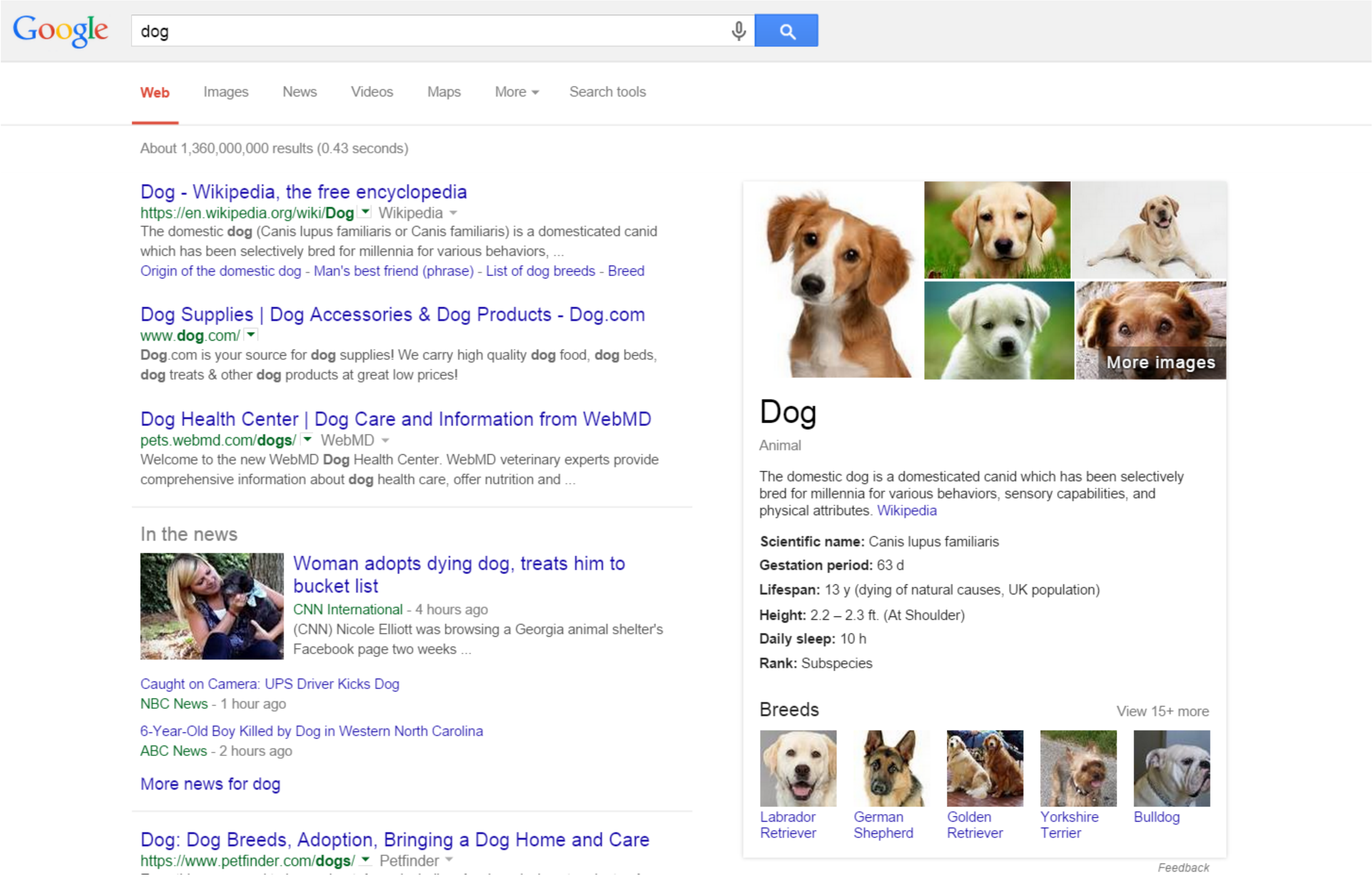
Learning to Rank

Jasmine Wilkerson

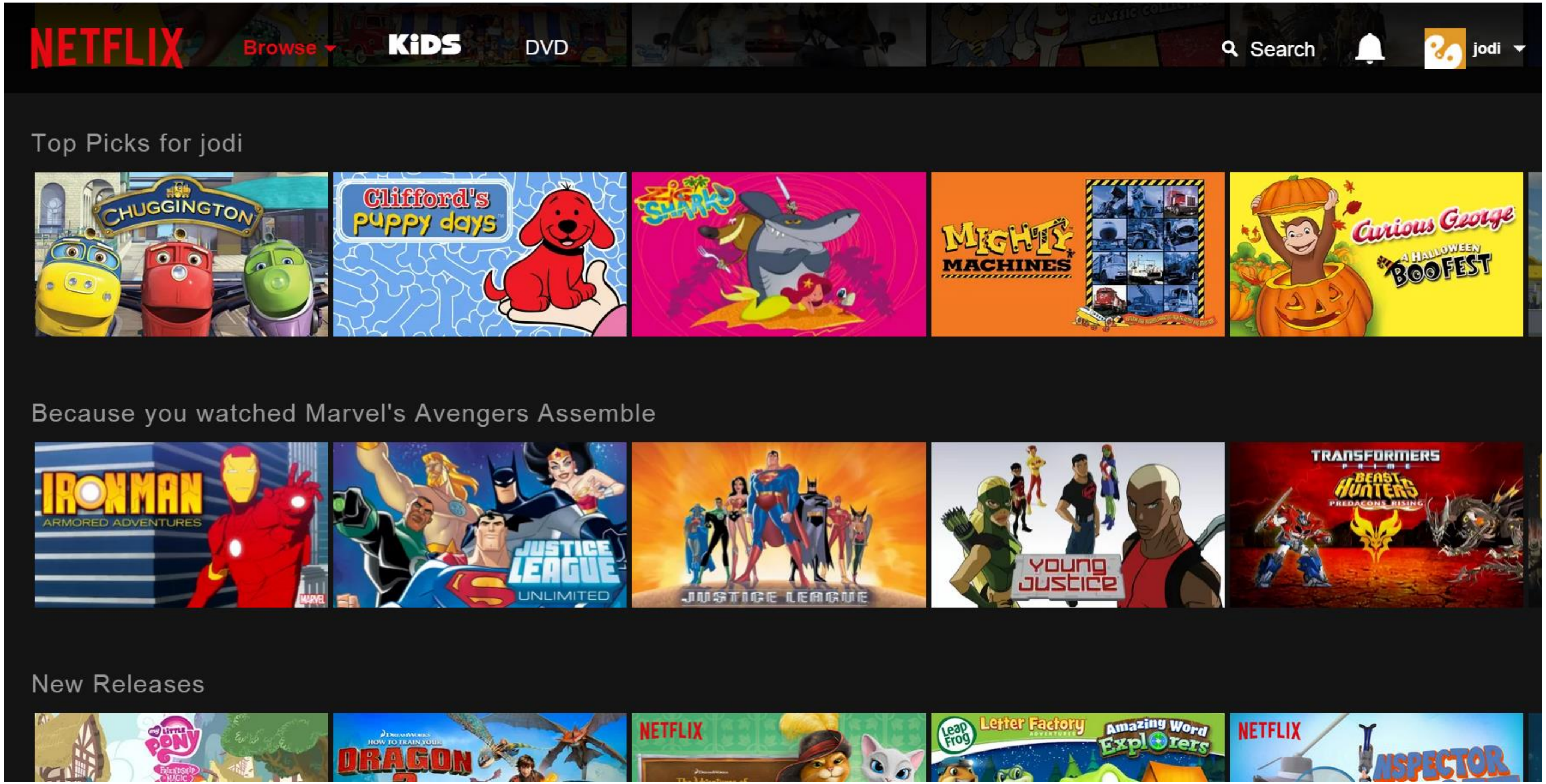
Overview

- Introduction to Ranking Problem
  - Using Information Retrieval (IR)-Document retrieval as example.
- General setup of a Ranking Model
- Quick recap of loss function and SVM
- Approaches
  - Pointwise
  - Pairwise
  - Listwise
- Further reading recommendation

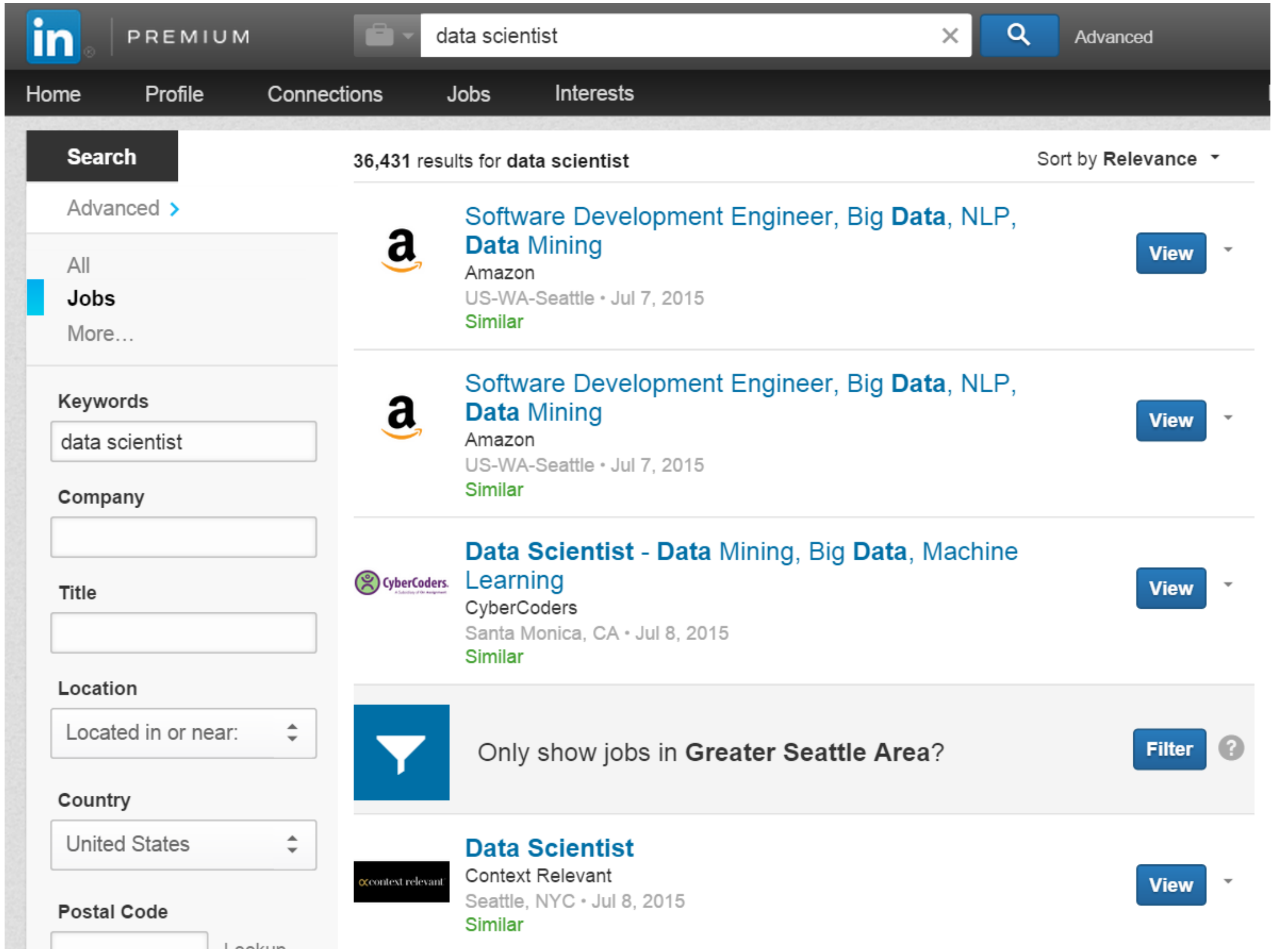
Example


Google
dog
Web
Images
News
Videos
Maps
More
Search tools
About 1,360,000,000 results (0.43 seconds)
Dog - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Dog Wikipedia
The domestic dog (Canis lupus familiaris or Canis familiaris) is a domesticated canid which has been selectively bred for millennia for various behaviors, ...
Origin of the domestic dog - Man's best friend (phrase) - List of dog breeds - Breed
Dog Supplies | Dog Accessories & Dog Products - Dog.com
www.dog.com/
Dog.com is your source for dog supplies! We carry high quality dog food, dog beds, dog treats & other dog products at great low prices!
Dog Health Center | Dog Care and Information from WebMD
pets.webmd.com/dogs/ WebMD
Welcome to the new WebMD Dog Health Center. WebMD veterinary experts provide comprehensive information about dog health care, offer nutrition and ...
In the news
Woman adopts dying dog, treats him to bucket list
CNN International - 4 hours ago
(CNN) Nicole Elliott was browsing a Georgia animal shelter's Facebook page two weeks ...
Caught on Camera: UPS Driver Kicks Dog
NBC News - 1 hour ago
6-Year-Old Boy Killed by Dog in Western North Carolina
ABC News - 2 hours ago
More news for dog
Dog: Dog Breeds, Adoption, Bringing a Dog Home and Care
https://www.petfinder.com/dogs/ Petfinder
More images
Dog
Animal
The domestic dog is a domesticated canid which has been selectively bred for millennia for various behaviors, sensory capabilities, and physical attributes. Wikipedia
Scientific name: Canis lupus familiaris
Gestation period: 63 d
Lifespan: 13 y (dying of natural causes, UK population)
Height: 2.2 – 2.3 ft. (At Shoulder)
Daily sleep: 10 h
Rank: Subspecies
Breeds
View 15+ more
Labrador Retriever
German Shepherd
Golden Retriever
Yorkshire Terrier
Bulldog
Feedback

Example


NETFLIX
Browse
KIDS
DVD
Search
jodi
Top Picks for jodi
CHUGGINGTON
Clifford's puppy days
MIGHTY MACHINES
Curious George
A HALLOWEEN BOOFEST
Because you watched Marvel's Avengers Assemble
IRON MAN
ARMORED ADVENTURES
JUSTICE LEAGUE
UNLIMITED
JUSTICE LEAGUE
young justice
TRANSFORMERS
BEAST HUNTERS
PREDACONS RISING
New Releases
NETFLIX
Letter Factory
Amazing Word Explorers
NETFLIX

Example


PREMIUM
data scientist
Advanced
Home
Profile
Connections
Jobs
Interests
Search
Advanced
All
Jobs
More…
Keywords
data scientist
Company
Title
Location
Located in or near:
Country
United States
Postal Code
36,431 results for data scientist
Sort by Relevance
Software Development Engineer, Big Data, NLP, Data Mining
Amazon
US-WA-Seattle • Jul 7, 2015
Similar
View
Software Development Engineer, Big Data, NLP, Data Mining
Amazon
US-WA-Seattle • Jul 7, 2015
Similar
View
Data Scientist - Data Mining, Big Data, Machine Learning
CyberCoders
Santa Monica, CA • Jul 8, 2015
Similar
View
Only show jobs in Greater Seattle Area?
Filter
Data Scientist
Context Relevant
Seattle, NYC • Jul 8, 2015
Similar
View

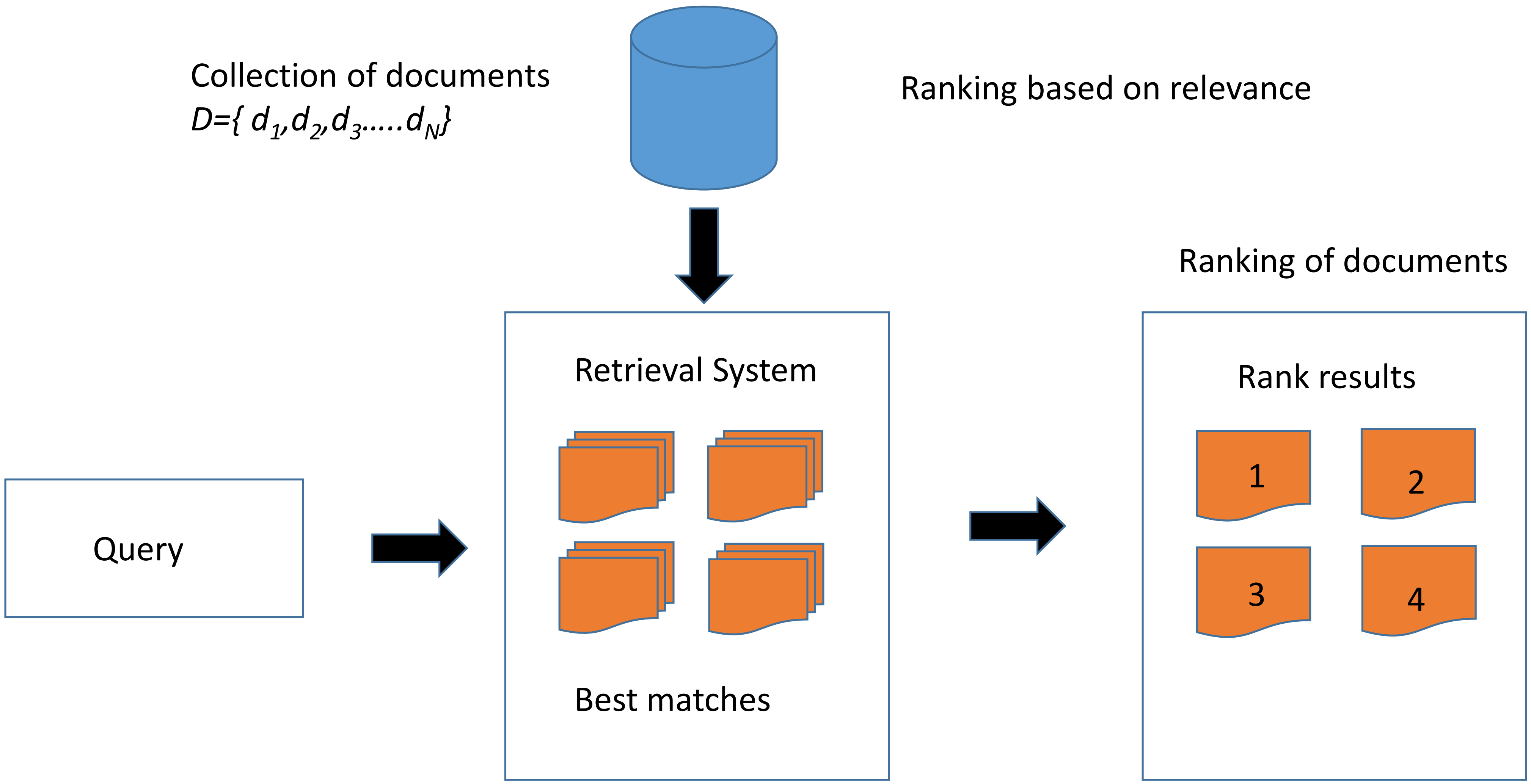
Introduction

- Learning to rank is a machine learning techniques for training model in ranking task
- Learning to rank can be employed to variety to application such as Information Retrieval (IR), Natural Language Processing (NLP), Data Mining (DM)
- Eg. Document retrieval which will be used as example through-out the presentation

Task of Retrieval and Ranking

- Given a query, system retrieves documents containing query words from document collection, ranks the documents and returns top ranked documents

Document Retrieval



Traditional Ranking Model

- Traditional ranking model $f(q,d)$ is created without training

- Eg. BM25
  - Conditional probability of relevancy of query to documents

    $P(r|q,d)$ r = 1 when relevant ; r= 0 when irrelevant

- Eg. Language Model for IR (LMIR)
  - conditional probability distribution by calculated the words appearing in the query and document

    $P(q,d)$

New Ranking Model Approach

- New trend, particularly in web search is to employ machine learning techniques to automatically construct the ranking model $f(q,d)$
- Motivated by existence of many signals to represent relevance in web search
  - Anchor text, PageRank score of a webpage
- These features are incorporated into the ranking model.

Other Features

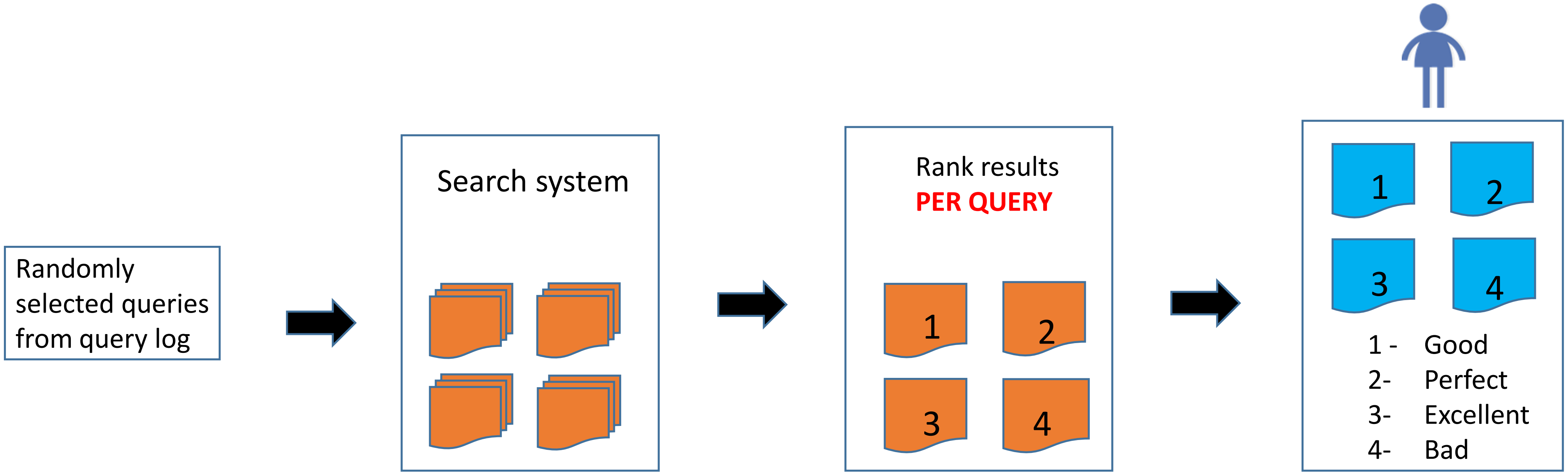
- Modern systems – especially on the Web – use a great number of features:
  - Log frequency of query word in anchor text?
  - Query word in color on page?
  - # of images on page?
  - # of (out) links on page?
  - PageRank of page?
  - URL length?
  - URL contains "~"?
  - Page length?
- The *New York Times* (2008-06-03) quoted Amit Singhal as saying Google was using over 200 such features.

Learning to Rank Model

- Supervised learning
- Labeled Dataset
  - Split into training and testing sets
- Algorithm
- Evaluate

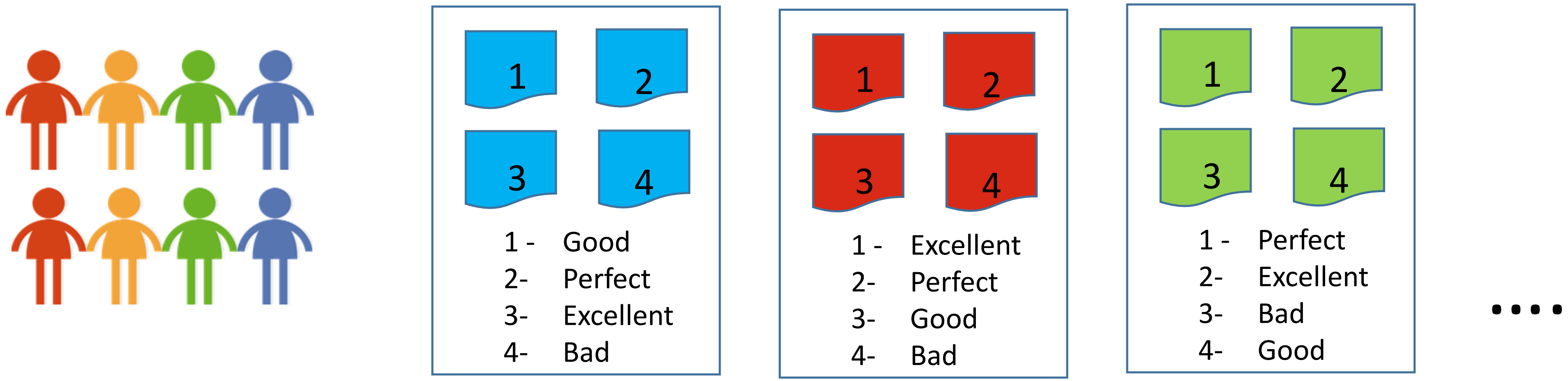
Obtain Labeled Data

- Explicit feedback by human judgements



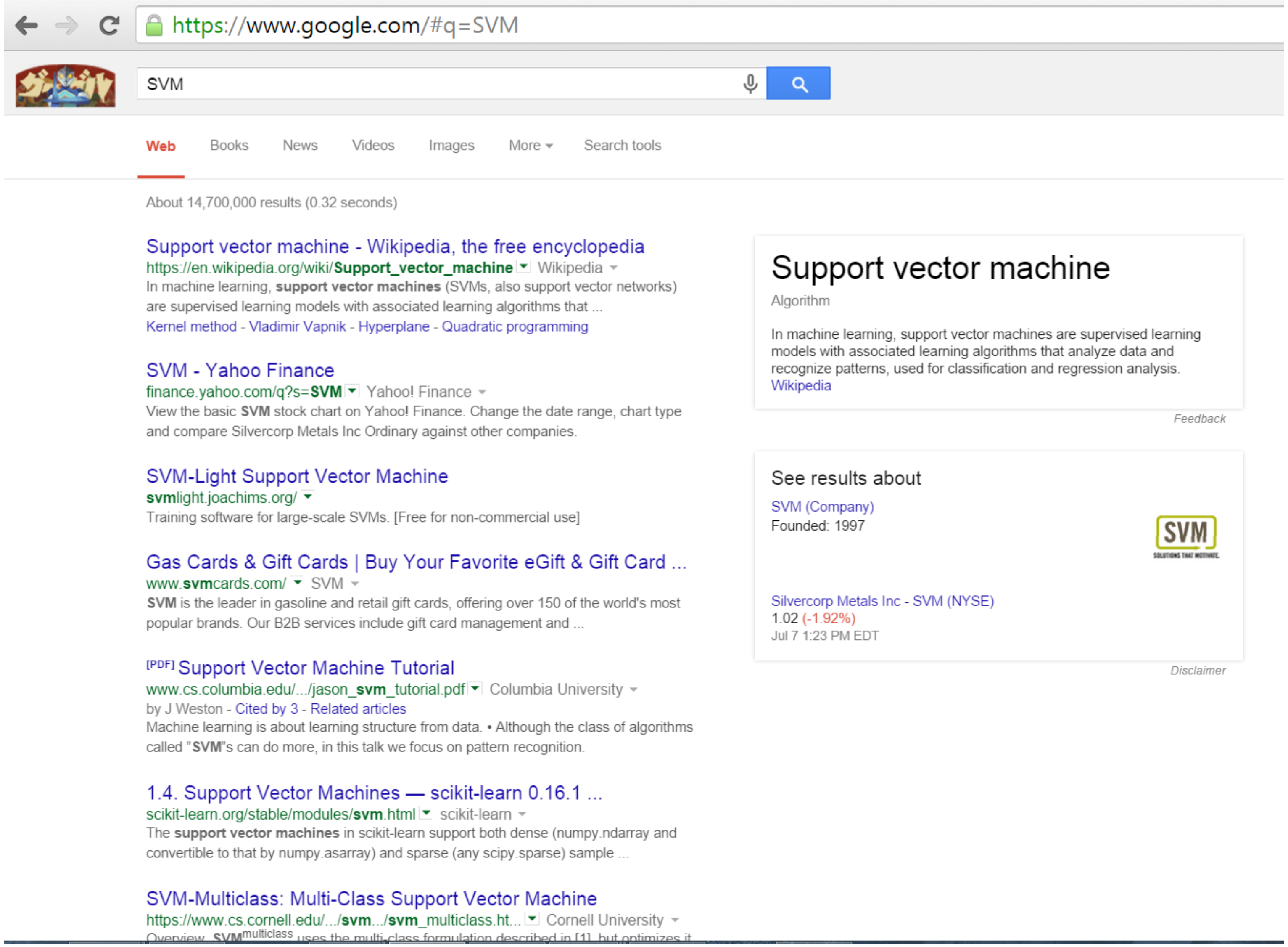
Obtain Labeled Data

PER QUERY



- Labels representing relevance are assigned to the query documents pairs
- Conduct majority voting with multiple judges.
- Expensive
- Overhead for users

How would you judge the rank?



Obtain Labeled Data

- Implicit feedback derived from log search data
- Eg. Click-Through-Rate, mousing, scrolling, eye-tracking
- No need to hire human judge
- No overhead for users
- More difficult to interpret

Training and Testing

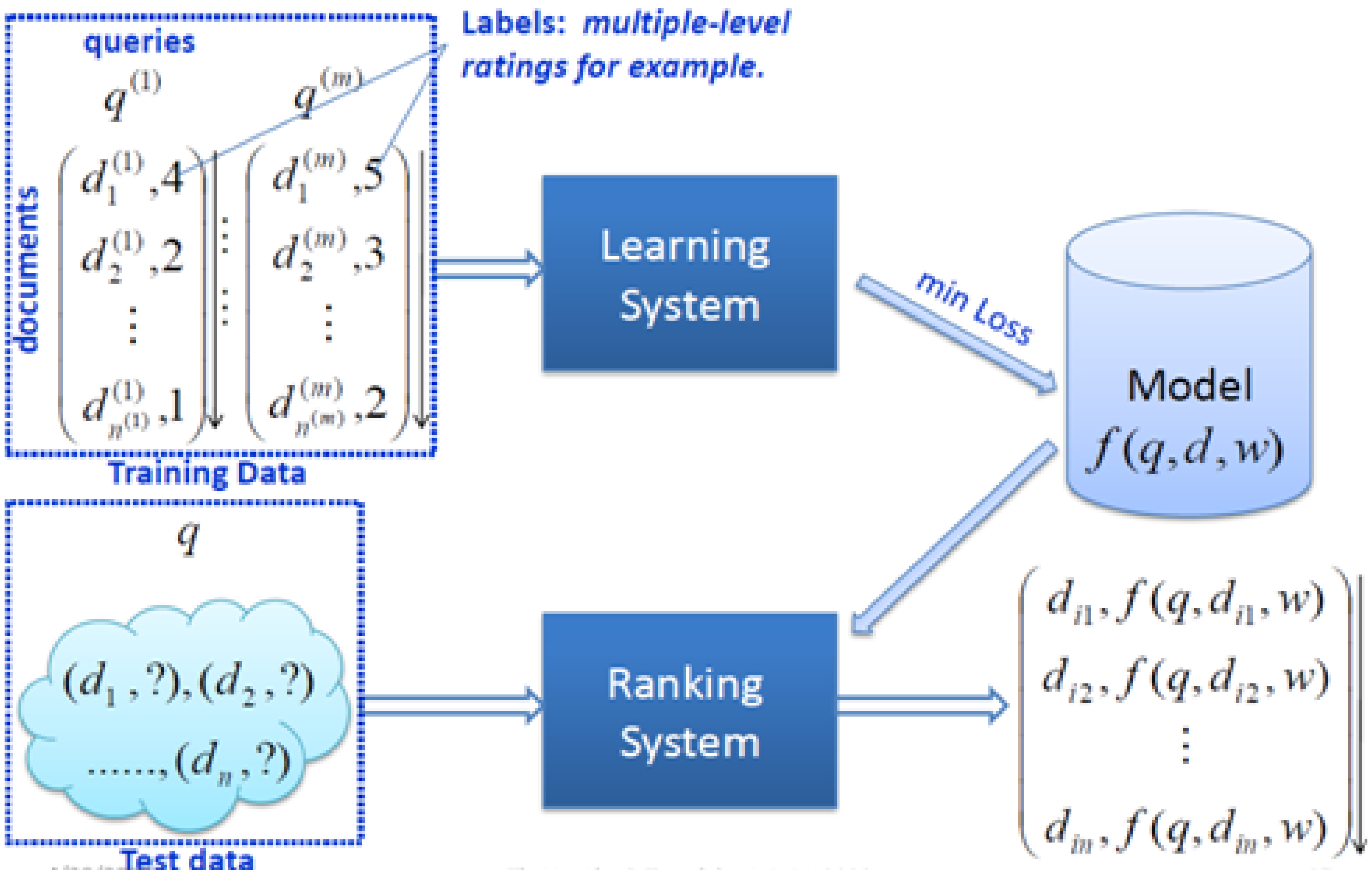


Image is from http://fangdonghao.com/

Evaluation

- Discounted Cumulative Gain (DCG)
- Normalized Discounted Cumulative Gain (NDCG)

Discounted Cumulative Gain (DCG)

- Uses graded relevance as a measure usefulness from a document
- Gain is accumulated starting at the top of the ranking, or discounted, at lower ranks
- DCG is the total gain accumulated at a particular rank p:

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

DCG Example

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$$

- 10 ranked documents judged on 0-3 relevance scale:

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- Discounted Gain (DC):
  - 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- Discounted Cumulative Gain:
  - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

IDCG and nDCG

- Idealized discounted cumulative gain (IDCG)
- The ideal ranking returns document with the highest relevance level, follow by next highest relevance level, etc.
- Normalized Discounted Cumulative Gain (nDCG)

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

Example of nDCG

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

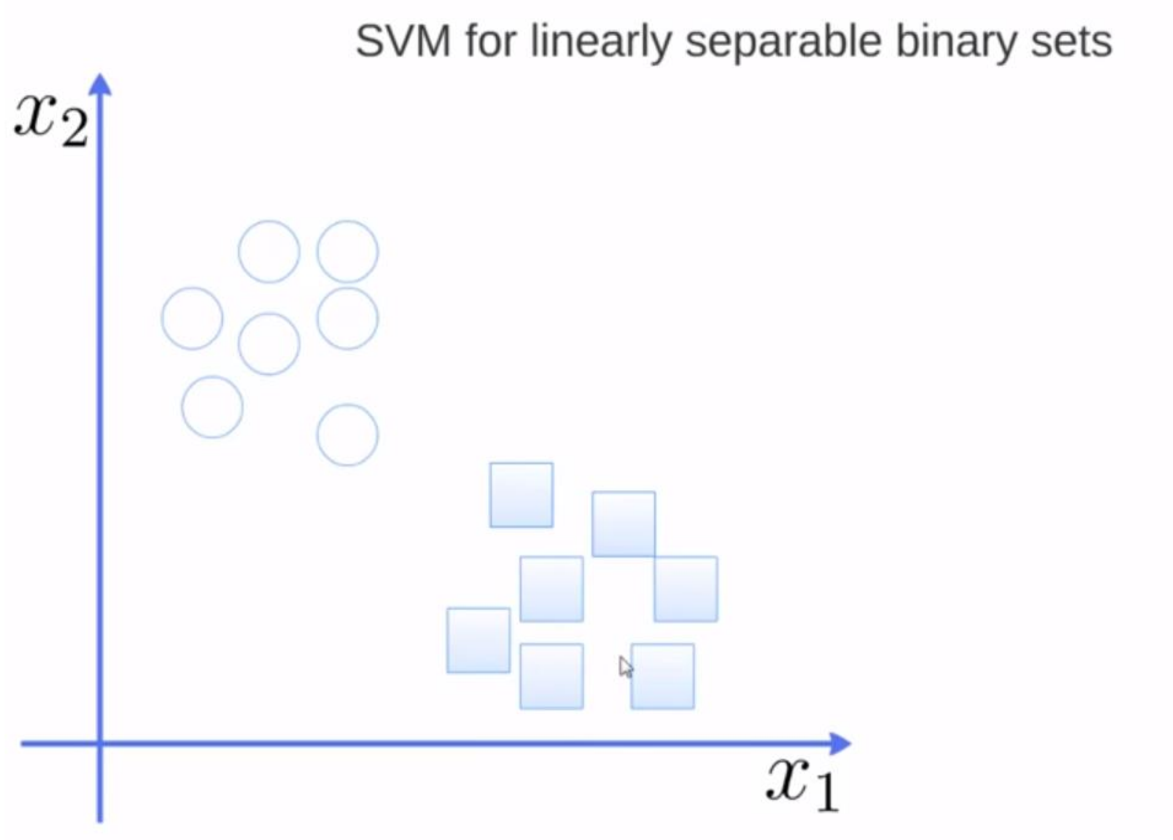
$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

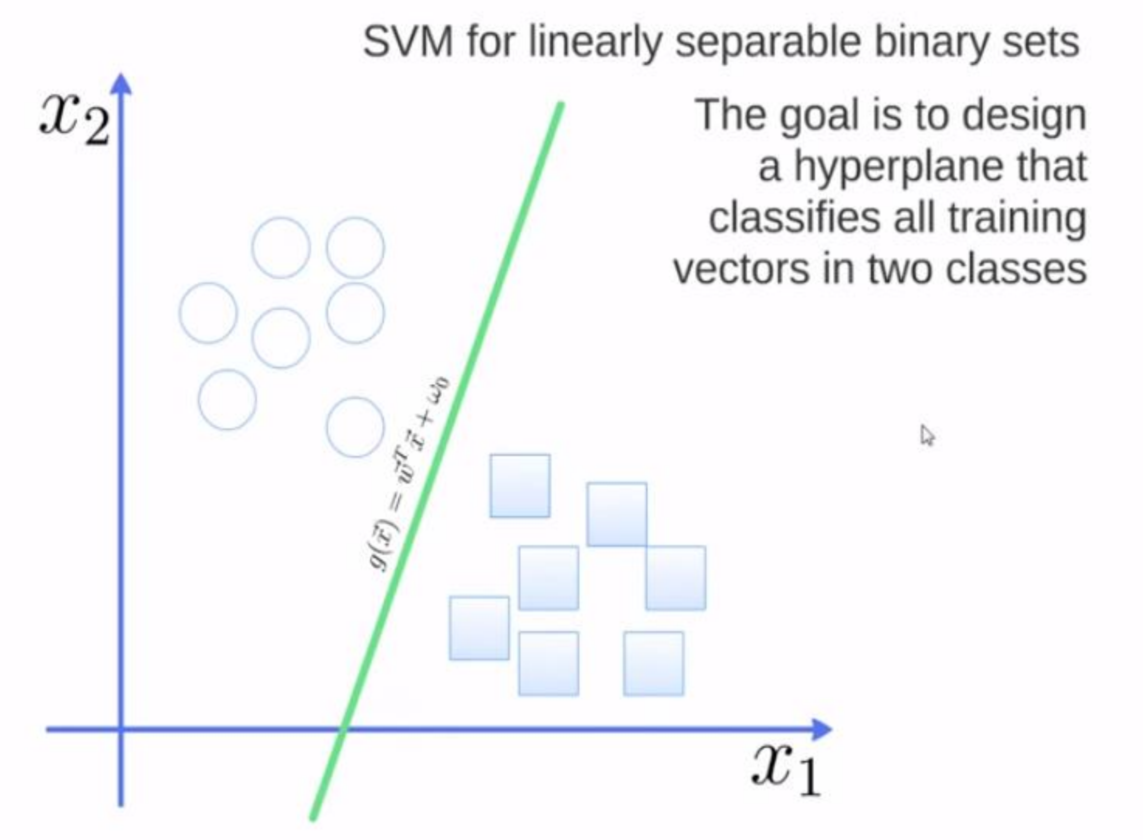
$$MaxDCG = DCG_{GT} = 4.6309$$

Recap

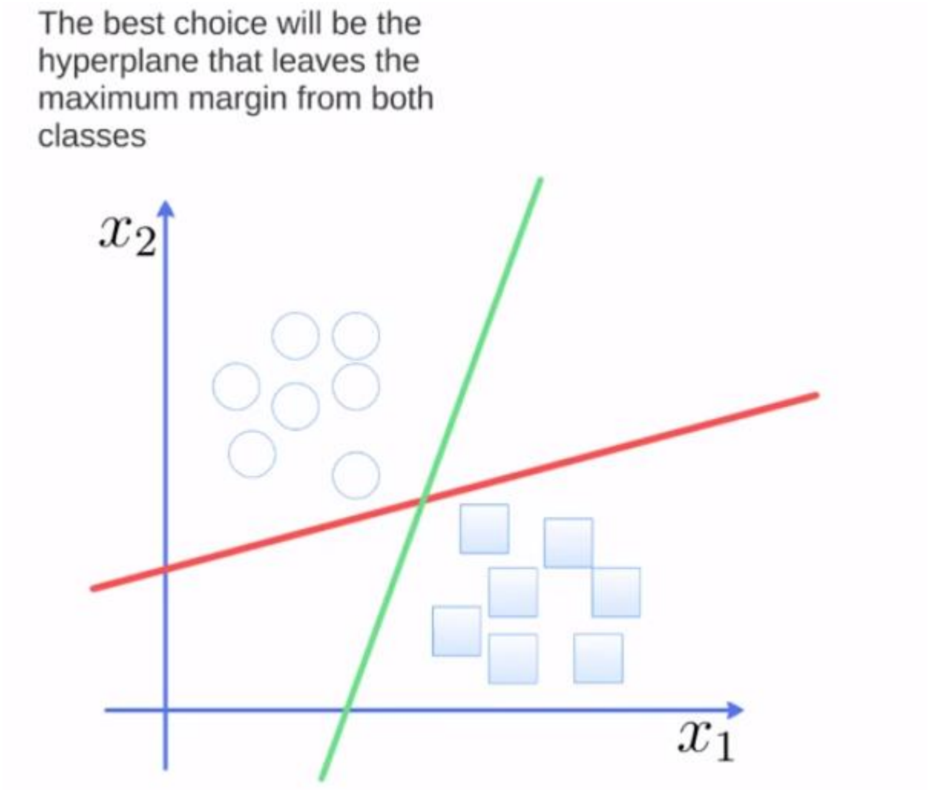
- Relationship of Hypothesis Function and Loss Function
- Recap of SVM

SVM recap

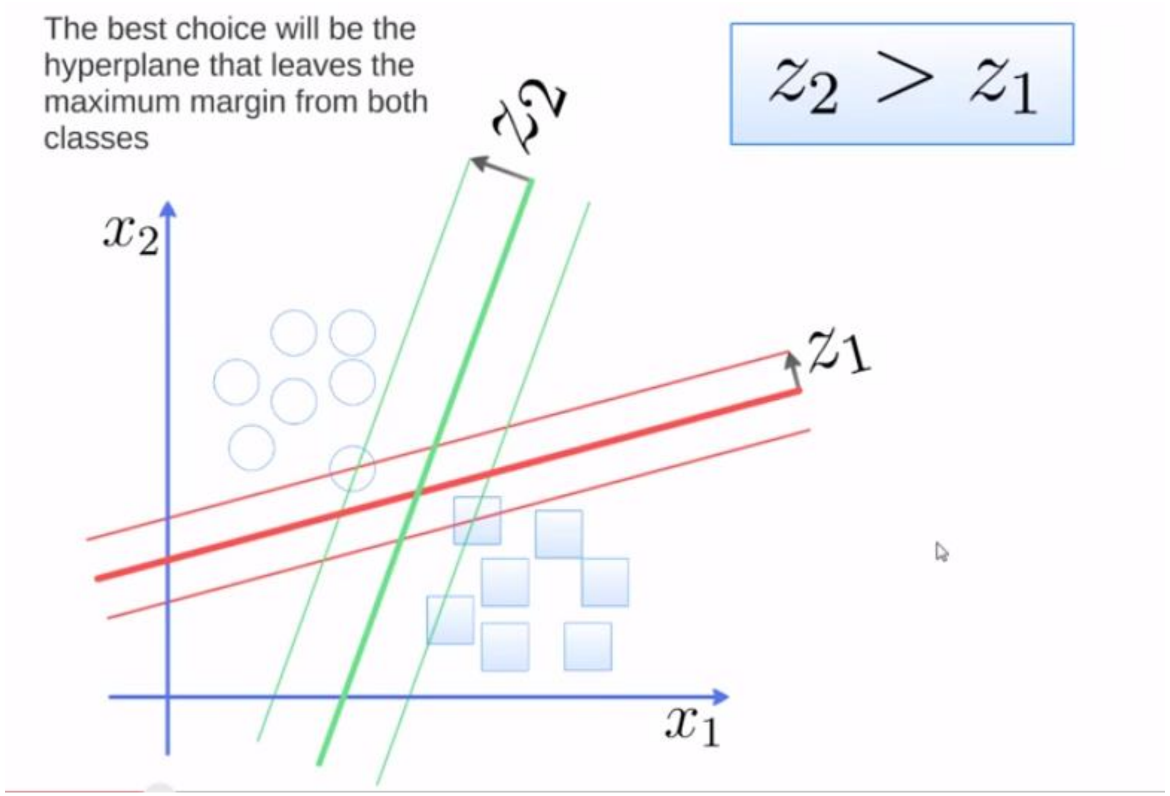

SVM for linearly separable binary sets
x_2
x_1


SVM for linearly separable binary sets
The goal is to design a hyperplane that classifies all training vectors in two classes
$g(\vec{x}) = \vec{w}^T\vec{x} + \omega_0$
x_2
x_1

SVM recap



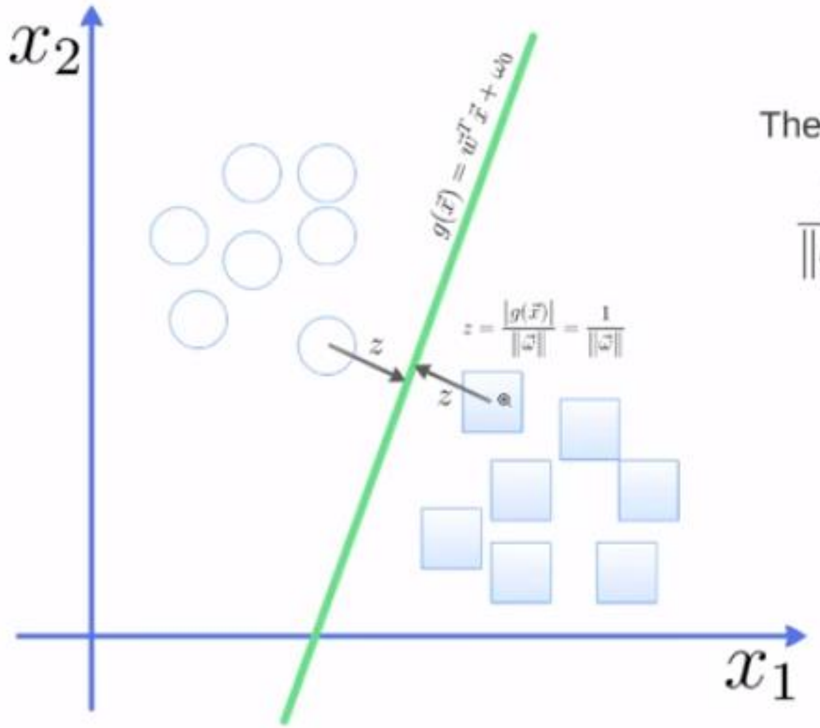
The best choice will be the hyperplane that leaves the maximum margin from both classes
x_2
x_1



The best choice will be the hyperplane that leaves the maximum margin from both classes
$z_2 > z_1$
z_2
z_1
x_2
x_1

SVM recap

$$g(\vec{x}) \geqslant 1, \quad \forall \vec{x} \in \text{class } 1$$
$$g(\vec{x}) \leqslant -1, \quad \forall \vec{x} \in \text{class } 2$$



The total margin is computed by

$$\frac{1}{\|\vec{\omega}\|} + \frac{1}{\|\vec{\omega}\|} = \frac{2}{\|\vec{\omega}\|}$$

Minimizing this term will maximize the separability

Learning to Rank Approach

- Pointwise
- Pairwise
- Listwise

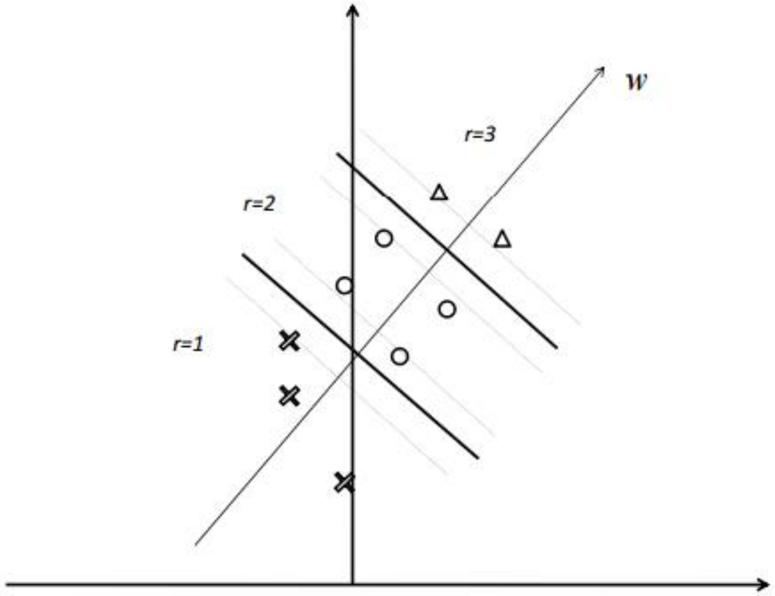
Pointwise Approach

- Transforming ranking to regression, classification, or ordinal regression.
- Technique includes Subset Ranking, McRank, Prank, OC SVM
- Supposed there is a group of objects (documents associated with a query) in the feature space
- Suppose that there are three grades (levels)
- Example: x_1, x_2 and x_3 in the first group are at different grades.
- The weight vector w corresponding to the linear function

  $f(x) = \langle w,x \rangle$, which can score and rank the objects.

Ordinal Classification SVM

- Learns the parallel hyperplanes by the large margin principle.



Loss Function

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{r=1}^{l-1}\sum_{i=1}^{m_r}(\xi_{r,i} + \xi^*_{r+1,i})$$
$$\text{s. t. } \langle w, x_{r,i}\rangle + b_r \geq 1 - \xi_{r,i}$$
$$\langle w, x_{r+1,i}\rangle + b_r \leq 1 - \xi^*_{r+1,i}$$
$$\xi_{r,i} \geq 0, \quad \xi^*_{r+1,i} \geq 0$$
$$i = 1, \cdots, m_r, \quad r = 1, \cdots, l-1$$
$$m = m_1 + \cdots + m_l,$$

Pairwise Approach

- Transforms ranking to pairwise classification problem
- Techniques are Ranking SVM, RankBoost, RankNet, IR SVM etc

Pairwise Approach

- Idea: Learn a ranking function, so that number of violated pair-wise training preferences is minimized.
- What is pairwise?

  Given a query q → Document d_i is more relevant than document dj

  Thus, $(q, d_i) > (q, d_j)$
- Form of Ranking Function: sort by

  $U(q, d_i) = w_1$* (#of query words in title of d_i) +

  w_2* (#of query words in anchor) +

  … +

  w_n* (page-rank of d_i) = $w*\phi(q, d_i)$

if user prefers d_i to d_j for query q, then $U(q, d_i) > U(q, d_j)$

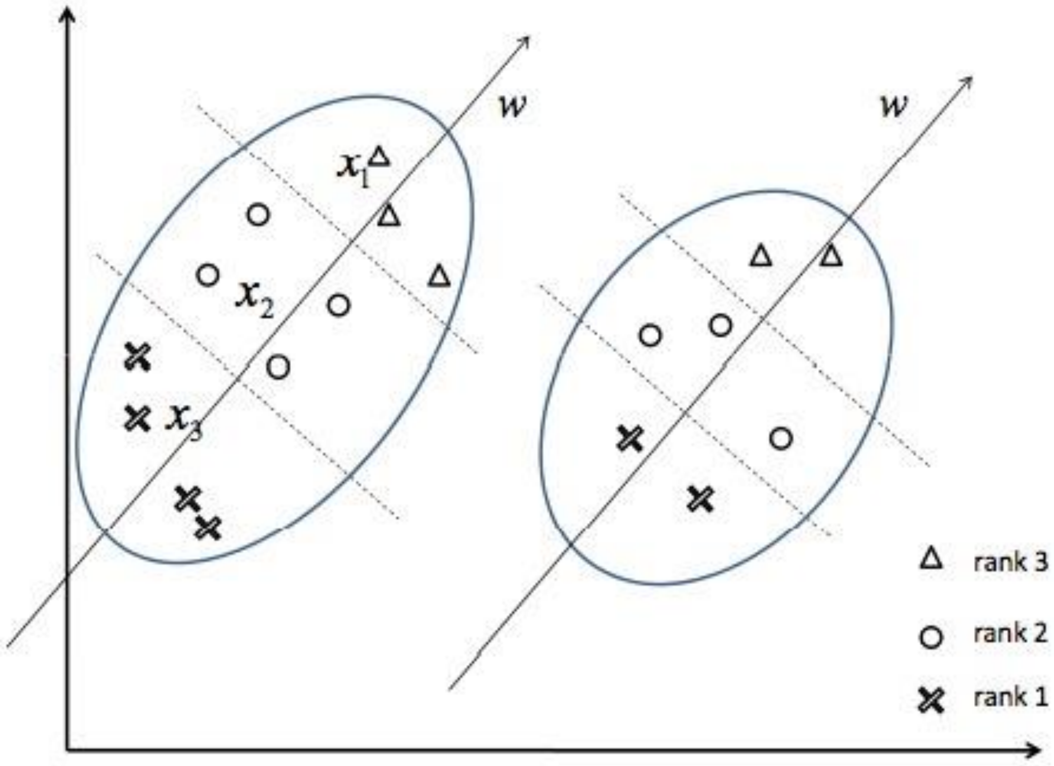
Pairwise Approach Example

- Supposed there are groups of objects (documents associated with two queries) in the feature space
- Suppose that there are three grades (levels)
- Example: x_1, x_2 and x_3 in the first group are at different grades.
- The weight vector w corresponding to the linear function

  $f(x) = \langle w,x \rangle$, which can score and rank the objects.

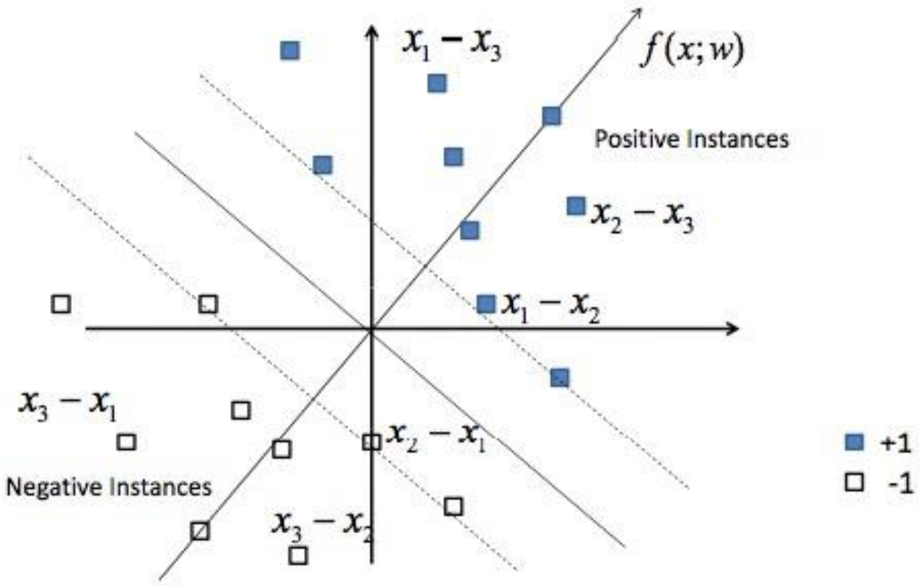
Pairwise Approach

- For a given query q, document $d_1 > d_2 > d_3$
- related, x_1, x_2, x_3 are d_1, d_2, d_3 characteristics
- Using machine learning methods to sort then transformed into a classification problem.



Pairwise Approach- Ranking SVM

- Define a new training sample such that
  - $x_1 - x_2$, $x_1 - x_3$, $x_2 - x_3$ is a positive sample,
  - $x_2 - x_1$, $x_3 - x_1$, $x_3 - x_2$ is negative sample
- Training a two classifier with support vector machine to these new training samples.



Ranking SVM

- using SVM to classify by minimizing loss function

$$
\begin{gathered}
\min_{w,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i \\
\text{s. t. } y_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i \\
\xi_i \geq 0 \\
i = 1, \ldots, m,
\end{gathered}
$$

where w is the parameter vector, x is a feature of the document, y is the document to the relative correlation, ξ is slack variable.

Future Reading

Large Scale Learning to Rank

D. Sculley
Google, Inc.
dsculley@google.com

Abstract

Pairwise learning to rank methods such as RankSVM give good performance, but suffer from the computational burden of optimizing an objective defined over $O(n^2)$ possible pairs for data sets with n examples. In this paper, we remove this super-linear dependence on training set size by sampling pairs from an implicit pairwise expansion and applying efficient stochastic gradient descent learners for approximate SVMs. Results show orders-of-magnitude reduction in training time with no observable loss in ranking performance. Source code is freely available at: http://code.google.com/p/sofia-ml

Combined Regression and Ranking

D. Sculley
Google, Inc.
Pittsburgh, PA USA
dsculley@google.com

;STRACT

ıy real-world data mining tasks require the achievement wo distinct goals when applied to unseen data: first, nduce an accurate preference *ranking*, and second to : good *regression* performance. In this paper, we give ɛfficient and effective Combined Regression and Ranking hod (CRR) that optimizes regression and ranking ob-ives simultaneously. We demonstrate the effectiveness of R for both families of metrics on a range of large-scale ːs, including click prediction for online advertisements. ults show that CRR often achieves performance equiv-ıt to the best of both ranking-only and regression-only roaches. In the case of rare events or skewed distribu-s. we also find that this combination can actuallv im-

for producing predicted values with the same pairwise ordering $y'_1 > y'_2$ as the true values $y_1 > y_2$ for a pair of given examples.

In many settings good performance on both families together is needed. An important example of such a setting is the prediction of clicks for sponsored search advertising. In real-time auctions for online advertisement placement, ads are ranked based on $bid * pCTR$ where $pCTR$ is the predicted click-through rate (CTR) an ad. Predicting a good ranking is critical to efficient placement of ads. However, it is also important that the $pCTR$ not only give good ranking value, but also give good regression estimates. This is because online advertisements are priced using next-price auctions, in which the price for a click on an ad at rank i

References

- A Short Introduction to Learning to Rank IEICE TRANS. INF. & SYST., VOL.E94–D, NO.10 OCTOBER 2011
- https://www.youtube.com/watch?v=1NxnPkZM9bc lecture on SVM by **Thales Sehn Körting**
- http://fangdonghao.com/
- http://users.cis.fiu.edu/~lzhen001/activities/KDD_USB_key_2010/docs/p979.pdf
- http://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/35662.pdf