SCAM: A Copy Detection Mechanism for Digital Documents

Narayanan Shivakumar, Hector Garcia-Molina
Department of Computer Science
Stanford University
Stanford, CA

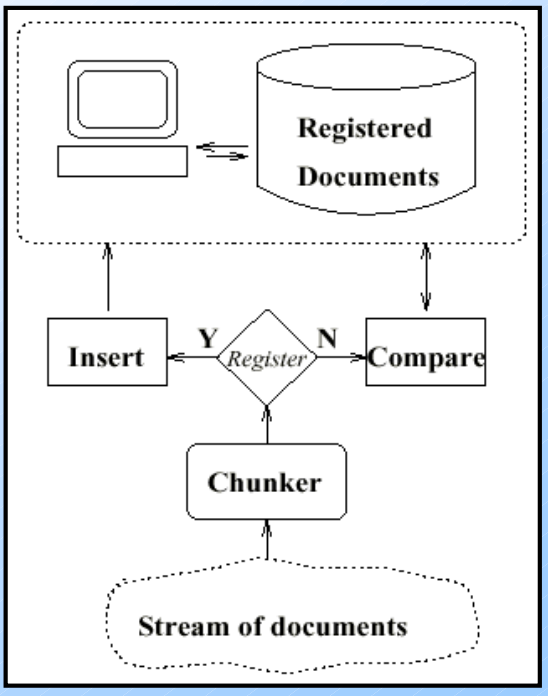
Presented by Dennis Peper

Introduction

- Digital libraries may become popular in the future.
- We don't know how to prevent illegal copying effectively.
- We can detect illegal copying. Methods are:
  - Signature based (e.g. Watermarking.)
  - Registration based (SCAM's approach)
    - Original document is registered with a server.
    - Subsequent documents are compared with pre-registered documents.

Copy Detection Overview

- Document is broken up into chunks.
  - A chunk could be a word, sentence or sequence of sentences, depending on the detection method.
  - SCAM is based on word chunking.
- The chunks of a registered document are stored in a database.
- A new document is broken up and compared to the chunks in the database.
- If overlap exceeds a certain threshold (depending on detection system used), then the author is notified.



Measuring Overlap

- Vocabulary V = {a,b,c,d,e,f,g,h} are the unique chunks in our library.
- Given a registered document R=<a,b,c>. What should our system return, when comparing R to other documents D_1, D_2,…?

Document	Comparison with R
D_1=<a,b>	Quite similar
D_2=<a,b,c>	Exact replica
D_3=<a^k> k≥1	Somewhat similar at low k, not very similar for high k. → e.g. when k=2, D_3=<a,a>
D_4=<a,b,c,d,e,f,g,h>	Significant overlap

Vector Space Model

- We need a method to compare registered document R to incoming documents.
- F(d) is the frequency vector. (e.g. F(R)=<1,1,1,0,…>)
- Dot product of normalized frequency vectors.
  - R=<a,b,c> V(R)=<1/3,1/3,1/3,0,...>
  - D_1=<a,b> V(D_1)=<1/2,1/2,0,0,...>
  - V(R)•V(D_1)=1/3x1/2+1/3x1/2=1/3

There is 1 'a' chunk, 1 'b' chunk, etc...

Document	Dot Product
D_2={a,b,c}	V(R)•V(D_2)=1/3
D_3={a^k} k≥1	V(R)•V(D_3)=1/3
D_4={a,b,c,d,e,f,g,h}	V(R)•V(D_4)=1/8

- The overlap of R with D_2 and D_4 is more significant than that with D_1 or D_3.
- The dot product does not report this.

Vector Space Model (Cont.)

2. Cosine Similarity measure. If R and Q are documents:

$$sim(R,Q) = \frac{\sum_{i=1}^{N} \alpha_i^2 * F_i(R) * F_i(Q)}{\sqrt{\sum_{i=1}^{N} \alpha_i^2 * F_i^2(R) * \sum_{i=1}^{N} \alpha_i^2 F_i^2(Q)}}$$

α_i is a weighting associated with the i^{th} chunk, set to 1 in the example.

- The higher the frequency of the word, the less the word contributes to the similarity result.
- R={a,b,c} F(R)=<1,1,1,0,...>
- D_1={a,b} F(D_1)=<1,1,0,0,...>
- sim(R,D_1)=0.2

HIDDEN SLIDE

Document	Cosine Similarity Measure
D_2={a,b,c}	sim(R,D_2)=1
D_3={a^k} k≥1	sim(R,D_3)=0.58 → F(D_3)=<k,0,0,…>
D_4={a,b,c,d,e,f,g,h}	sim(R,D_4)=0.61

- Similarity for D_3 is is independent of k. Ideally similarity decreases as k increases.
- Similarity for D_4 is low considering the entirety of R contained within D_4.

Relative Frequency Model - SCAM(?)

- For documents R and S define a closeness set c(R,S) to contain all words w_i with similar frequencies $F_i(R)$ and $F_i(S)$.
  - "Similar" is defined by $s_i = Fi(R)/Fi(S) + Fi(S)/Fi(R) < \varepsilon$.
  - ε must be chosen in the range 2.0001 to infinity.
  - If either $F_i(S)$ or $F_i(R) = 0$, then don't include s_i in c(R,S)
- Note: if a word occurs the same number of times in both documents, it is added.
- If $F_i(R)=3$ and $F_i(S)=2$, then c(R,S) contains w_i if $\varepsilon < 2.17$.
- We can see that ε is a tolerance factor.

Relative Frequency Model (Cont.)

- Subset measure. How much of a subset is R wrt S? (my understanding.)
- $subset(R,S) = \left(\sum_{w_i \in c(R,S)} \alpha_i^2 F_i(R)F_i(S) \right) / \left(\sum_{i}^{N} \alpha_i^2 F_i^2(R) \right)$
- α_i is a weighting associated with the i^{th} chunk, this allows adding an importance factor.
- $sim(R,S)=\max\{subset(R,S), subset(S,R)\}$
- If $sim(R,S) \geq 1$, set the result to 1, no extra information is gained as both documents are very similar anyway.
- $R=\{a,b,c\}$ $D_1=\{a,b\}$
- $sim(R,D_1)=\max\{(1x1+1x1)/3,(1x1+1x1)/2\}=1$ (ε=2.0001 or 3)

Document	sim result,ε=2.0001	sim result,ε=3
$D_2=\{a,b,c\}$	$sim(R,D_2)=1$	$sim(R,D_2)=1$
$D_3=\{a^k\}$ $k\geq 1$	$sim(R,D_3)=1,k=1$	$sim(R,D_3)=1,k=1$
	$sim(R,D_3)=0,k>1$	$sim(R,D_3)=2/3,k=2$
		$sim(R,D_3)=0,k\geq 3$
$D_4=\{a,b,c,d,e,f,g,h\}$	$sim(R,D_4)=1$	$sim(R,D_4)=1$

Conclusion

- What about detecting copying from multiple sources?

- Compared against another method, COPS, and results favored SCAM.
- Has higher false positive rate.

Thank you, any questions?