BTS | Barcelona Technology School

Data Driven Business

Recommendation Systems

Master in Big Data Solutions

Trainer's name: Atilla YARDIMCI

Email address: atilla.yardimci@bts.tech

BTS | Barcelona Technology School

Today…

09:00 – 09:30 = Recommendation Systems

09:30 – 10:50 = Data-Driven Use Case (Ebru Umut Deniz)

11:00 – 13:00 = Recommendation Systems

Today…


Understand the definition, benefits and techniques of
Recommendation Systems…

BTS | Barcelona Technology School

Actually, recommendation systems are everywhere, and we can see them in our whole life. When we are shopping or watching movies or when we use our social media accounts such as LinkedIn, Instagram or Facebook.

"Other Movies You May Enjoy"
Netflix

"Jobs You May Be Interested In"
LinkedIn

"We are leaving the **age of information** and entering the **age of recommendation**"
Chris Anderson - The Long Tail

"People You May Know"
Facebook

"Customer who bought this item also bought …"
Amazon

BTS | Barcelona Technology School

Recommendation Systems

- A decision-making strategy for users under complex information environments.
- A tool that helps users search through records of knowledge which is related to users' interest and preference.
- A system able to suggest item to end-users.

...

- Most of e-commerce institutions such as Google, Amazon, Facebook and so forth use recommendation systems..

…. what is this recommendation system ?

- First of all, it is a decision-making strategy for users under complex information environments.
- Secondly, it is a tool that helps users search through records of knowledge which is related to users' interest and preference
- And finally it is a system able to suggest an item to end-users.

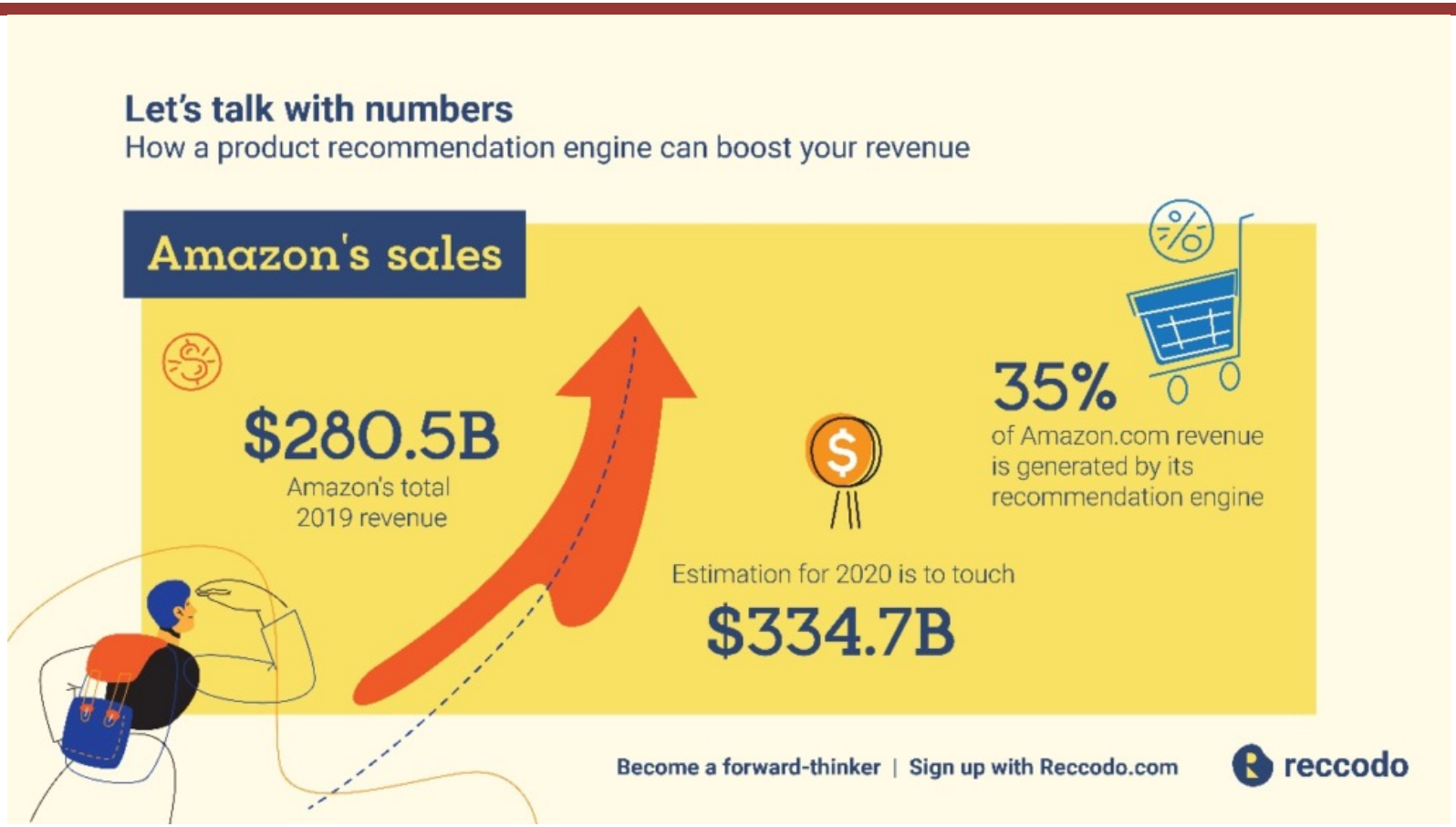
BTS | Barcelona Technology School

Benefits…

https://research.aimultiple.com/recommendation-system/

The majority of experiences and studies indicate that using recommendation systems can result in increased revenue or conversion, increased user satisfaction, increased loyalty, and reduced churn for your company / any company.

Benefits…

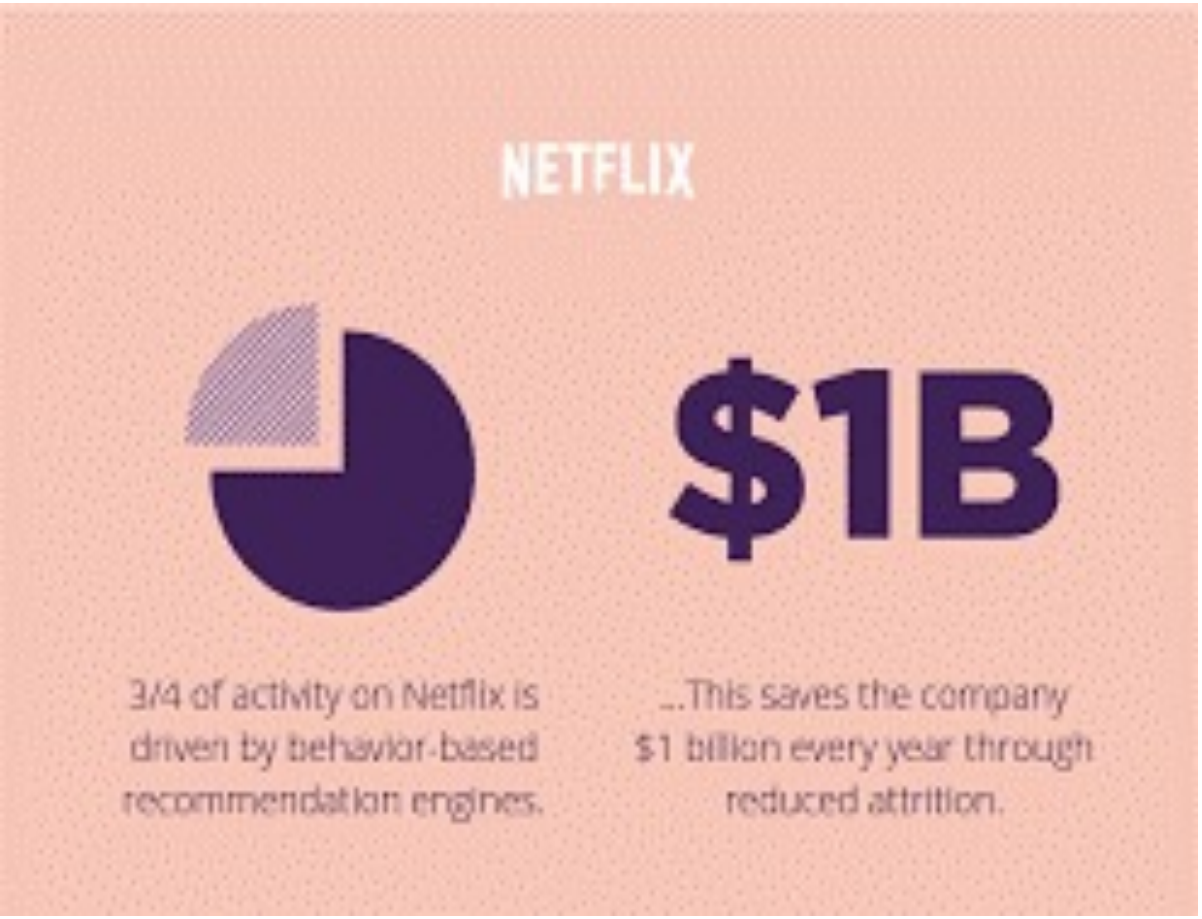


https://reccodo.com/amazons-recommendation-system/

For instance, 35% of Amazon's sales, which equals to 280.5 Billion USD, is generated by using recommendation system and also they expect their revenue to hit 334.7 billion dollars in 2020.

BTS | Barcelona Technology School

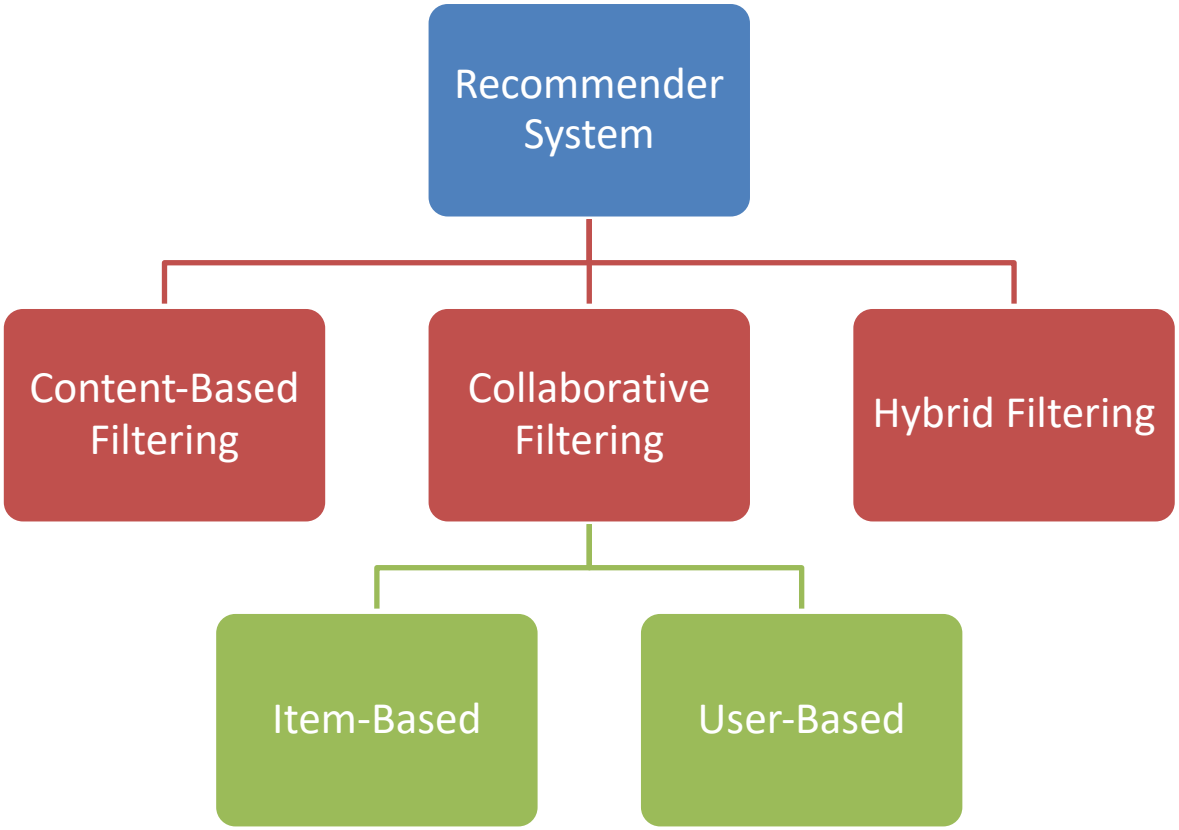
Benefits…



Another example is Netflix… 75% of its activity is generated by its recommendation system and this system assists them in saving $1 billion every year..

https://www.plytix.com/blog/revenue-optimization

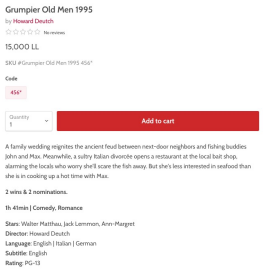
Techniques

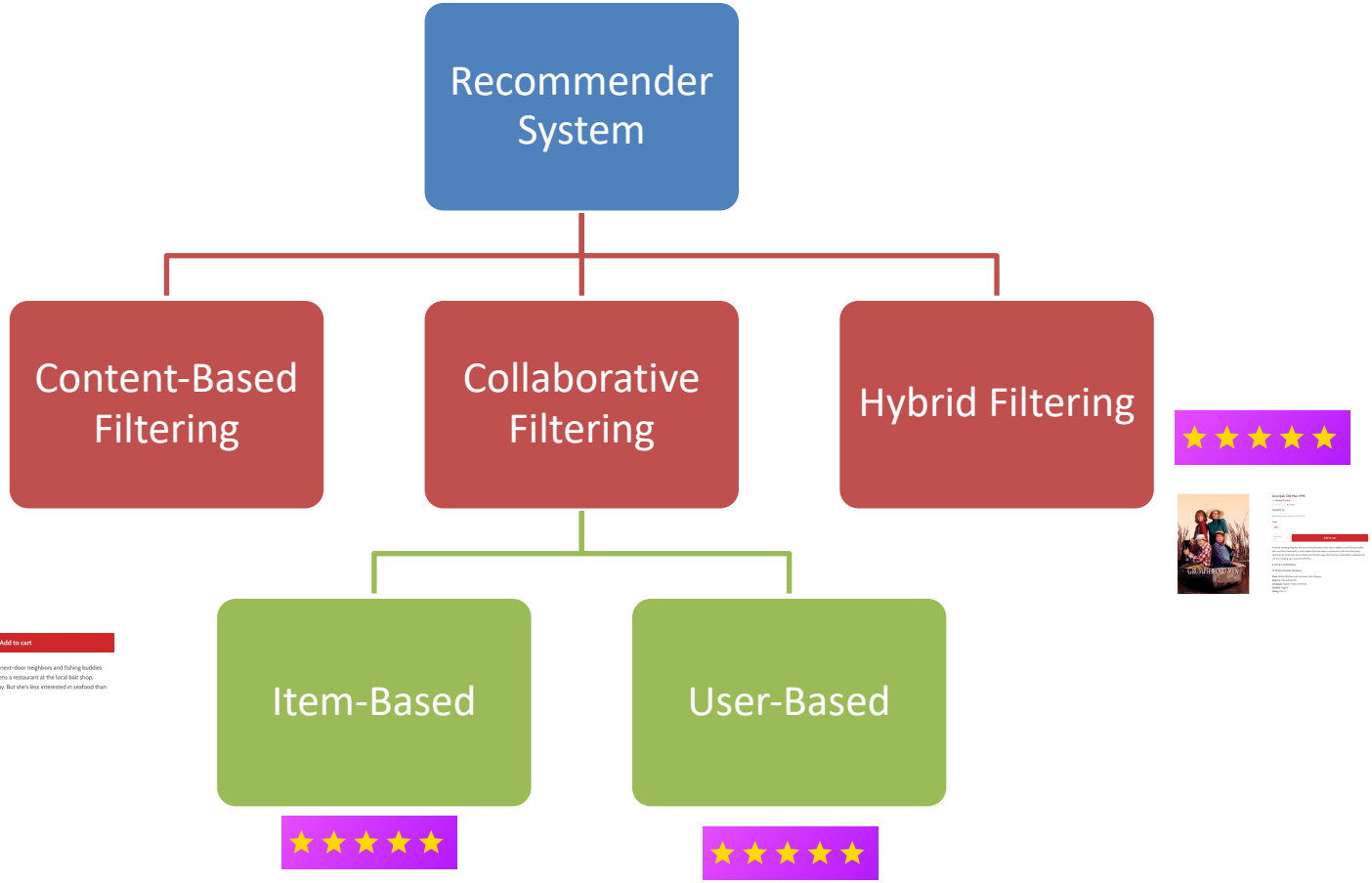


There are basically 3 approaches of RC Systems which are Content-Based Filtering, Collaborative Filtering, and Hybrid Filtering. We will deep into of these approaches in the next slides.

Techniques

https://www.youtube.com/watch?v=n3RKsY2H-NE



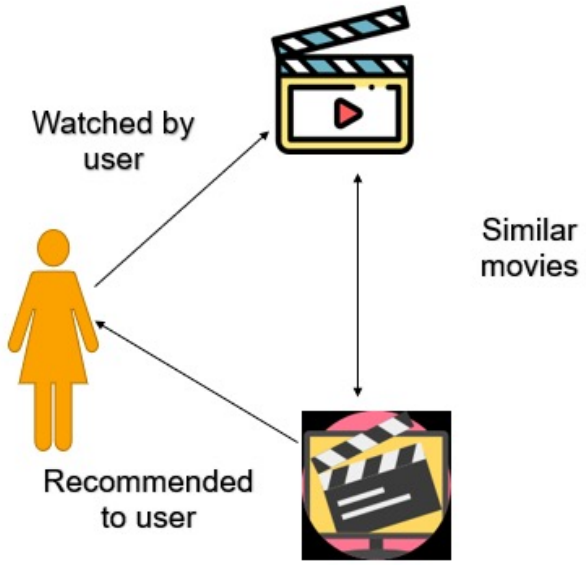


There are basically 3 approaches of RC Systems which are Content-Based Filtering, Collaborative Filtering, and Hybrid Filtering. We will deep into of these approaches in the next slides.

BTS | Barcelona Technology School

Content-Based Filtering

- Recommendations based on information on the content of items.
- Builds a model of the users' preferences from instances based on a featural description of content by using machine learning algorithms.
- Tries to recommend items similar to another item a given user has liked previously.



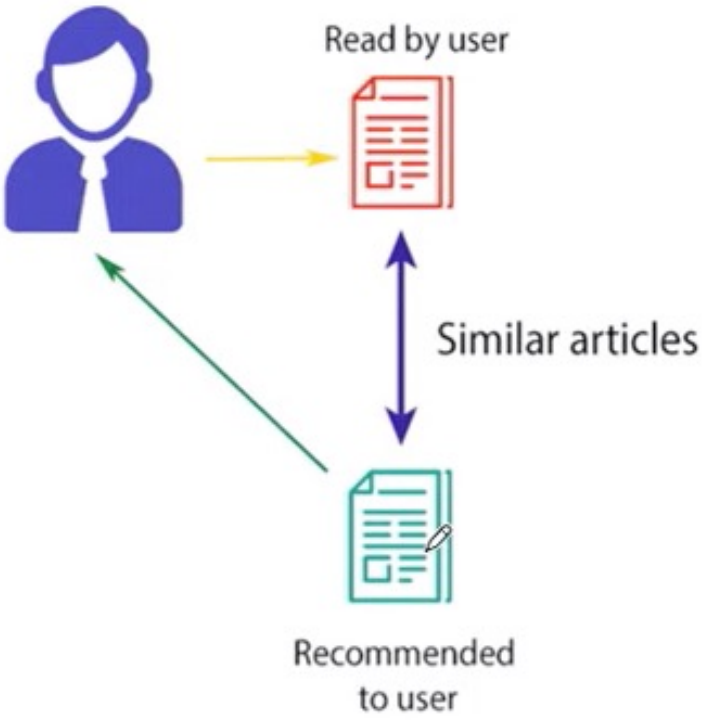
This filtering method is based on the description of a product and the content of the user's preferred options. In a content-based recommendation system, keywords are used to describe the product. Algorithms try to recommend products like what a user has liked in the past. The Content Based Filtering idea is based on the idea that if you like a product, you will also like another product like this product.

For example, If you like Titanic, you can also like Romeo and Juliet. The assuption is that their contents are similar, that is, they are both romantic movies.

BTS | Barcelona Technology School

Content-Based **Filtering**

Filtering is important here.

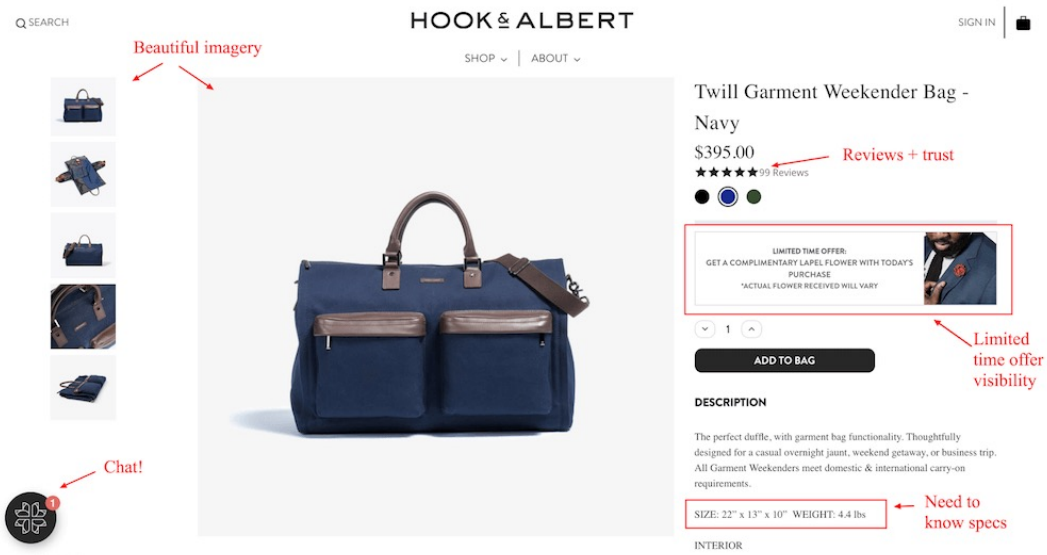


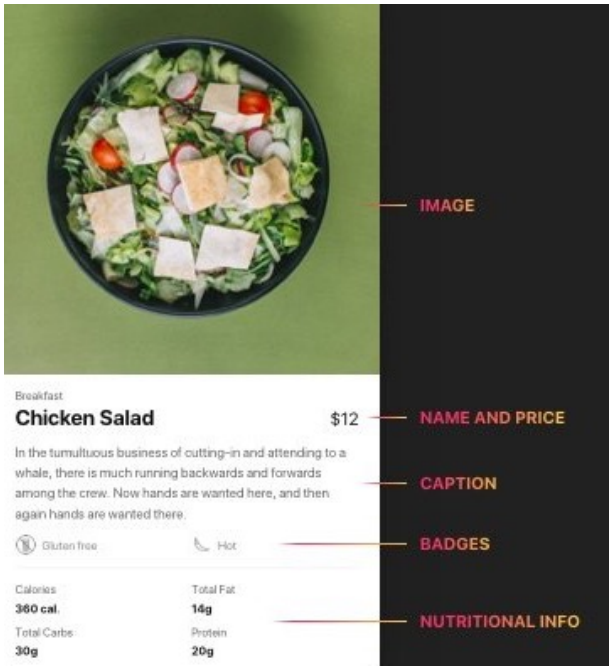
Top words in highly rated romance novels
(By most- to least-used)

kiss
nod
mouth
grin
leg
lift
sex
damn
lower
sweet
chuckle
send
tongue
low
chin
growl
moan
ass
clench
fully
sensation
threat
wince
anticipation

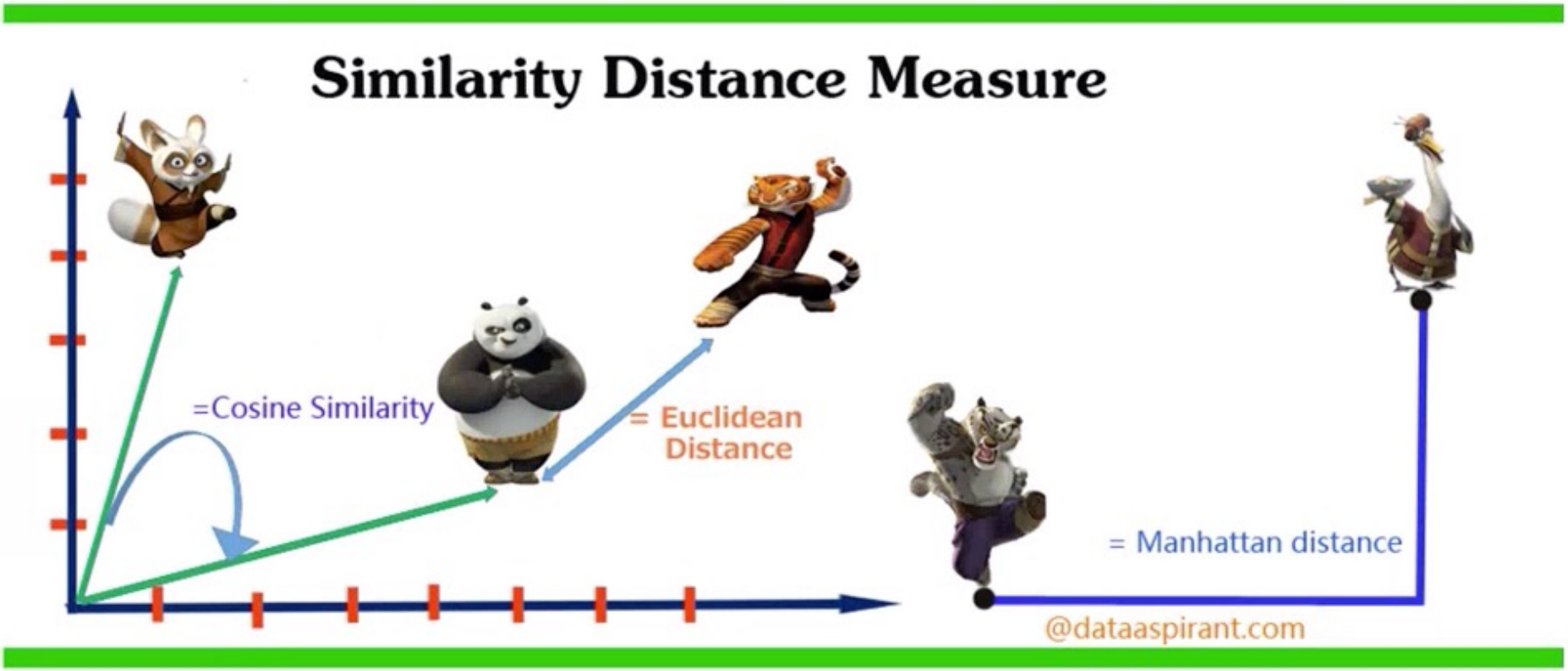
BTS | Barcelona Technology School

Content-Based **Filtering**





Content-Based **Filtering**



$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}^\intercal}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} \mathbf{x}_i \cdot \mathbf{y}^\intercal{}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{x}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{y}_i)^2}}$$

BTS | Barcelona Technology School

Content-Based Filtering

One of the best-known measure is the Term Frequency/Inverse Document Frequency (TF-IDF)

Number of times term *t* appears in a document *d*.

$$\mathrm{tf}(t, d) = \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in D\}|}$$

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot idf(t, D)$$

Number of times of total documents in documents (d) with term t.

https://www.youtube.com/watch?v=G1bof7UL9RU

- TF-IDF method is the most-known measure in Content-Based Filtering.
- TF gives us the frequency of term in a document.
- IDF means inverse document frequency. If the term *t* appears in every document of the corpus, idf of t is equal to 0.
- And finally, by multipliying TF and IDF results we found, we can reach TF-IDF values for each t in d.

BTS | Barcelona Technology School

Content-Based Filtering

Total number of words in Document1 = 7

Document1 : '' The man is walking on the road.''

Document2 : '' The woman is running on the sidewalk.''

Total number of words in Document2 = 7

	Count		TF		IDF	TF x IDF	
Words	D1	D2	D1	D2		D1	D2
The	1	1	1/7	1/7	log(2/2) = 0	0	0
man	1	0	1/7	0	log(2/1) = 0.3	0.043	0
woman	0	1	0	1/7	log(2/1) = 0.3	0	0.043
is	1	1	1/7	1/7	log(2/2) = 0	0	0
walking	1	0	1/7	0	log(2/2) = 0	0	0
walking	1	1	1/7	1/7	log(2/2) = 0	0	0
on	1	1	1/7	1/7	log(2/2) = 0	0	0
the	1	1	1/7	1/7	log(2/2) = 0	0	0
road	1	0	1/7	0	log(2/1) = 0.3	0.043	0
sidewalk	0	1	0	1/7	log(2/1) = 0.3	0	0.043

- We have 2 documents such as : **'' The man is walking on the road.'' and '' The woman is running on the sidewalk.''**
- If we want to calculate of TF-IDF for each word (term, t), we should firstly calculate frequencies of all words which means TF.
- For TF of the term ''The'':
  We can easily see that the term ''The'' occurs 1 time in D1, the term ''man'' occurs 1 time in D1, the term ''woman'' appears 0 time in d1, and so on...
  TF result of the term ''The'' is 1 over 7 because we have 7 words in D1 and the term ''The'' appears 1 time in D1
- For IDF of the term ''The'':
  Term ''The'' , as you see in table, N is equal to 2 because we have 2 sentences (document), and also we have 2 sentences (document) which have term ''The'',
  So the log value of 2 over 2 is equal to zero. If we multiply 1 over 7 by zero, we can reach the TF-IDF value of the term ''The'' which is equal to zero.
- And we easily apply this formula for all terms in our documents.
- We can understand that TF-IDF values of common words are equal to 0, which shows they are not significant. But, the TF-IDF of "man", "woman", "road", and "sidewalk" are greater than 0, so we can understand that these words (items) have more significance than the others/

BTS | Barcelona Technology School

Content-Based **Filtering**

Count Vectorizer

	blue	bright	sky	sun
Doc1	1	0	1	0
Doc2	0	1	0	1

TD-IDF Vectorizer

	blue	bright	sky	sun
Doc1	0.707107	0.000000	0.707107	0.000000
Doc2	0.000000	0.707107	0.000000	0.707107

It is to represent the **TEXT** with numbers.

BTS | Barcelona Technology School

https://www.jeremyjordan.me/identifying-related-bodies-of-text-using-tf-idf-vectorization/

Content-Based **Filtering**

Term Counts

	"TFIDF"	"IS"	"A"	"SIMILARITY"	"WEATHER"	"TO"
Document 1	4	9	12	6	0	3
Document 2	5	7	15	3	0	2
Document 3	0	8	19	1	3	4
Total	9	24	46	10	3	9

Bias occurs for high-frequency words.

Let's say we have a collection of documents:

- **Document 1:** "TF-IDF vectorization is a useful tool for text analytics. It allows you to quantify the similarity of different documents."
- **Document 2:** "You can use cosine similarity to analyze TF-IDF vectors and cluster text documents based on their content."
- **Document 3:** "The weather today is looking quite bleak. It's a great day to write a blog post!"

BTS | Barcelona Technology School

https://www.jeremyjordan.me/identifying-related-bodies-of-text-using-tf-idf-vectorization/

Content-Based **Filtering**

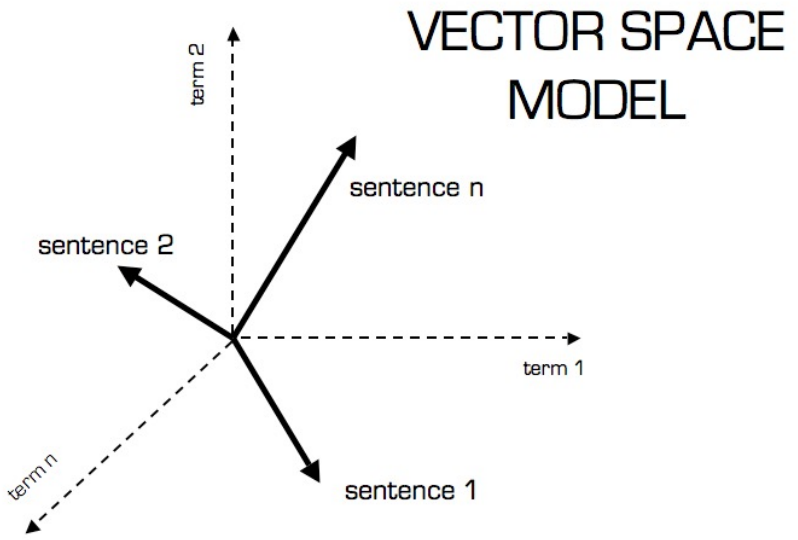
Term Counts

	"TFIDF"	"IS"	"A"	"SIMILARITY"	"WEATHER"	"TO"
Document 1	4	9	12	6	0	3
Document 2	5	7	15	3	0	2
Document 3	0	8	19	1	3	4
Total	9	24	46	10	3	9

Cosine Similarity

Now that we have a TF-IDF vector for each document, we can quantify the similarity between two vectors using cosine similarity. The image below is quite instructive for providing the motivation behind this technique.

$$\begin{bmatrix} 4/34 \\ 9/34 \\ 12/34 \\ 6/34 \\ 0/34 \\ 3/34 \end{bmatrix} \circ \begin{bmatrix} 101/9 \\ 101/24 \\ 101/46 \\ 101/10 \\ 101/3 \\ 101/9 \end{bmatrix} = \text{TF-IDF vector for Document 1}$$

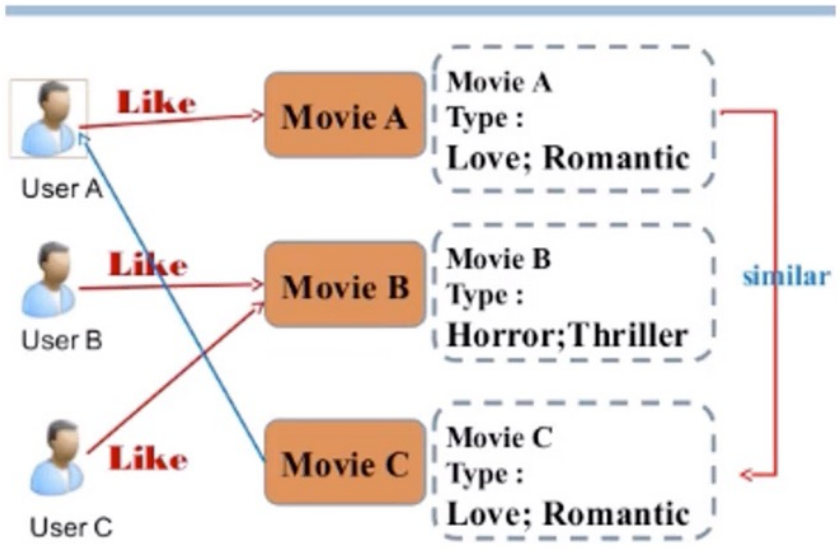


BTS | Barcelona Technology School

Collaborative Filtering

Filtering is important here.

If the user like a movie, the system recommends similar movies



Restorants

BTS | Barcelona Technology School

Collaborative Filtering

	Movie 1	Movie 2	Movie 3	.	.	.	Movie n
User 1	1	2					4
User 2	5		3				3
User 3	3	5	5				
.							
.							
.							
User m		2	1				1

The objective here is to fill the blanks

BTS | Barcelona Technology School

Collaborative Filtering

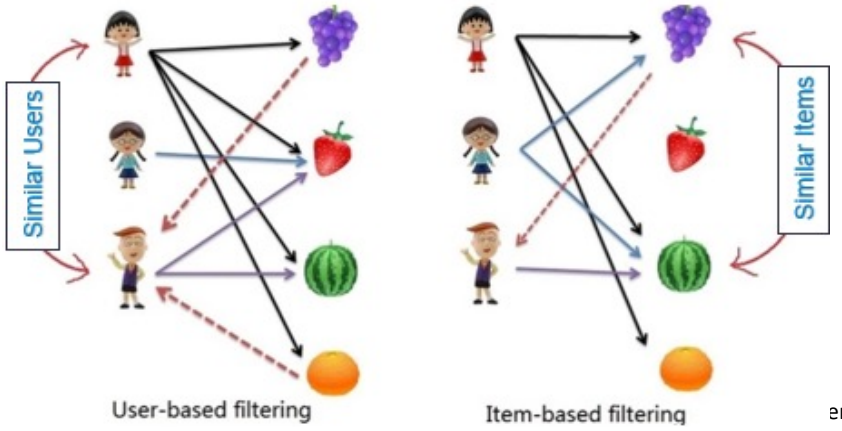
The Corretation…

	Movie 1	Movie 2	Movie 3	.	.	.	Movie n
User 1	1	2					4
User 2	5		3				3
User 3	3	5	5				
.							
.							
.							
User m		2	1				1

BTS | Barcelona Technology School

Collaborative Filtering

- A tool for filtering out items that a user might like based on the reactions of other users.
- In User-Based Filtering, the recommendation based on similarity of users.
- In Item-Based Filtering, the recommendation based on similarity of items.



ence.wordpress.com/2016/07/22/recommender-systems-101/

For user-based filtering:

- Assume there are three users: A, B, and C. User A and C are close in user-based filtering since they all like Strawberry and Watermelon. Now, user A enjoys grapes and oranges as well. As a result, the user-based filtering will recommend Grapes and Orange to user C.

For item-based filtering:

- Grapes and Watermelon will form the neighborhood of the related object in item-based filtering, which means that regardless of users, different items that are similar will form a neighborhood.

Calculating user similarity

A most used similarity calculated method in user-based filtering is Pearson correlation.

$$sim(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P}(r_{a,p} - \bar{r}_a)^2}\sqrt{\sum_{p \in P}(r_{b,p} - \bar{r}_b)^2}}$$

Where

a, b = users
$r_{a,p}$ = rating of user *a* for item *p*
$\bar{r}_a$ = *average rating of user a*
$r_{b,p}$ = rating of user *b* for item *p*
$\bar{r}_b$ = *average rating of user b*
P = set of items, rated both by *a* and *b*

Possible similarity values are between -1 and 1.

- 1 shows a strong positive relationship.
- -1 a strong negative relationship.
- 0 shows no relationship at all

BTS | Barcelona Technology School

Calculating user similarity

Let's make an example…

	Item1	Item2	Item3	Item4	Item5	Pearson correlation
User A	5	3	4	4	?	
User B	3	1	2	3	5	Sim(A,B)=0.85
User C	4	3	4	2	3	sim(A,C)=0.43
User D	3	3	1	5	4	sim(A,D)=0.00
User E	1	5	5	2	1	sim(A,E)=-0.79

User A is much more similar with User B than the other users…

We can recommend Item 5 to User A…

Calculating item similarity

A most used similarity calculated method in item-based filtering is Cosine Similarity Measure.

$$\text{sim}(x, y) = \frac{x.y}{||x||.||y||}$$

Where

$||x||$ = The Euclidean norm of vector x= $(x_1, x_2, x_3, \ldots, x_n)$, defined as $\sqrt{x_1^2 + x_2^2 + x_3^2 \ldots + x_n^2}$

$||y||$ = The Euclidean norm of vector x= $(y_1, y_2, y_3, \ldots, y_n)$, defined as $\sqrt{y_1^2 + y_2^2 + y_3^2 \ldots + y_n^2}$

Possible cosine similarity values are between 0 and 1.

- If cosine similarity value is equal to 0, then the two vectors have no match.
- If cosine similarity value is equal to 1, then the smaller the angle and the greater the match between vectors.

Cosine similarity measures the similarity between two n-dimensional vectors based on the angle between them. Cosine-based measure is also, widely used in the fields of texts mining to compare two text documents that are represented as vectors of terms.

BTS | Barcelona Technology School

Calculating item similarity

$$\text{sim}(x, y) = \frac{x.y}{||x||.||y||}$$

Suppose that we have two documents such as *x* and *y*. Let's find how similar they are…

x= (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

x.y = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) . (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)
= 5.3 + 0.0 + 3.2 + 0.0 +2.1 + 0.1 + 0.0 + 2.1 + 0.0 + 0.1
= 25

$$||x|| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$||y|| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

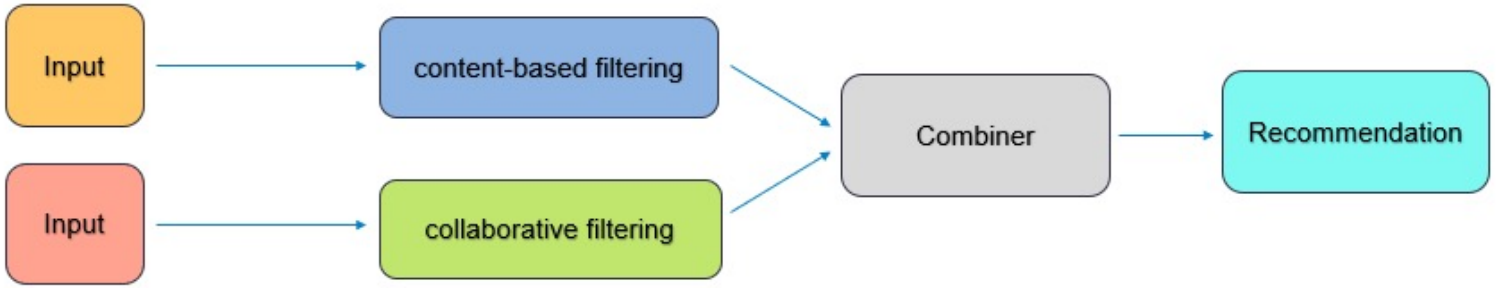
$$\text{sim}(x, y) = \frac{25}{6.48 \times 4.12} = 0.94$$

These two documents are quite **similar!**

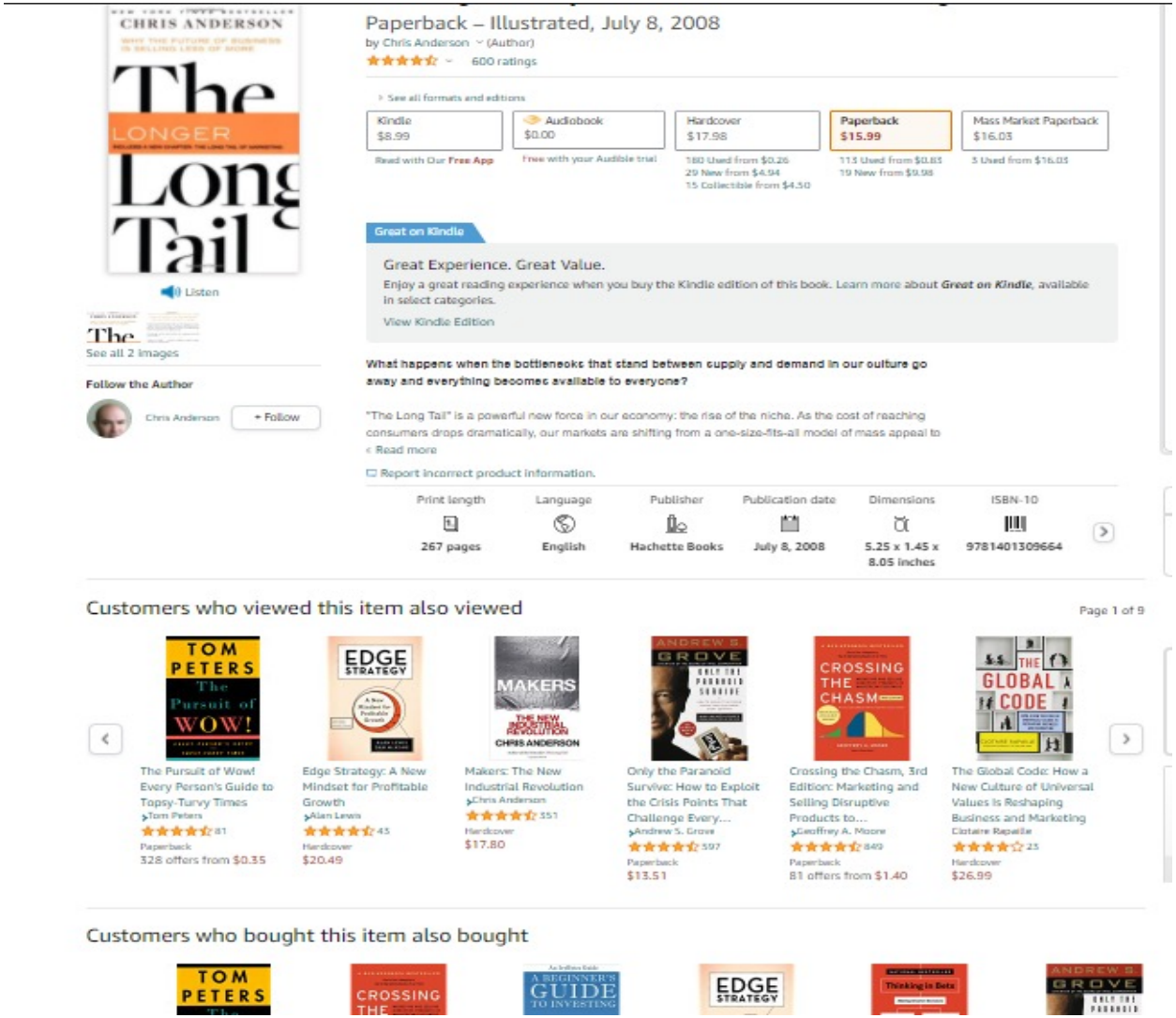
- Let's make an example… Suppose that we have two documents such as x and y. Let's find how similar they are…
- We know that x and y are vectors.
- Firstly, we apply matrix multiplication, and we find the result as 25.
- Secondly, we calculate the root of the sum of all squared items in document x. And we also apply this formula for also all items in document y.
- Finally, we divide 25 by the multiplication of 6.48 by 4.12
- And the final result is 0.94 and it is very close to 1 which means these two documents are very similar.

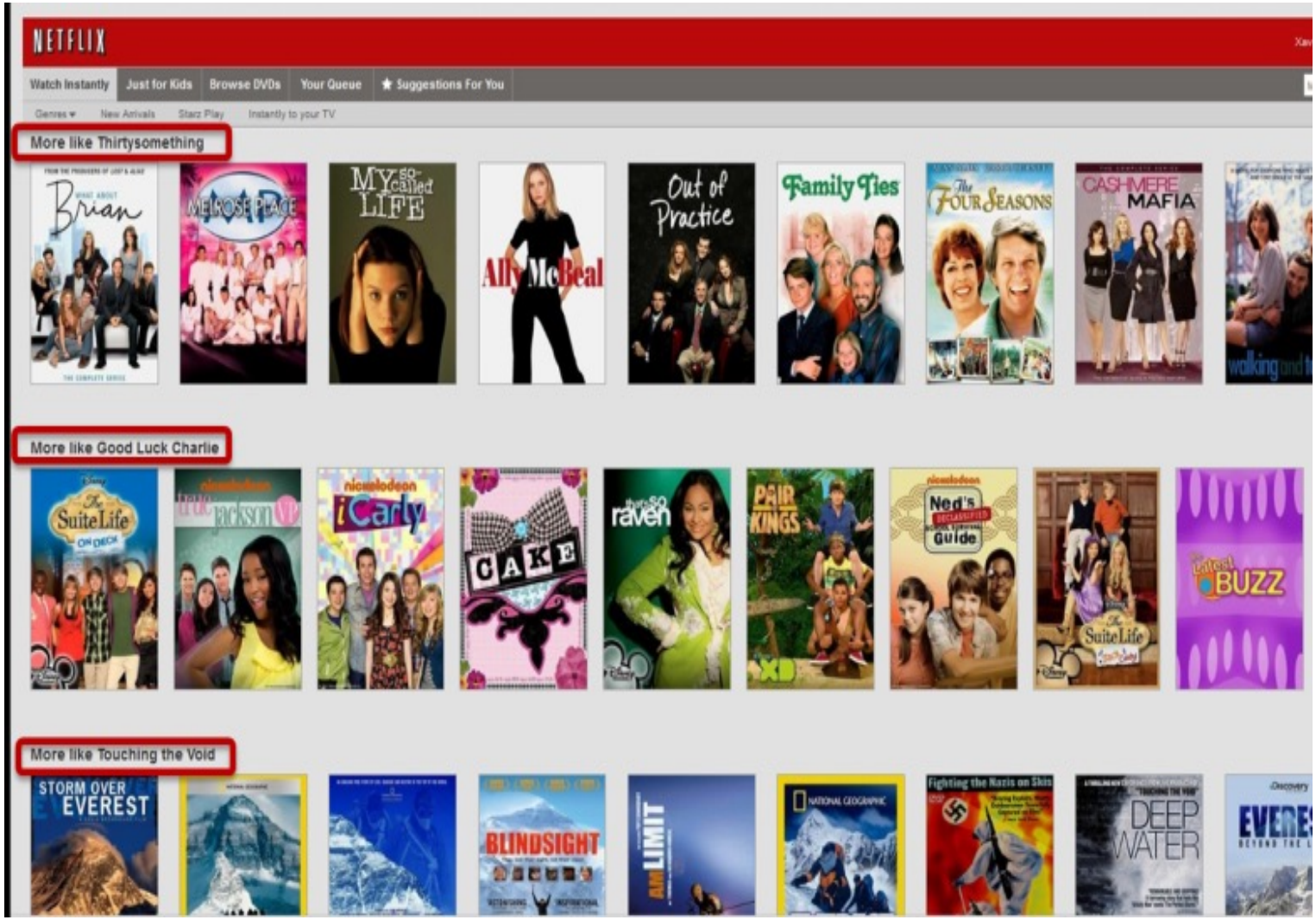
Hybrid Filtering

- A technique which combines content-based filtering algorithms and collaborative filtering algorithms.
- Netflix is a good example of the hybrid recommendation system.
- Netflix makes recommendations based on similar users' surfing and search patterns (i.e. collaborative filtering) as well as recommending movies of similar content that a user previously ranked highly.(i.e. content-based filtering).



http://dataconomy.com/2015/03/an-introduction-to-recommendation-engines/


CHRIS ANDERSON
The LONGER Long Tail
Listen
See all 2 images
Follow the Author
Chris Anderson
+ Follow
Paperback – Illustrated, July 8, 2008
by Chris Anderson (Author)
600 ratings
See all formats and editions
Kindle $8.99
Audiobook $0.00
Hardcover $17.98
Paperback $15.99
Mass Market Paperback $16.03
Great on Kindle
Great Experience. Great Value.
Enjoy a great reading experience when you buy the Kindle edition of this book. Learn more about Great on Kindle, available in select categories.
View Kindle Edition
What happens when the bottlenecks that stand between supply and demand in our culture go away and everything becomes available to everyone?
"The Long Tail" is a powerful new force in our economy: the rise of the niche. As the cost of reaching consumers drops dramatically, our markets are shifting from a one-size-fits-all model of mass appeal to
Read more
Report incorrect product information.
Print length 267 pages
Language English
Publisher Hachette Books
Publication date July 8, 2008
Dimensions 5.25 x 1.45 x 8.05 inches
ISBN-10 9781401309664
Customers who viewed this item also viewed
Page 1 of 9
The Pursuit of Wow! Every Person's Guide to Topsy-Turvy Times
Edge Strategy: A New Mindset for Profitable Growth
Makers: The New Industrial Revolution
Only the Paranoid Survive: How to Exploit the Crisis Points That Challenge Every...
Crossing the Chasm, 3rd Edition: Marketing and Selling Disruptive Products to...
The Global Code: How a New Culture of Universal Values Is Reshaping Business and Marketing
Customers who bought this item also bought

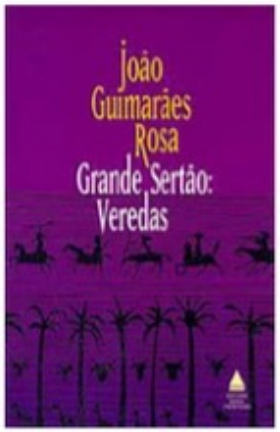

NETFLIX
Watch Instantly
Just for Kids
Browse DVDs
Your Queue
Suggestions For You
Genres
New Arrivals
Starz Play
Instantly to your TV
More like Thirtysomething
More like Good Luck Charlie
More like Touching the Void

Recommendations > Top Shelf Shelf

Here are some books we think you'll like based on the books you've added to this shelf. Other readers with similar interests have enjoyed them. How to improve your recommendations...

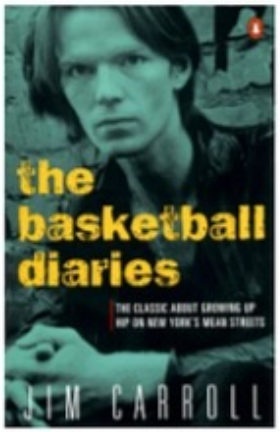
updated: Jul 21, 2017 08:36AM

View: covers | list



Want to Read

Not interested

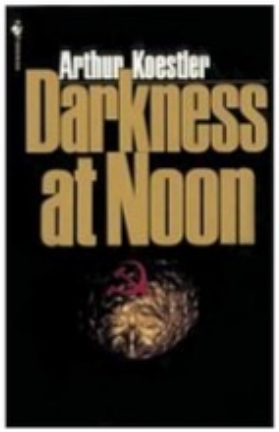


Want to Read

Not interested

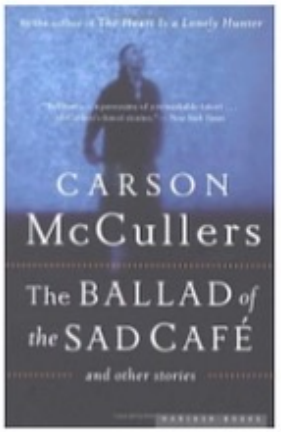
Want to Read

Not interested



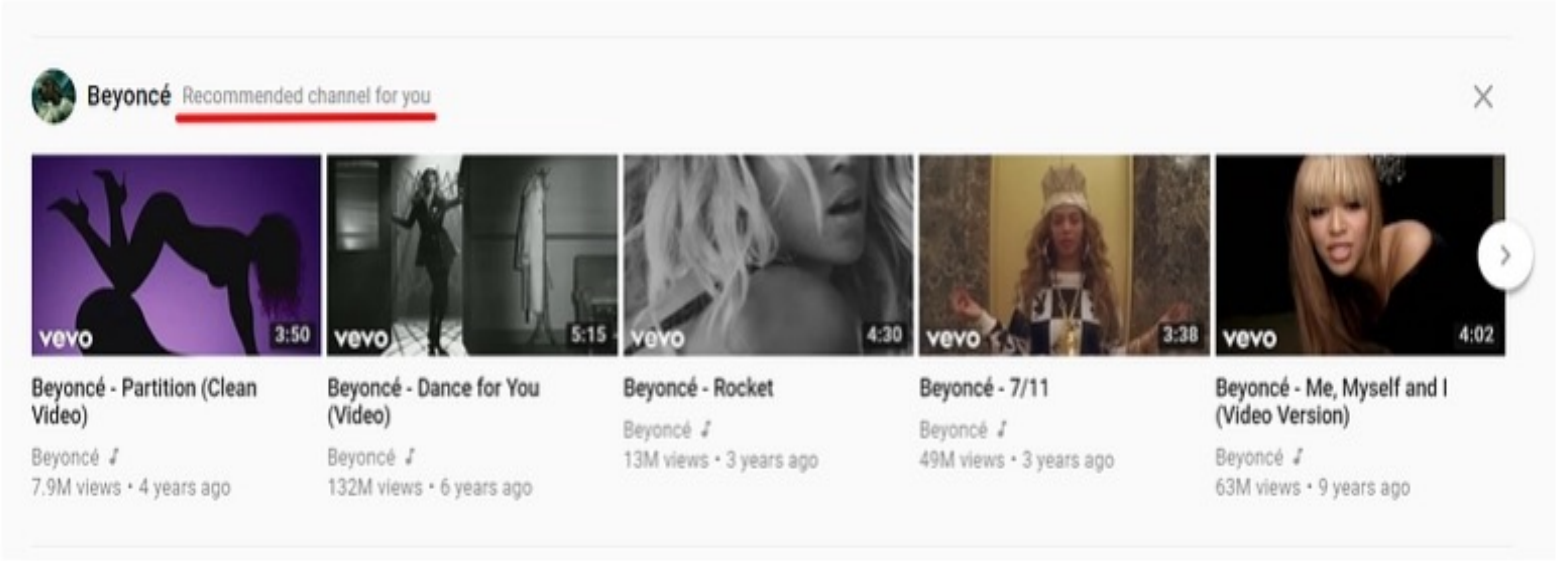
Want to Read

Not interested



Want to Read

Not interested

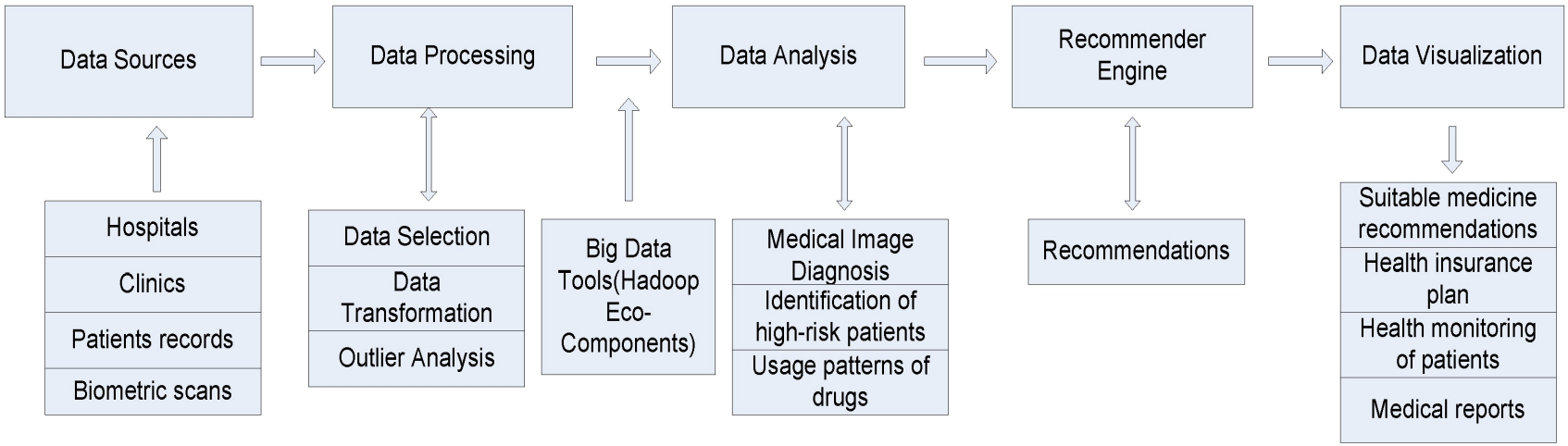


https://www.iteratorshq.com/blog/an-introduction-recommender-systems-9-easy-examples/

SOME EXAMPLES OF OTHER INDUSTRIES USING RECOMMENDATIONS SYSTEMS…

BTS | Barcelona Technology School

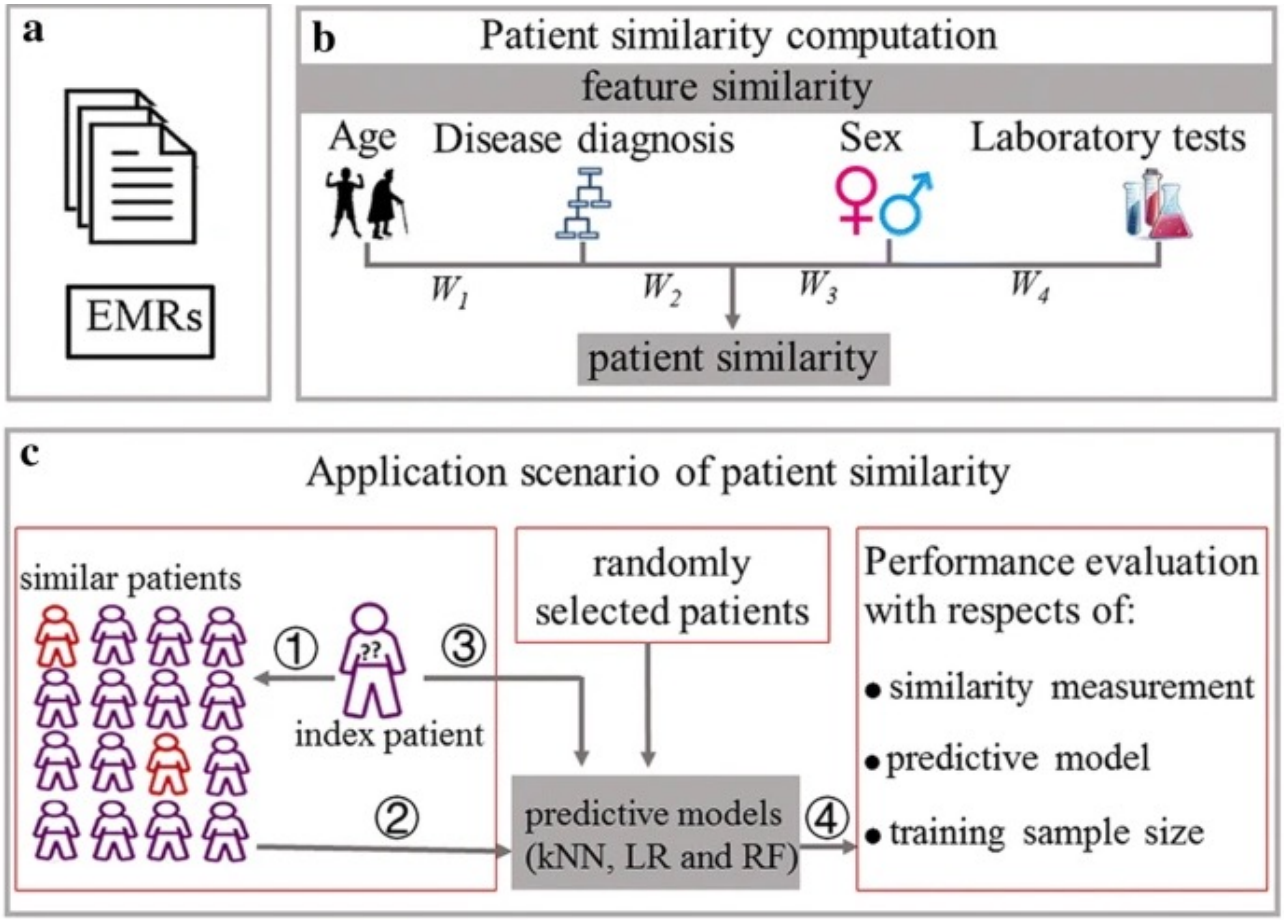
Recommendation Systems in HealthCare



https://www.mdpi.com/2079-3197/7/2/25/htm

BTS | Barcelona Technology School

Recommendation Systems in Disease Diagnosis



Main steps of the workflow. ***A***>Retrieving analysed data from EMRs data.

B>Calculation of four types of feature similarities and patient similarity.

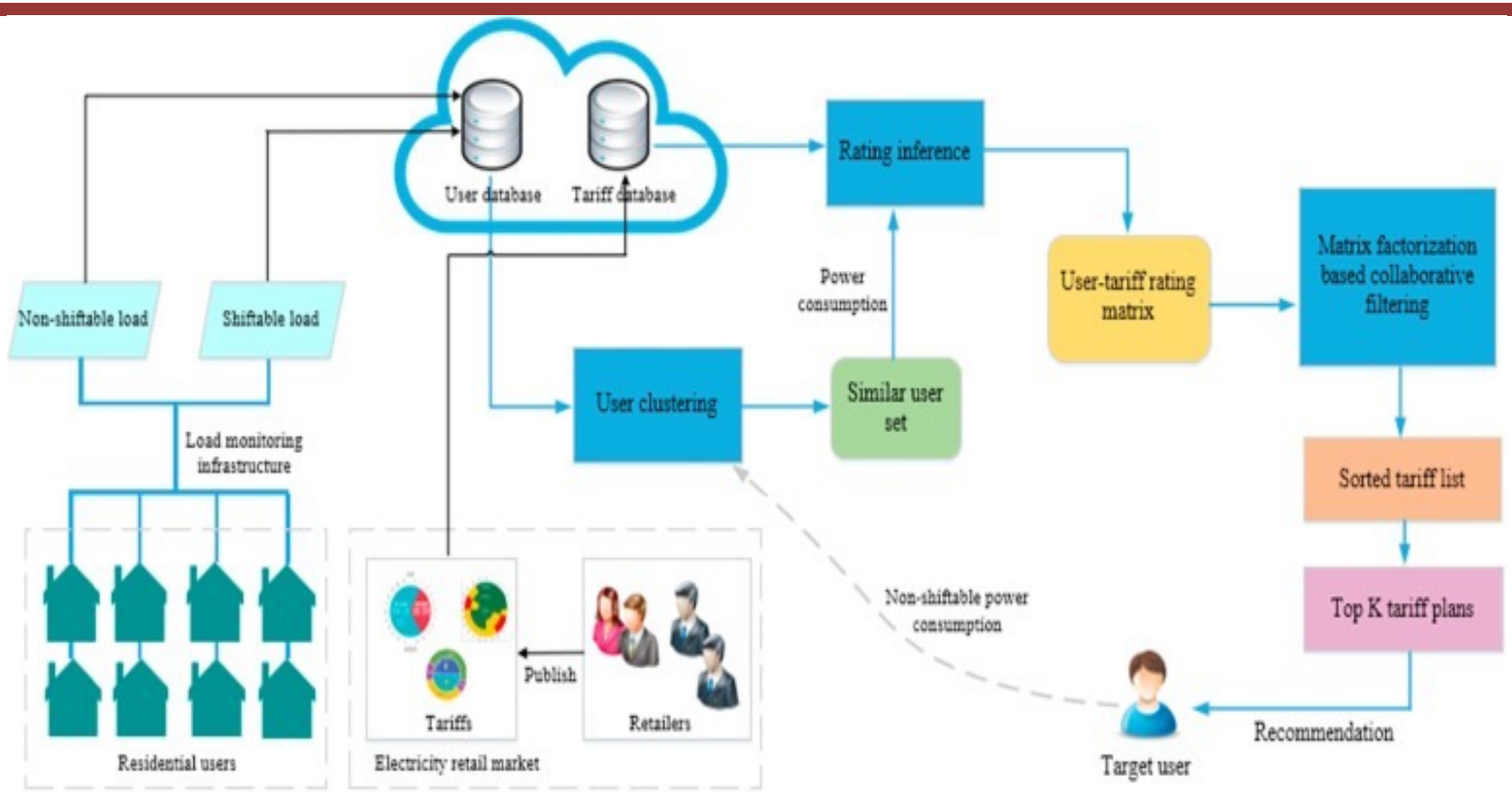
C>Application of patient similarity into personalized predictive model for future diabetes status prediction. *kNN* *k*-nearest neighbor, *LR* logistic regression, *RF* random forest, *EMRs* electrical medical records

The successful application of patient similarity in predicting a patient's diabetes status provided useful references for diagnostic decision-making support by investigating the evidence on similar patients.

https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-019-0718-2

BTS | Barcelona Technology School

Recommendation Systems in Energy Industry



https://www.sciencedirect.com/science/article/pii/S0142061518340924?casa_token=enieceCKHBcAAAAA:GVJtfuUbqYGXDf3UlaWDBA6qsNyGGfFIrD1mEBQrN3VyfG_5yVHgWRZ4gGTeo-GvU_gguW67jA

BTS | Barcelona Technology School