Quick intro to min-hash

UCU Winter School 2017

James Pritts

Czech Technical University

January 20, 2017

Jaccard similarity

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- Jim an Nancy have watched 100 movies and saw 50 of the same movies
- if Jim's movies are A, and Nancy's movies are B, what is the $J(A, B)$?

Uses

- News aggregators
- Near duplicate detection
- Motion segmentation, multi-model fitting (avoid greedily choosing best model)
- Image Retrieval for near-duplicate detection

Images as sets

- Jaccard similarity is suitable to the bag-of-words model
- The representation is easily constructed by considering all non-zero visual words

MinHash signatures

- Problem, computing J(A,B) is quadratic
- Set intersection and union are expensive operations
- Min-Hash signatures are computed for fast comparison

Hashing to shuffle

- Let x be the largest element in the set visual word.
- $h(x) = (ax + b) \bmod c$
- Randomly choose a, b less the max value of x
- Choose a prime number c greater than max value of x.
- Every integer x mapped to unique integer.
- Equivalent to quickly shuffling numbers.

Example

- Set $A = \{32, \mathbf{3}, 22, 5, \mathbf{15}, \mathbf{11}\}$
- Set $B = \{\mathbf{15}, 30, 7, \mathbf{11}, 28, \mathbf{3}, 17\}$
- $J(A, B) = 0.3$

- What is the probability that the min-hash is the same for both sets?
- MinHash is equivalent to shuffling $A \cup B$ and taking the minimum.
- $J(A, B) = 0.3$

Min-hash signatures

- Generate a lot of hash functions
- Calculate min-hash for each signature
- Calculate the proportion of equal min-hash signatures
- Approx equal to Jaccard similarity.