SIMILARITY HASH

MIN-HASH = minwise independent permutations

Permutations X = set of items $|X|!$ permutations

$X = \{a, b, c\}$

a b c
a c b

b a c
b c a

c a b
c b a

3! permutations hv

Each permutation induces an order (eg. c a b → c < a < b)

We take the minimum of a permutation

$h(X)$ = permutation of X

$\min h(X) \overset{\text{def}}{=} \arg\min_{a \in X} h(a)$

$\min h(X) : X = U \to a \in X$

example:

	c a b
h()	1 2 3

$\min h(X) = c$

Given a set $X \subseteq U$, permutation $h : U \to [|U|]$ (ranking)

family $\mathcal{H} = \{h : X \to [|X|] \text{ bijective}\}$ is MIN-WISE INDEPENDENT

if $\forall X \subseteq U \quad \forall a \in X$

$$\Pr_{h \in \mathcal{H}}[a = \min h(X)] = \frac{1}{|X|}$$

we like this because of Jaccard's index for similarity

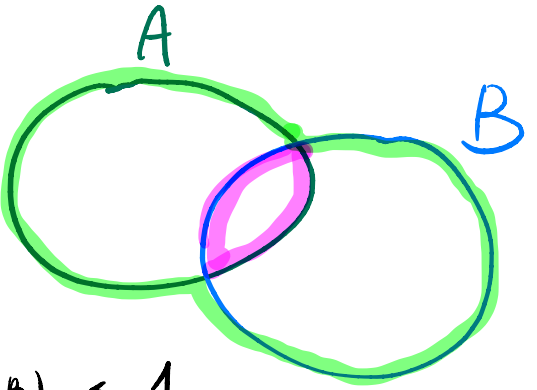
(see Broder & Altavista)

document = bag of words = set of "lemmatized" words

$U = \{$ all possible words $\}$

$A \subseteq U$ (dictionary)

$A, B \subseteq U$ Jaccard's index



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

$$0 \leq J(A,B) \leq 1$$

0: $A \cap B = \phi$

1: $A = B$

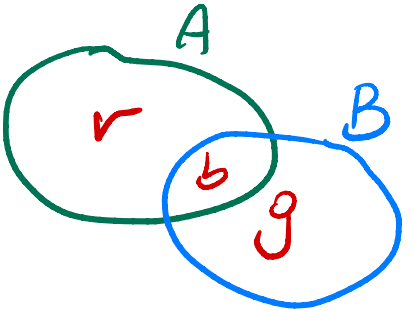
After sorting A, B: $J(A,B)$ takes $O(|A|+|B|)$ time

Big data: $J(A,B)$ is too expensive for zillions of A, B

MIN-HASH (SIM-HASH) helps!

Interesting property: $\Pr_{h \in \mathcal{H}}\left(\min h(A) = \min h(B)\right) = J(A,B) \quad \forall A,B$

TOO SEE WHY



$r = |A \setminus B|$
$b = |A \cap B|$
$g = |B / A|$

$|A| = r + b$
$|B| = b + g$
$|A \cap B| = b$
$|A \cup B| = r + b + g$

$$\Pr\left(\min h(A) = \min h(B)\right) = \Pr\left[\exists y \in A \cap B : y = \min h(X)\right]$$

$X = A \cup B$
$|X| = r + b + g$

$$= \underset{\text{UNION BOUND}}{\sum_{y \in A \cap B}} \underbrace{\Pr\left[y = \min h(X)\right]}_{\frac{1}{|X|}} = \frac{b}{r+b+g} = \frac{|A \cap B|}{|A \cup B|} = J(A,B)$$

MIN-HASH

$$\Pr\left[a = \min h(X)\right] = \frac{1}{|X|}$$

Some observations

- Storing $h(U)$ is unfeasible in many applications
- One single $h \in \mathcal{H}$ could not give enough precision

Fix

- It has been shown that we can "safely" replace permutation family $\mathcal{H}$ with our UNIVERSAL family $\mathcal{H}$
- Use k hash functions $h_1, h_2, \ldots, h_k$

k-min hash $\quad A \rightarrow \langle \min h_1(A), \min h_2(A), \ldots, \min h_k(A) \rangle = S(A)$ sketch

approximate Jaccard : $\tilde{J}(A,B) = \dfrac{|S(A) \cap S(B) \cap S(A \cup B)|}{|S(A \cup B)|}$

bottom-k hash:
instead of using k hash functions we use a single h, and take the k smallest ranked items

$S(A \cup B) = S(S(A) \cup S(B)) = \{a, b, c, d\}$

$S(A) = \{a, c, d, f\}$

$S(B) = \{a, b, f, g\}$

$k=4$

$a < b < c < d < e < f < g$

Given $S(A)$, $S(B)$, then $J(A,B)$ can be computed in $O(k)$ time

$\Rightarrow$ <u>step 1</u> compute $S(A)$ for each set A

<u>step 2</u> for each pair A, B : compute $\tilde{J}(A,B)$

much faster than computing $J(A,B)$

Price to pay: it's an approximation. Use indicator variables

k-min hash : given A, B

$$X_i = \begin{cases} 1 & \text{if } \min h_i(A) = \min h_i(B) \\ 0 & \text{otherwise} \end{cases} \qquad pr = J(A,B)$$

$E[X_i] = J(A,B)$ UNBIASED ESTIMATOR

$Y = \frac{\sum_{i=1}^{k} X_i}{k}$ UNBIASED ESTIMATOR $E[Y] = \frac{\sum_{i=1}^{k} E[X_i]}{k} = \frac{k\,J(A,B)}{k} = J(A,B)$

$Y = J(A,B)$

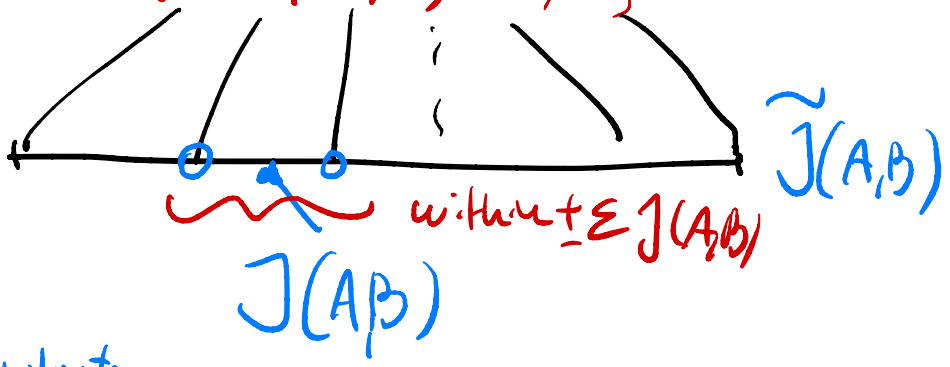
$K=1 \rightarrow Y \in \{0,1\}$

$K=2 \rightarrow Y = \{0, \frac{1}{2}, 1\}$

$K=3 \rightarrow Y \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$

$K=4 \rightarrow Y \in \{0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}$

$K \ldots \rightarrow Y \in \{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\}$



Chernoff's bounds helps us to evaluate

CHERNOFF'

$$Y' = \sum_{i=1}^{k} X_i \qquad (Y' = kY) \Rightarrow \mu' = E[Y'] = k\,\underbrace{E[Y]}_{\mu} = k\,J(A,B)$$

$$\Pr[|Y-\mu| \geq \varepsilon\mu] \overset{\text{multiply by } k}{=} \Pr[|Y'-\mu'| \geq \varepsilon\mu'] \underset{CB}{\leq} 2\cdot e^{\frac{-\varepsilon^2\mu'}{3}} = 2e^{\frac{-\varepsilon^2 k\mu}{3}}$$

relative error

Is this good! We want to bound this probability by $\delta < 1$

$$2e^{\frac{-\varepsilon^2 k\mu}{3}} < \delta \Rightarrow k = O\left(\varepsilon^{-2}\lg\delta^{-1}\mu^{-1}\right)$$

not good! $\mu = J(A,B)$

$\mu^{-1} \sim |A \cup B|$

Let's use a more suitable version of CBs (concentration bounds)

AZUMA-HOEFFDING bound (AH)

$X_1, X_2, \ldots, X_k$ i.i.d. random vars, $\mu = E\left[\frac{\sum_{i=1}^{k} X_i}{k}\right]$

$a_i \le X_i \le b_i$

(in our $X_i \in \{0,1\}$)

our Y

$$Pr[|Y-\mu| \ge \varepsilon] \le 2 \cdot e^{-\frac{2k^2\varepsilon^2}{\sum_{i=1}^{k}(b_i-a_i)^2}}$$

no μ here

ABSOLUTE ERROR ☹

IN OUR CASE:

$$2 \cdot e^{-\frac{2k^2\varepsilon^2}{k}} = 2e^{-2k\varepsilon^2} < \delta \Rightarrow k = O(\varepsilon^{-2} \log \delta^{-1})$$

$a_i = 0$ $b_i = 1$ $\Rightarrow \sum_{i=1}^{k}(b_i-a_i)^2 = k$

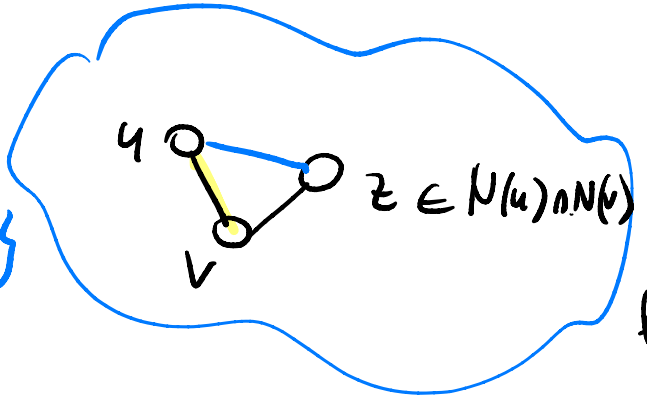
☺

NETWORK ANALYSIS : social networks "triangle" is a measure of sociality

$G = (V, E)$

$x \in V$

$N(x) = \{y : xy \in E\}$



Counting # triangles

BASIC ALGORITHM :

```
C = 0
for each edge uv ∈ E :
    C += |N(u) ∩ N(v)|
return C/3
```

$O(|E|^{3/})$

Let's estimate instead C = # triangles

we use Jaccard and min-hash

$A = N(u) \quad B = N(v)$ $\boxed{|A \cap B|}$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad \dots \rightarrow \boxed{|A \cap B|} = \frac{J(A,B) \cdot (|A| + |B|)}{1 + J(A,B)}$$

using min-hash, we get an estimation

$$\widetilde{|A \cap B|} = \frac{\tilde{J}(A,B)\,(|A| + |B|)}{1 + \tilde{J}(A,B)}$$

note: error is both at the numerator and denominator

TIME IS LINEAR $O(|E|)$