

Chapter 5, Exercise 5.22 (page 123)

There are $|X \cap Y|$ similar elements and $n - |X \cap Y|$ different elements in the sets X and Y . Let $BF_k^m(x)$ be the Bloom filter of set x with k hash functions and m bits. By definition of the Bloom filter,

$$BF(X) = BF(X - X \cap Y) \text{ OR } BF(X \cap Y) \text{ (similarly } BF(Y)),$$

where OR is bitwise oring.

Some of the bits where $BF(X - X \cap Y)$ and $BF(Y - X \cap Y)$ differ are masked if they are 1's in $BF(X \cap Y)$. Thus, the number of bits where $BF(X)$ and $BF(Y)$ differ is the number of different bits between $BF(X - X \cap Y)$ and $BF(Y - X \cap Y)$ minus the number of masked bits.

Let A be the number of different bits in $BF(X - X \cap Y)$ and $BF(Y - X \cap Y)$, B the number of masked bits, and C the number of different bits in $BF(X)$ and $BF(Y)$. Because $C = A - B$, by linearity of expectation,

$$\mathbf{E}[C] = \mathbf{E}[A] - \mathbf{E}[B].$$

We first compute $\mathbf{E}[A]$. A bit is different in $BF(X - X \cap Y)$ and $BF(Y - X \cap Y)$ if it is 0 in one Bloom filter and 1 in the other. Thus, $\Pr(\text{bit } i \text{ is different}) = 2 \cdot \Pr(\text{bit } i = 1) \cdot \Pr(\text{bit } i = 0)$. Define p as $\Pr(\text{bit } i = 0) = (1 - \frac{1}{m})^{k(n - |X \cap Y|)}$. $\Pr(\text{bit } i = 1) = 1 - p$. Thus, $\mathbf{E}[A] = m \cdot \Pr(\text{bit } i \text{ is different}) = m \cdot 2p(1 - p)$.

Next, we compute $\mathbf{E}[B]$. A bit is masked if it is different between $BF(X - X \cap Y)$ and $BF(Y - X \cap Y)$ and is 1 in $BF(X \cap Y)$. From previous analysis, we know that $\Pr(\text{bit } i \text{ is different}) = 2p(1 - p)$. Define p' as $\Pr(\text{bit } i \text{ is 1 in } BF(X \cap Y)) = 1 - (1 - \frac{1}{m})^{k|X \cap Y|}$. Thus, $\Pr(\text{bit } i \text{ is masked}) = 2p(1 - p) \cdot p'$. Therefore, $\mathbf{E}[B] = m \cdot 2p(1 - p) \cdot p'$.

Finally, the expected number of bits where Bloom filters differ is $\mathbf{E}[C] = \mathbf{E}[A] - \mathbf{E}[B] = m \cdot 2p(1 - p) - m \cdot 2p(1 - p) \cdot p' = m \cdot 2p(1 - p)(1 - p')$, where both p and p' are functions of n , k , and $|X \cap Y|$, as shown above.

The number of bits where Bloom filters of song lists differ is an estimator of music taste similarity. The running time of comparing Bloom filters is $O(m)$, which is less than $O(n \lg n)$, the running time of directly comparing song lists.