78	79	80	81	82	83	84
##	26	24	25	19	22	26	27	21	29	13	29	21
##	85	86	87	88	89	90	91	92	93	94	95	96
##	20	13	15	10	21	26	14	16	23	13	15	19
##	97	98	99	100	101	102	103	104	105	106	107	108
##	14	20	21	12	12	14	17	16	15	12	13	18
##	109	110	111	112	113	114	115	116	117	118	119	120
##	11	11	15	16	14	10	12	14	5	9	17	5
##	121	122	123	124	125	126	127	128	129	130	131	132
##	10	11	11	9	17	16	14	12	6	10	10	8
##	133	134	135	136	137	138	139	140	141	142	143	144
##	8	9	10	11	8	5	7	12	9	6	6	7
##	145	146	147	148	149	150	151	152	153	154	155	156
##	7	9	7	6	6	5	8	10	7	11	6	9
##	157	158	159	160	161	162	163	164	165	166	167	168
##	7	3	8	7	8	6	7	9	4	9	6	6
##	169	170	171	172	173	174	175	176	177	178	179	180
##	1	10	6	6	6	2	7	8	3	9	5	5
##	181	182	183	184	185	186	187	188	189	190	191	192
##	4	9	6	8	3	2	9	5	5	6	7	4
##	193	194	195	196	197	198	199	200	201	202	203	204
##	7	1	6	8	4	2	7	4	9	6	4	2
##	205	206	207	208	209	210	212	213	214	215	216	217
##	4	4	4	10	2	4	1	3	3	4	3	5
##	218	219	220	221	222	223	224	225	226	227	228	229
##	1	5	5	8	2	4	5	3	7	3	2	2
##	230	231	232	233	234	235	236	237	238	239	240	241
##	6	6	5	3	8	5	5	6	4	6	2	4
##	242	243	244	245	246	247	248	249	250	251	252	253
##	5	5	1	2	1	1	4	1	2	1	4	2
##	254	255	256	257	258	259	260	261	262	263	264	265
##	2	2	3	1	3	4	2	4	3	4	1	1
##	266	267	268	269	270	271	272	275	276	277	278	279
##	2	4	5	2	5	4	3	6	4	1	1	2
##	280	281	282	283	284	285	286	287	288	289	290	291
##	1	3	5	2	2	2	3	2	4	1	1	4
##	293	295	296	297	298	300	301	302	304	305	306	307
##	1	1	2	1	4	1	3	1	1	1	2	1
##	308	309	310	311	313	314	315	316	317	319	320	321
##	3	4	2	1	2	3	1	2	3	3	1	1
##	322	323	324	325	326	327	328	329	330	331	332	333
##	3	4	1	2	2	1	1	6	1	2	2	2
##	334	336	337	338	339	340	341	342	343	344	346	347
##	2	2	1	1	1	1	1	2	2	3	2	2
##	348	349	350	352	354	356	357	358	359	360	363	364
##	1	1	1	2	1	1	1	1	1	1	1	1
##	366	367	368	369	370	372	373	374	375	376	377	378
##	3	1	1	2	1	2	1	2	1	2	2	1

##	379	381	382	383	384	385	386	387	388	389	390	391
##	1	2	2	1	4	2	2	1	3	1	1	2
##	392	393	394	395	396	397	398	399	401	403	406	407
##	4	1	2	1	2	1	4	1	1	3	1	1
##	409	410	412	415	416	417	418	419	420	421	422	425
##	3	1	1	1	3	1	4	1	1	2	2	1
##	427	428	430	435	436	437	438	440	441	442	444	445
##	1	1	1	3	1	1	2	2	4	1	1	1
##	446	447	448	451	453	455	456	457	458	459	460	464
##	1	1	2	1	1	2	2	1	1	2	1	2
##	466	468	472	476	480	481	485	487	489	492	493	494
##	1	2	1	1	1	1	2	4	1	1	1	1
##	496	497	498	507	508	510	512	515	516	517	520	522
##	1	2	1	1	1	1	1	1	1	1	2	2
##	523	524	525	526	527	528	532	534	537	539	540	542
##	1	2	1	1	1	2	2	1	1	1	2	1
##	543	544	547	562	566	570	572	573	575	577	580	590
##	1	1	1	2	1	1	1	1	2	1	2	1
##	591	596	597	599	600	601	602	603	611	613	620	624
##	2	2	2	1	1	1	1	3	2	1	2	1
##	625	626	627	628	629	649	653	656	657	658	661	662
##	1	1	1	2	1	1	1	1	1	1	1	1
##	664	665	666	667	670	673	675	690	692	696	698	705
##	1	1	1	1	1	1	1	1	1	1	1	1
##	707	709	713	721	724	725	727	732	734	736	740	742
##	1	1	1	1	1	1	2	1	1	2	1	2
##	745	746	748	765	767	769	775	781	782	785	790	796
##	2	1	1	1	1	1	2	1	1	1	1	1
##	802	804	805	807	812	814	818	820	829	834	837	849
##	1	1	1	1	1	1	1	1	1	1	1	1
##	855	858	870	878	880	887	894	897	903	917	922	933
##	1	2	1	1	1	1	1	1	1	1	1	3
##	945	948	950	953	961	965	968	971	978	985	1009	1010
##	1	1	1	1	1	1	1	1	1	1	1	1
##	1016	1018	1025	1030	1035	1036	1039	1054	1074	1076	1077	1078
##	1	1	1	1	1	1	1	1	1	2	2	1
##	1079	1080	1083	1088	1089	1111	1132	1133	1136	1149	1152	1153
##	1	1	1	1	1	1	1	1	1	1	1	1
##	1157	1171	1183	1190	1199	1202	1234	1241	1242	1253	1255	1260
##	1	1	1	1	1	1	1	1	1	1	1	1
##	1264	1270	1275	1285	1293	1308	1312	1317	1320	1326	1340	1351
##	1	1	1	1	1	1	2	1	1	1	1	1
##	1359	1395	1433	1463	1481	1496	1511	1512	1514	1518	1542	1578
##	1	1	1	1	1	1	1	1	1	1	1	1
##	1582	1621	1626	1631	1668	1687	1705	1734	1799	1802	1808	1840
##	1	1	1	1	1	1	1	1	1	1	1	1
##	1966	2040	2060	2124	2190	2197	2202	2215	2270	2272	2291	2295
##	1	1	1	1	1	1	1	1	1	1	1	1
##	2305	2364	2373	2413	2542	2773	2864	2884	2908	3236	3897	4187
##	1	1	1	1	1	1	1	1	1	1	1	1
##	5440	5608	5683	6196	10253							
##	1	1	1	1	1							
##												
##	Min.	1st Qu.	Median		Mean	3rd Qu.		Max.				

```
##      1.0      1.0      1.0      11.1      4.0 10250.0
##
## includes extended item information - examples:
##                                labels
## 1                                'N' Is for Moose
## 2 ' Allo 'Allo: the War Diaries of Rene Artois
## 3                                ' Boule De Suif
##
## includes extended transaction information - examples:
## transactionID
## 1              10
## 2              1000
## 3             100001
```

The density is 0.005%, this means most of values in transaction dataset are 0. The image graph shows sparse matrix of the sample of 100 transactions.

```
image(sample(bookbaskets, 100))
```



To answer the first question “what is the top selling books” we need to do the following

- First we calculate how many of the same books customers bought. This number is also called support.
 $Support = Count(X)$
- Then we calculate frequency/relative support of books by dividing its support to total count.
 $RelativeSupport = Count(X)/N$
- Then multiple relative support to total number of books with sorting to get top twenty most popular books.

```

basketSizes<-size(bookbaskets)
# Calculate the support for each book title
bookFreq<-itemFrequency(bookbaskets)
# Get the absolute count of book occurrences.
bookCounts <- (bookFreq/sum(bookFreq))*sum(basketSizes)
# High frequency items
orderedBooks = sort(bookCounts, decreasing = TRUE)
#head(orderedBooks, 10)

orderedBooks_df <- as.data.frame(orderedBooks)
orderedBooks_df$title <- rownames(orderedBooks_df)
rownames(orderedBooks_df) <- 1:nrow(orderedBooks_df)
colnames(orderedBooks_df) <- c("Support", "Title")
orderedBooks_df <- orderedBooks_df[c("Title", "Support")]
kable(orderedBooks_df[1:20,], caption = "Top 20 Books")

```

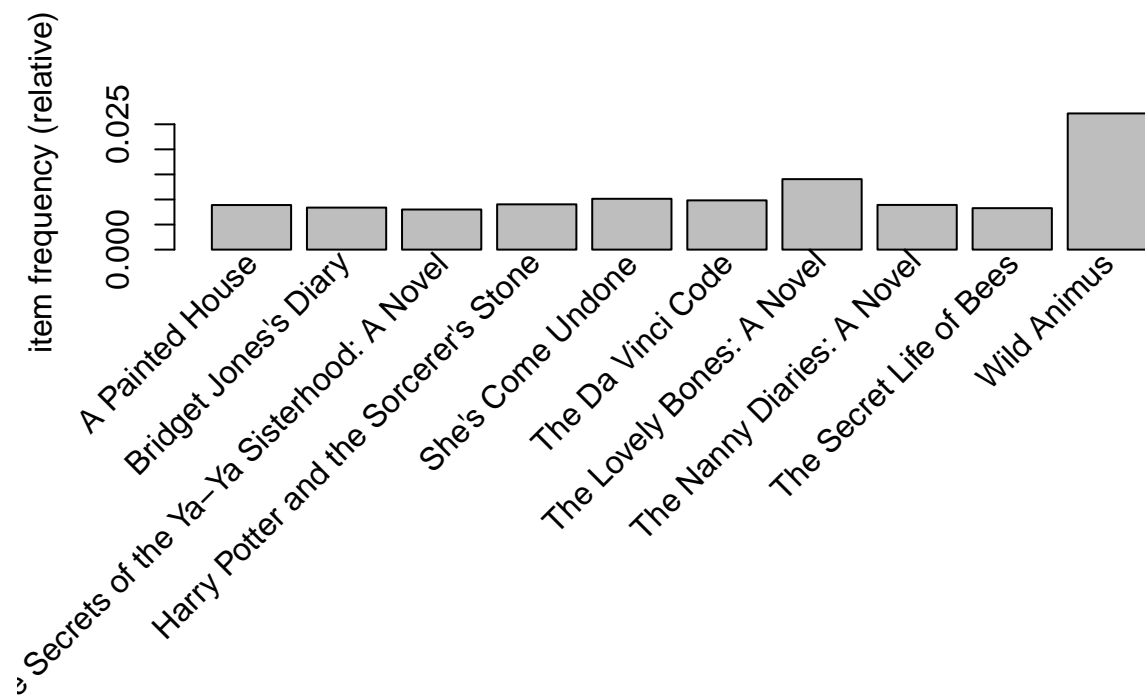
Table 1: Top 20 Books

Title	Support
Wild Animus	2502
The Lovely Bones: A Novel	1295
She's Come Undone	934
The Da Vinci Code	905
Harry Potter and the Sorcerer's Stone	832
The Nanny Diaries: A Novel	821
A Painted House	819
Bridget Jones's Diary	772
The Secret Life of Bees	762
Divine Secrets of the Ya-Ya Sisterhood: A Novel	737
The Red Tent	727
White Oleander : A Novel	713
Angels & Demons	661
Snow Falling on Cedars	661
Life of Pi	658
The Summons	651
Girl with a Pearl Earring	647
Where The Heart Is	625
The Notebook	620
The Testament	600
####Support plots	
The first graph shows books with relative support	above 0.8%.
The below second graph shows top twenty books purc	hased.

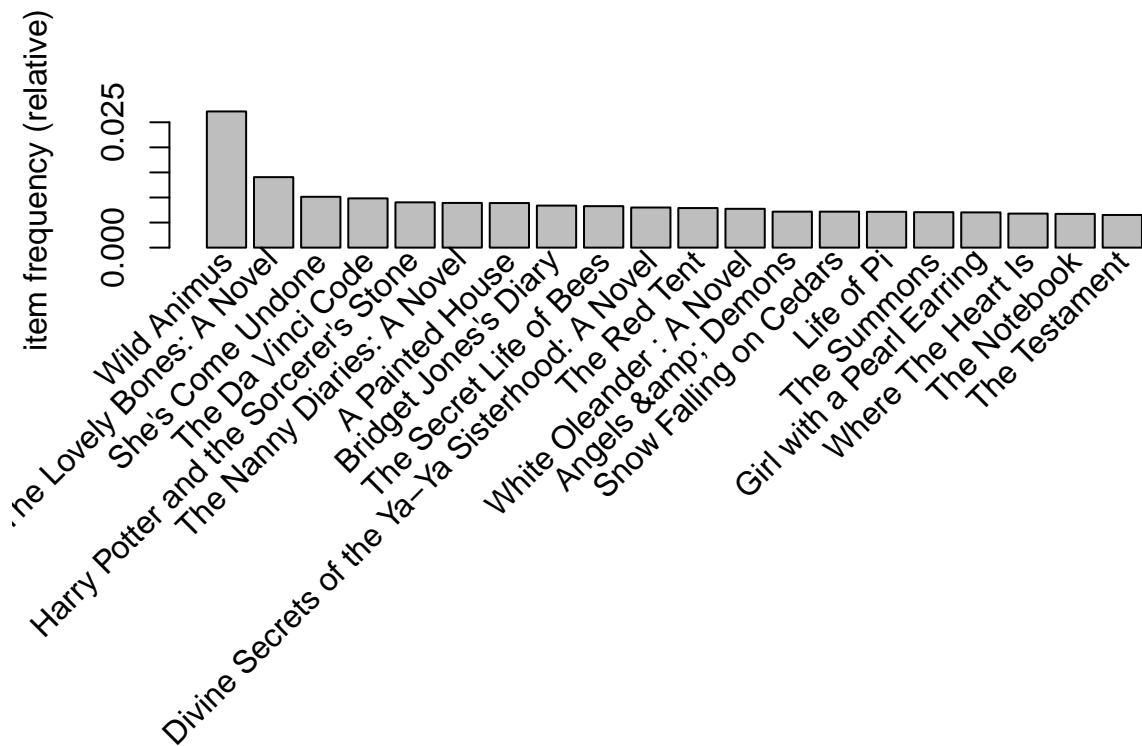
```

#Let's plot the frequency of items
# Support number low because lower frequency on books inside dataset.
# Let's impose a rule. Let's say we only want to see items with at least 0.8% support
itemFrequencyPlot(bookbaskets, support = 0.008)

```



#Now let's say we only want to see "top 30" items (i.e. 20 items most frequently purchased)
`itemFrequencyPlot(bookbaskets, topN = 20)`



Model

We need an association rule model to list strong rules. For example, a customer buys book 1 also buys book 2. This is similar to recommendation feature on Amazon. We also want to remove large transactions with more than 200 books. Those larger transactions with more than 200 books are treated as outliers. They are not typical book store customers' transactions. We will remove it. We also removed single book transaction because association rule needs at least two items. Association Rules by Apriori method is good for this purpose. Apriori method works well on any transaction data set. It creates a search tree of item set then prune the tree use minimal support and confidence value. The method eventually returns rules of associated items. Rules follow the format like items on left hand side return Items on the right hand side. For this rule we know associated items.

To start Apriori method first we want to keep only those transactions when customer bought more than one book.

A single book transaction would not be useful for this model.

```
#Only keep transactions with more than one book and let than 200 books purchased.
bookbaskets_use<-bookbaskets[basketSizes>1 & basketSizes<200]
basketSizes<-size(bookbaskets_use)
```

Question 2 ask us to create a list of book for customers who follows Oprahs Book Club.

There are seventy-six books in Oprah Book Club.

We want to know how many of those same books also existed in the transaction data set. The Intersect two list contains fifty-six titles.


```
oprahBookClub <- read.csv("C:/Users/lj015625/Desktop/DataMining Class/HW4/data/oprahBookClub.csv")
orderedBooks_df <- as.data.frame(orderedBooks)
length(oprahBookClub$Title)
```

```
## [1] 76
```

```
oprahBookClub_intersect <- intersect(rownames(orderedBooks_df), oprahBookClub$Title)
length(oprahBookClub_intersect)
```

```
## [1] 56
```

If we use Oprah Book Club books on the left hand side rule of the Apriori algorithm this would give us books we should display for a typical Oprah Book Club reader. We did trial and errors and end up picking support at 0.002%, confidence at 60%, minimum length at 2, maximum length at 5. The support concept is stated in previous section. We also specify a minimal confidence number. While support measures How often does the rule happen. Confidence measures how often is the rule correct. $confidence(X \rightarrow Y) = (support(X, Y)) / (support(X))$

```
oprahRule_df_total_df <- data.frame(rules=factor(),
                                   support=double(),
                                   Factors=double(),
                                   lift=double(),
                                   stringsAsFactors=FALSE)
oprahRule <- apriori(bookbaskets_use, parameter = list(support = 0.0002, confidence = 0.6, minlen = 2, maxlen = 5,
                                                       appearance=list(lhs=c(oprahBookClub_intersect), default="rhs"))
```

```
## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##          0.6   0.1   1 none FALSE                TRUE         5 0.0002      2
## maxlen target  ext
##          5 rules FALSE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##      0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 8
##
## set item appearances ...[56 item(s)] done [0.00s].
## set transactions ...[142363 item(s), 40011 transaction(s)] done [0.53s].
## sorting and recoding items ... [9150 item(s)] done [0.08s].
## creating transaction tree ... done [0.06s].
## checking subsets of size 1 2 3 4 done [0.60s].
## writing ... [28 rule(s)] done [0.14s].
## creating S4 object ... done [0.05s].
```

```
oprahRule_sorted <- sort(oprahRule, by = c("lift", "confidence"))
oprahRule_df <- as(oprahRule_sorted, "data.frame")
```

We can sort rules by lift. The table below shows titles in the transaction data set could appeal to readers who belong to Oprah's Book Club.

```
kable(oprahRule_df[order(-oprahRule_df$lift), ], caption = "Oprah Book Club Association Rules")
```

Table 2: Oprah Book Club Association Rules

	rules	support
15	{Open House,Vinegar Hill} => {Drowning Ruth}	0.000224
6	{House of Sand and Fog,Songs in Ordinary Time} => {Drowning Ruth}	0.000224
21	{Black and Blue,The Book of Ruth} => {Drowning Ruth}	0.000224
5	{Songs in Ordinary Time,While I Was Gone} => {Drowning Ruth}	0.000224
16	{I Know This Much Is True,Open House} => {Drowning Ruth}	0.000224
10	{Jewel,The Rapture of Canaan} => {The Pilot's Wife : A Novel}	0.000299
13	{House of Sand and Fog,The Rapture of Canaan} => {The Pilot's Wife : A Novel}	0.000224
14	{A Lesson Before Dying,Jewel} => {Where The Heart Is}	0.000224
3	{Icy Sparks,Songs in Ordinary Time} => {Where The Heart Is}	0.000249
26	{I Know This Much Is True,Songs in Ordinary Time,The Reader} => {White Oleander : A Novel}	0.000224
2	{The Bluest Eye,The Reader} => {Where The Heart Is}	0.000224
19	{Icy Sparks,The Reader} => {White Oleander : A Novel}	0.000274
9	{The Rapture of Canaan,We Were the Mulvaney's} => {Divine Secrets of the Ya-Ya Sisterhood: A Novel}	0.000224
28	{I Know This Much Is True,The Reader,While I Was Gone} => {She's Come Undone}	0.000249
27	{The Reader,Vinegar Hill,While I Was Gone} => {She's Come Undone}	0.000224
1	{Mother of Pearl,The Reader} => {White Oleander : A Novel}	0.000224
4	{Jewel,Songs in Ordinary Time} => {White Oleander : A Novel}	0.000249
12	{I Know This Much Is True,The Rapture of Canaan} => {She's Come Undone}	0.000224
11	{The Rapture of Canaan,While I Was Gone} => {She's Come Undone}	0.000274
18	{Vinegar Hill,While I Was Gone} => {She's Come Undone}	0.000374
7	{Songs in Ordinary Time,While I Was Gone} => {She's Come Undone}	0.000249
24	{I Know This Much Is True,While I Was Gone} => {She's Come Undone}	0.000324
23	{The Reader,While I Was Gone} => {She's Come Undone}	0.000374
20	{Jewel,While I Was Gone} => {She's Come Undone}	0.000299
25	{House of Sand and Fog,While I Was Gone} => {She's Come Undone}	0.000299
17	{Open House,While I Was Gone} => {The Lovely Bones: A Novel}	0.000224
8	{Songs in Ordinary Time,While I Was Gone} => {The Lovely Bones: A Novel}	0.000224
22	{Black and Blue,While I Was Gone} => {The Lovely Bones: A Novel}	0.000224

Question 3 asks us to identify additional books We want to recommend for display.

We can run Apriori Association Rules using all books in transaction data set and using a higher relative support value. This would give us a list of books customers most likely to read. We sorted resulting rules by lift and confidence. Lift measures “how much more likely an item is to be purchased relative to its typical purchase rate, given that you know another item has been purchased” (Lantz 2013, p. 261).

$Lift = Confidence / Support$

```
bookbasketrules <- apriori(bookbaskets_use, parameter = list(support = 0.001, confidence = 0.8, minlen =
```

```
## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##          0.8    0.1    1 none FALSE                TRUE         5   0.001      2
## maxlen target  ext
##          10  rules FALSE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##          0.1 TRUE TRUE  FALSE TRUE     2    TRUE
##
```

```
## Absolute minimum support count: 40
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[142363 item(s), 40011 transaction(s)] done [0.62s].
## sorting and recoding items ... [1472 item(s)] done [0.03s].
## creating transaction tree ... done [0.02s].
## checking subsets of size 1 2 3 4 5 done [0.05s].
## writing ... [98 rule(s)] done [0.03s].
## creating S4 object ... done [0.03s].

#print(bookbasketrules)
summary(bookbasketrules)

## set of 98 rules
##
## rule length distribution (lhs + rhs):sizes
## 2 3 4 5
## 4 60 30 4
##
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 2.000  3.000   3.000   3.347   4.000   5.000
##
## summary of quality measures:
##      support      confidence      lift
## Min.   :0.001025  Min.   :0.8000  Min.   : 60.34
## 1st Qu.:0.001075  1st Qu.:0.8476  1st Qu.:124.48
## Median :0.001200  Median :0.9000  Median :258.56
## Mean   :0.001481  Mean   :0.8925  Mean   :259.63
## 3rd Qu.:0.001612  3rd Qu.:0.9372  3rd Qu.:339.17
## Max.   :0.003574  Max.   :1.0000  Max.   :584.58
##
## mining info:
##      data ntransactions support confidence
## bookbaskets_use      40011  0.001      0.8

# sort rules by lift and confidence
bookbasketrules_sorted <- sort(bookbasketrules, by = c("lift", "confidence"))
#inspect(bookbasketrules_sorted[1:10])

bookbasketrules_sorted_df <- as(bookbasketrules_sorted, "data.frame")
#length(bookbasketrules_sorted_df$rules)
kable(bookbasketrules_sorted_df[1:10,], caption = "Top Ten Association Rules")
```

Table 3: Top Ten Ass

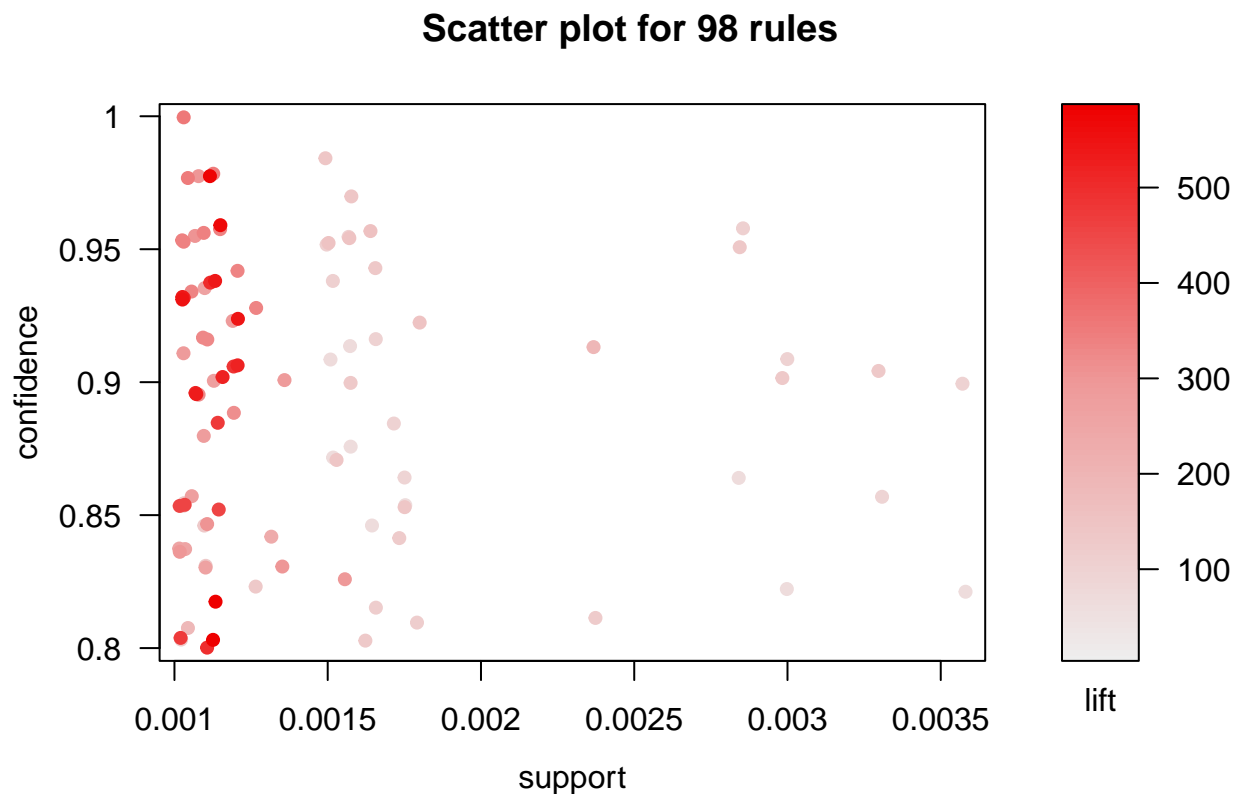
	rules
2	{Key of Knowledge} => {Key of Valor}
1	{Key of Valor} => {Key of Knowledge}
67	{Apollyon: The Destroyer Is Unleashed,Nicolae: The Rise of Antichrist,Tribulation Force: The Continuing Drama of T
6	{Apollyon: The Destroyer Is Unleashed,Nicolae: The Rise of Antichrist} => {Soul Harvest: The World Takes Sides}
71	{Apollyon: The Destroyer Is Unleashed,Left Behind: A Novel of the Earth's Last Days,Tribulation Force: The Contin
9	{Apollyon: The Destroyer Is Unleashed,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul H
68	{Nicolae: The Rise of Antichrist,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of T
72	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides,Tribulation Force: The Contin
11	{Apollyon: The Destroyer Is Unleashed,Left Behind: A Novel of the Earth's Last Days} => {Soul Harvest: The Worl

rules

10 {Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Apollyo

There are ninety-eight rules. We can plot them on a scatter plot. We want to pick higher confidence and higher lift rules, basically all those rules on the top right corner. Most of rule has greater than 0.1% relative support and 80% confidence.

```
plot(bookbasketrules_sorted)
```



We can enhance rules scatter plot by adding order color shading. Higher order rules have higher color shading. The highest number of order is 5.

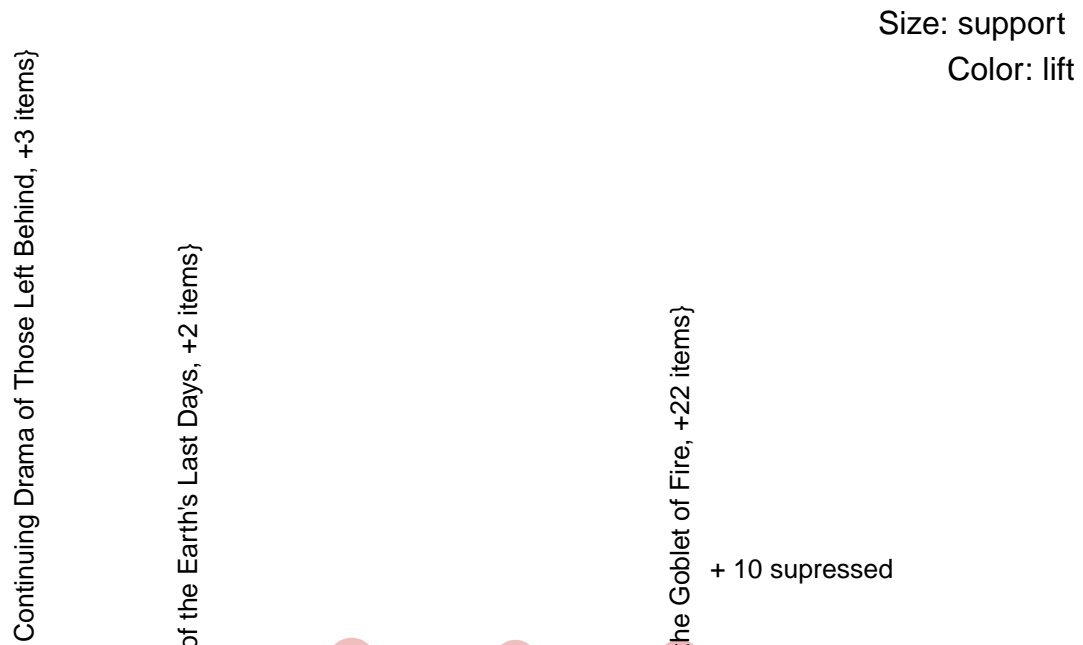
```
##A two-key plot  
plot(bookbasketrules_sorted, shading="order", control=list(main="Two-key plot"))
```



We can also create a grouped matrix plot limit to top five rules.

```
## Grouped Matrix Plot  
#plot(bookbasketrules_sorted, method="grouped")  
plot(bookbasketrules_sorted, method="grouped", control=list(k=5))
```

Grouped Matrix for 98 Rules

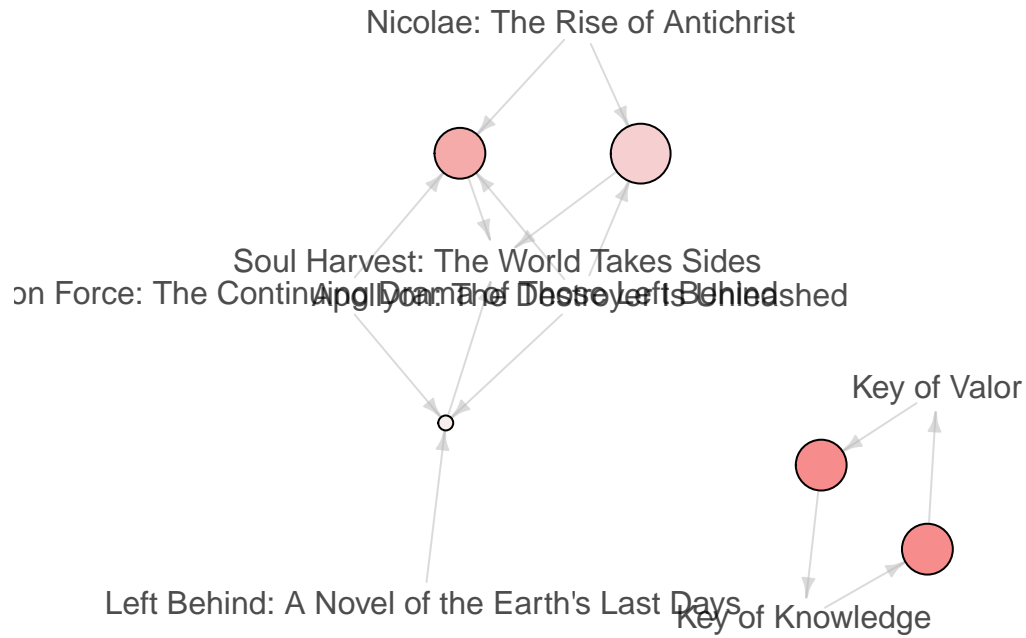


Graph Based Visualizations are helpful to visualize smaller set of rules. We can use it to display first ten rules. For example, “Key of Knowledge” on the left hand side leads to “key of Light” on the right hand side.

```
plot(bookbasketrules_sorted[1:5], method="graph")
```

Graph for 5 rules

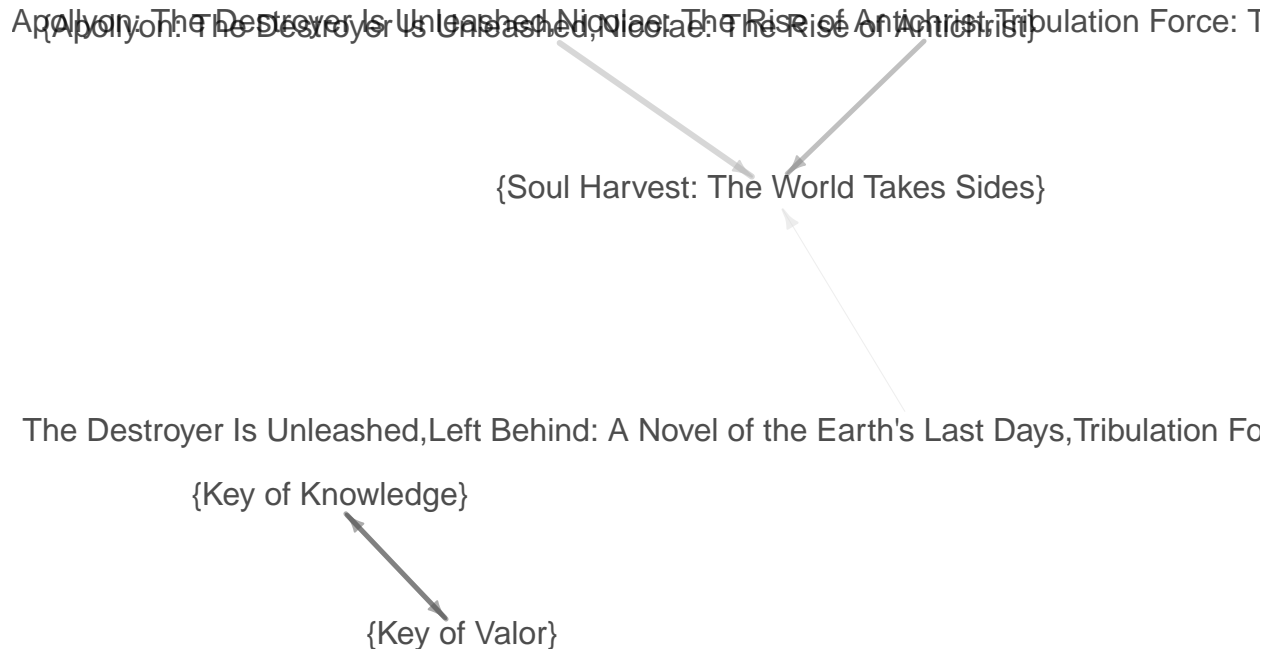
size: support (0.001 – 0.001)
color: lift (548.279 – 584.576)



```
plot(bookbasketrules_sorted[1:5], method="graph", control=list(type="itemsets"))
```

Graph for 5 rules

width: support (0.001 – 0.001)
color: lift (548.279 – 584.576)



If we want to exclude title “Wild Animus” then we can rerun Apriori without this title.

```
bookbasketsRule_noAnimus <- apriori(bookbaskets_use, parameter = list(support = 0.001, confidence = 0.8,
appearance = list(none = c("Wild Animus"))))
```

```
## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##          0.8   0.1   1 none FALSE                TRUE     5   0.001    2
## maxlen target  ext
##      10  rules FALSE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##    0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 40
##
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[142363 item(s), 40011 transaction(s)] done [0.63s].
## sorting and recoding items ... [1471 item(s)] done [0.05s].
## creating transaction tree ... done [0.05s].
## checking subsets of size 1 2 3 4 5 done [0.05s].
## writing ... [98 rule(s)] done [0.03s].
## creating S4 object ... done [0.05s].
```



```
bookbasketsRule_noAnimus_sorted <- sort(bookbasketsRule_noAnimus, by = c("lift", "confidence"))
#inspect(bookbasketsRule_noAnimus_sorted[1:20])
bookbasketsRule_noAnimus_sorted_df <- as(bookbasketsRule_noAnimus_sorted, "data.frame")
kable(bookbasketsRule_noAnimus_sorted_df[1:20,], caption = "Top Twenty Books without 'Wild Animus'")
```

Table 4: Top Twenty Books without 'Wild Animus'

	rules
2	{Key of Knowledge} => {Key of Valor}
1	{Key of Valor} => {Key of Knowledge}
67	{Apollyon: The Destroyer Is Unleashed,Nicolae: The Rise of Antichrist,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
6	{Apollyon: The Destroyer Is Unleashed,Nicolae: The Rise of Antichrist} => {Soul Harvest: The World Takes Sides}
71	{Apollyon: The Destroyer Is Unleashed,Left Behind: A Novel of the Earth's Last Days,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
9	{Apollyon: The Destroyer Is Unleashed,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
68	{Nicolae: The Rise of Antichrist,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
72	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
11	{Apollyon: The Destroyer Is Unleashed,Left Behind: A Novel of the Earth's Last Days} => {Soul Harvest: The World Takes Sides}
10	{Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Apollyon: The Destroyer Is Unleashed}
7	{Nicolae: The Rise of Antichrist,Soul Harvest: The World Takes Sides} => {Apollyon: The Destroyer Is Unleashed}
12	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides} => {Apollyon: The Destroyer Is Unleashed}
66	{Apollyon: The Destroyer Is Unleashed,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
75	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
3	{Key of Knowledge} => {Key of Light}
20	{Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Nicolae: The Rise of Antichrist}
76	{Left Behind: A Novel of the Earth's Last Days,Nicolae: The Rise of Antichrist,Tribulation Force: The Continuing Drama of Those Left Behind} => {Nicolae: The Rise of Antichrist}
14	{Apollyon: The Destroyer Is Unleashed,Tribulation Force: The Continuing Drama of Those Left Behind} => {Nicolae: The Rise of Antichrist}
16	{Apollyon: The Destroyer Is Unleashed,Left Behind: A Novel of the Earth's Last Days} => {Nicolae: The Rise of Antichrist}
22	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides} => {Nicolae: The Rise of Antichrist}

We also want to exclude those books in a series. Then we can rerun the apriori exclude books in a series If we also want to exclude those books in a series then we can rerun the Apriori algorithm and exclude those books in a series In this case when we exclude the Stephanie Plum crime series, Lord of the Ring series, Harry Potter series, The Green Miles series. We used the same 0.1% support and 80% confidence threshold.

```
excludeList <- c("Wild Animus", "Three To Get Deadly : A Stephanie Plum Novel", "Four to Score", "Seven
               "The Fellowship of the Ring", "The Return of the King", "The Two Towers",
               "Harry Potter and the Sorcerer's Stone", "Harry Potter and the Order of the Phoenix",
               "Harry Potter and the Chamber of Secrets", "Harry Potter and the Goblet of Fire",
               "The Green Mile: Night Journey", "The Green Mile: Coffey on the Mile", "The Green Mile
bookbasketsRule_series <- apriori(bookbaskets_use, parameter = list(support = 0.001, confidence = 0.8, m

## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##           0.8   0.1   1 none FALSE                TRUE         5   0.001     2
## maxlen target  ext
##           5 rules FALSE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##       0.1 TRUE TRUE  FALSE TRUE     2    TRUE
```

```
##
## Absolute minimum support count: 40
##
## set item appearances ...[18 item(s)] done [0.00s].
## set transactions ...[142363 item(s), 40011 transaction(s)] done [0.63s].
## sorting and recoding items ... [1457 item(s)] done [0.06s].
## creating transaction tree ... done [0.03s].
## checking subsets of size 1 2 3 4 done [0.08s].
## writing ... [40 rule(s)] done [0.02s].
## creating S4 object ... done [0.05s].

summary(bookbasketsRule_series)

## set of 40 rules
##
## rule length distribution (lhs + rhs):sizes
## 2 3 4
## 4 24 12
##
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      2.0      3.0      3.0      3.2      4.0      4.0
##
## summary of quality measures:
##      support      confidence      lift
## Min.      :0.001025  Min.      :0.8000  Min.      :121.4
## 1st Qu.:0.001025  1st Qu.:0.8504  1st Qu.:309.2
## Median :0.001100  Median :0.8989  Median :455.1
## Mean    :0.001117  Mean    :0.8906  Mean     :386.3
## 3rd Qu.:0.001150  3rd Qu.:0.9329  3rd Qu.:516.1
## Max.    :0.001550  Max.     :1.0000  Max.     :584.6
##
## mining info:
##      data ntransactions support confidence
## bookbaskets_use      40011  0.001      0.8

bookbasketsRule_series_sorted <- sort(bookbasketsRule_series, by = c("lift", "confidence"))
#inspect(bookbasketsRule_series_sorted)
bookbasketsRule_series_sorted_df <- as(bookbasketsRule_series_sorted, "data.frame")
kable(bookbasketsRule_series_sorted_df[1:15,], caption = "Top Books not in a series")
```

Table 5: Top Books not in a series

	rules
2	{Key of Knowledge} => {Key of Valor}
1	{Key of Valor} => {Key of Knowledge}
31	{Apollyon: The Destroyer Is Unleashed,Nicolae: The Rise of Antichrist,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
6	{Apollyon: The Destroyer Is Unleashed,Nicolae: The Rise of Antichrist} => {Soul Harvest: The World Takes Sides}
35	{Apollyon: The Destroyer Is Unleashed,Left Behind: A Novel of the Earth's Last Days,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
9	{Apollyon: The Destroyer Is Unleashed,Tribulation Force: The Continuing Drama of Those Left Behind} => {Soul Harvest: The World Takes Sides}
32	{Nicolae: The Rise of Antichrist,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Apollyon: The Destroyer Is Unleashed}
36	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Apollyon: The Destroyer Is Unleashed}
11	{Apollyon: The Destroyer Is Unleashed,Left Behind: A Novel of the Earth's Last Days} => {Soul Harvest: The World Takes Sides}
10	{Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Drama of Those Left Behind} => {Apollyon: The Destroyer Is Unleashed}
7	{Nicolae: The Rise of Antichrist,Soul Harvest: The World Takes Sides} => {Apollyon: The Destroyer Is Unleashed}
12	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides} => {Apollyon: The Destroyer Is Unleashed}

	rules
30	{Apollyon: The Destroyer Is Unleashed,Soul Harvest: The World Takes Sides,Tribulation Force: The Continuing Dram
39	{Left Behind: A Novel of the Earth's Last Days,Soul Harvest: The World Takes Sides,Tribulation Force: The Continui
3	{Key of Knowledge} => {Key of Light}

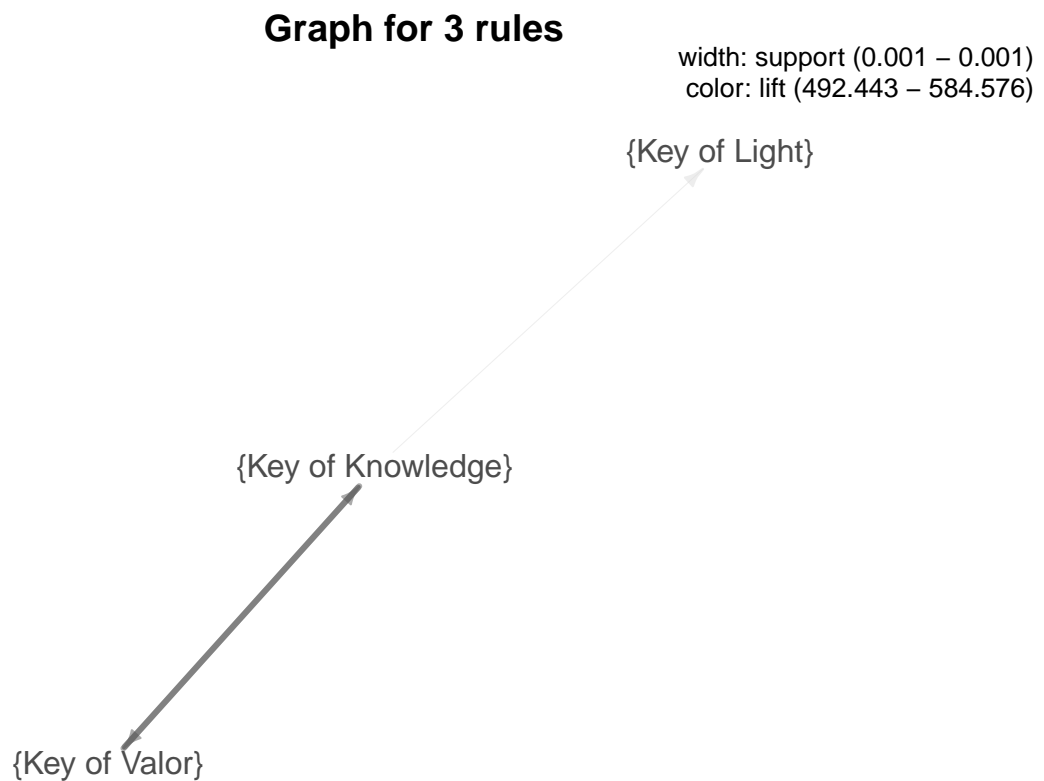
If we do a subset to list all rules from title “Key of Knowledge” then we get list of three rules related to title “Key of Knowledge”. We can do the same for other interesting titles. The itemset plot shows this subset of rules on one title very well.

```
subsetRules <- subset(bookbasketsRule_series_sorted, items %in% "Key of Knowledge")
subsetRules_df <- as(subsetRules, "data.frame")
kable(subsetRules_df, caption = "Key of Knowledge subset")
```

Table 6: Key of Knowledge subset

	rules	support	confidence	lift
2	{Key of Knowledge} => {Key of Valor}	0.0011247	0.8181818	584.5763
1	{Key of Valor} => {Key of Knowledge}	0.0011247	0.8035714	584.5763
3	{Key of Knowledge} => {Key of Light}	0.0010997	0.8000000	492.4431

```
plot(subsetRules, method="graph", control=list(type="itemsets"))
```



Commentary

- We listed top twenty bestselling books from customers' transaction data set.
- Oraph Book Club has fifty-six books in our transaction data set. From those fifty-six books we discovered rules that book club reader would like to read. We can use the list to appeal book club readers. The resulting rules are listed above.
- The book club rules have smaller support because they are not popular books in our customers' transaction data set. If we use transactions with more than one book and less than two hundred books, then we can get stronger rules, those with higher relative support and confidence. And if we excluded "Wild Animus" book, and books in a series then we get books from the same author or books on the same topic. For example, "Key of Knowledge" and "Key of Valor" both written by same author. Top fifteen associated rules are listed above.

END