Signature files (Faloutsos' 92)

- Quick and dirty, based on hashing
- Underestimate the distances (important!!)

Signature files

- Quick and dirty, based on hashing
- Underestimate the distances (important!!)

Kasim	00110101
Selcuk	10010110
(bit-or)	10110111

Signature files

- Quick and dirty, based on hashing
- Underestimate the distances (important!!)

Kasim	00110101
Selcuk	10010110
(bit-or)	10110111

- Length of the signature?
- Ratio of the 1s to 0s?

Signature files

- *l* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits

Signature files

- *l* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- If there are *b* keywords, the probability that a given bit is set is

$$1-\left(1-\frac{1}{B}\right)^{bl}$$

Signature files

- *l* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- If there are *b* keywords, the probability that a given bit is set is

$$1-\left(1-\frac{1}{B}\right)^{bl} \approx 1-e^{-b\alpha} \qquad \alpha=\frac{l}{B}$$

Signature files

- *I* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- Probability that *I* random bits in the query are also set in the signature

$$\left(1-e^{-b\alpha}\right)^{I} = \left(1-e^{-b\alpha}\right)^{\alpha B}$$

Signature files

- *I* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- Probability that *I* random bits in the query are also set in the signature

Minimized if $\alpha = \frac{\ln(2)}{b}$

$$\left(1-e^{-b\alpha}\right)^{I} = \left(1-e^{-b\alpha}\right)^{\alpha B}$$

Signature files

- *I* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- Optimal *I* is $B\frac{\ln(2)}{b}$

Signature files

- *l* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- Optimal *l* is $B\frac{\ln(2)}{b}$
- False hit rate under the optimal *l* is $\frac{1}{2^l}$

Signature files

- *l* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- Optimal *l* is $B\frac{\ln(2)}{b}$

There is a tradeoff between the length of the signature and the false hits!

- False hit rate under the optimal *l* is $\frac{1}{2^l}$

Signature files

- *l* of *B* bits are randomly set
  - too many 1s would cause false hits
- Minimize false hits
- Optimal *l* is $B\frac{\ln(2)}{b}$

There is a tradeoff between the length of the signature and the false hits!

- False hit rate under the optimal *l* is $\frac{1}{2^l}$

Experimentally: a good α is 0.5