CSE 321  Discrete Structures

March 3rd, 2010
Lecture 22 (Supplement): LSH
Jaccard

• Jaccard similarity: \( J(S, T) = \frac{|S \cap T|}{|S \cup T|} \)

• Problem: given large collection of sets \( S_1, S_2, \ldots, S_n \), and given a threshold \( s \), find all pairs \( S_i, S_j \) s.t. \( J(S_i, S_j) > s \)
Application: Collaborative Filtering

- We have n customers: 1, 2, ..., n
- Each customer i buys a set of items Si
- We would like to recommend items bought by customer j, if J(Si, Sj) > s
Example

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toothpaste</td>
<td>Floss</td>
<td>Ipod</td>
<td>Ipod</td>
<td>Floss toothpaste</td>
</tr>
<tr>
<td>floss</td>
<td>mouthwash</td>
<td>PowerBook</td>
<td>mouthwash</td>
<td>mouthwash</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VideoAdapter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

New customer buys mouthwash, floss; What do you recommend?
Application: Similar Documents

• Given \( n \) documents 1, 2, \( \ldots \), \( n \)

• Let \( S_i \) be the set of \( q \)-grams for document \( I \)

• Want to find all pairs of “similar” documents, i.e. for which \( J(S_i, S_j) > s \)
Example

• You work for a copyright violation detection company
• Customers: have documents 1, 2, 3, …., $10^6$
• Web: has pages 1, 2, 3, …., $10^{11}$

• Your job is to find “almost identical documents”
• What do you do ???
The Signature Method

• For each $S_i$, compute a signature $\text{Sig}(S_i)$ s.t.

\[ J(S_i, S_j) > s \iff \text{Sig}(S_i) \ast \text{Sig}(S_j) \neq \text{emptyset} \]

With high probability
The Signature Method

• Step 1: compute all pairs i,j for which Sig(Si) * Sig(Sj) ≠ emptyset
  – This is a join operation!

• Step 2: for all such pairs, return (i,j) if J(Si, Sj) > s
  – Hopefully only a few such pairs

Both false positives and false negatives are possible
The Signature Method

Will construct the signature in two steps:
1. Minhashes
2. LSH
Minhashing

• Let $\pi$ be an arbitrary permutation of the domain

• For each $i$, let:

\[
mh(S_i) = \{\text{the smallest element in } S_i \text{ according to } \pi\}
\]
Example

• The entire domain is \{a,b,c,d,e,f,g,h\}
• The set \(S_i\) is

\[
S_i = \{a,b,c,e,f\}
\]

• Suppose we choose the permutation:

\[
\pi = d,g,c,h,b,f,a,e
\]

Then what is \(mh(S_i) = ?\)
Minhashing

Main property:

\[
\text{Probability}(\text{mh}(S_i) = \text{mh}(S_j)) = J(S_i, S_j)
\]
Warmup Question

• Choose a *random* permutation $\pi$ of 
\{a,b,c...,z\}
• Consider the set $S=$\{a,b,c,d\}
• What is the probability that $mh(S) = c$?
Warmup Question

• Choose a \textit{random} permutation $\pi$ of 
$\{a, b, c, \ldots, z\}$

\[
S = \begin{array}{ccc}
    a & b & d \\
    c & & \\
\end{array}
\]

What is the probability that $mh(S) = c$?
Warmup Question

• Choose a *random* permutation $\pi$ of \{a,b,c…,z\}

$S=$

What is the probability that $mh(S) = c$?

Answer: $P = \frac{1}{4}$ (each of a, b, c, d can be the min)
Main Property

What is Prob(mh(Si) = mh(Sj))?
What is \( \text{Prob}(\text{mh}(S_i) = \text{mh}(S_j)) \) ?

Answer: \( \frac{2}{6} = J(S_i, S_j) \)
Computing Minhashes

- We use a hash function (which we assume is random)

```
compute_minhash(S) {
    \( v = \infty \);
    \text{for all } x \text{ in } S \text{ do }
    \quad \text{if } h(x) < v \text{ then } \{ v = h(x); y = x; \}
    \text{return } y;
}
```
Example

• The set $S_i$ is $S_i = \{a,b,c,e,f\}$

• Compute $h$:
  $h(a)=77, h(b)=55, h(c)=33, h(e)=88, h(f)=66$

• Then what is $m_h(S_i) =$ ?
Usage Idea

• Recall: we have $n$ sets $S_1, \ldots, S_n$

• Compute $mh(S_1), \ldots, mh(S_n)$

• Consider only those pairs for which $mh(S_i)=mh(S_j)$: compute their Jaccard similarity

• But too many false negatives!

• How can we improve?
Improvement

• Independent hash functions $h_1, \ldots, h_m$

• For each $S_i$, compute $MH(S_i) = \text{the } m \text{ minhashes for each } j=1,\ldots,m$
Example

• The set $S_i$ is $S_i = \{a, b, c, e, f\}$

• Compute $h_1$:
  $h_1(a) = 77, h_1(b) = 55, h_1(c) = 33, h_1(e) = 88, h_1(f) = 66$

• Compute $h_2$:
  $h_2(a) = 22, h_2(b) = 66, h_2(c) = 55, h_2(e) = 11, h_2(f) = 44$

• Then what is $MH(S_i) = \ ?$
Example

• The set $S_i$ is $S_i = \{a, b, c, e, f\}$

• Compute $h_1$:
  $h_1(a) = 77, h_1(b) = 55, h_1(c) = 33, h_1(e) = 88, h_1(f) = 66$

• Compute $h_2$:
  $h_2(a) = 22, h_2(b) = 66, h_2(c) = 55, h_2(e) = 11, h_2(f) = 44$

• Then what is $MH(S_i) =$ ?

• Answer: $MH(S_i) = (c, e)$ (an ordered pair)
Using Minhashes

• Compute MH(S1), …, MH(Sn)

• $J(S_i, S_j) \approx$ the fraction of positions where MH(Si) and MH(Sj) agree
Example

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Estimate $J(S1, S2) = ?$, then $J(S1, S3) = ?$
Note: Minhashes still require Jaccard

• How do we compute the fraction of positions where MHi and MHj agree?
• Annotate each position with its position number
• Then this is precisely $J(MHi, MHj)$
Note:
Minhashes still require Jaccard

\[ \text{Fraction of equal positions} = \frac{4}{m} \]

\[
\begin{array}{ccc}
1 & 1 & 4 \\
2 & 4 & 4 \\
3 & 8 & 8 \\
.. & 4 & 7 \\
.. & 8 & 8 \\
m & 6 & 6 \\
\end{array}
\]

\[
\text{MH1} = \{(1,1),(2,4),(3,8),(4,4),(5,8),(6,6)\}
\]

\[
\text{MH2} = \{(1,4),(2,4),(3,8),(4,7),(5,8),(6,6)\}
\]

\[
J(\text{MH1}, \text{MH2}) = \frac{4}{2m-4}
\]
Comments on Minhashes

• It is not a signature yet!
• We have only reduced the problem of computing \( J(S_i,S_j) \) to the problem of computing \( J \) on smaller sets, of size \( m \)

• The signature is provided by LSH (next)
Locality Sensitive Hashing

- We have \( n \) strings \( MH_1, \ldots, MH_n \), each of length \( m \)
- Compute a signature \( \text{Sig}(MH_i) \) for each string \( MH_i \)

Desired property:

\[
\text{Sig}(MH_i) * \text{Sig}(MH_j) \neq \text{empty} \iff J(MH_i, MH_j) > s
\]
LSH

- Each MHi consists of m values
- Each signature Sig(MHi) will have size b
- Divide the m values into b “bands” of size r “rows”, i.e. m = b*r
- For each band j=1,…,r, apply a hash function hj to the string of values in band j in MHi \( \Rightarrow hj \)
- Then Sig(MHi) = (h1, h2, …, hb)
Each rectangle (=band) becomes one new hash value.
Analysis

• Goal: want to compute the probability that $\text{Sig(MH(Si))} \ast \text{Sig(MH(Sj))} \neq \text{emptyset}$, as a function of $s = J(Si, Sj)$

• So let $s = J(Si, Sj) = \text{fraction of equal positions}$
Analysis

\[ J(S_i, S_j) = s \]

What is the probability that two entries are equal?
What is the probability that two entries are equal?

Answer: $s$
Analysis

J(S_i, S_j) = s

What is the probability that the bands are equal?
Analysis

\[ J(S_i, S_j) = s \]

What is the probability that the bands are equal?

Answer: \( s^r \)
What is the probability that some pair of bands are equal?

$J(S_i, S_j) = s$
Analysis

\[ J(S_i, S_j) = s \]

What is the probability that some pair of bands are equal?

Answer: \[ 1 - (1 - s^r)^b \]
This is precisely the probability that $\text{Sig}(S_i) \times \text{Sig}(S_j) \neq \text{emptyset}$

$J(S_i,S_j) = s$

What is the probability that some pair of bands are equal?

Answer: $1 - (1 - s^r)^b$
Analysis

Probability $1 - (1 - s^r)^b$
Putting it together

• You work for a copyright violation detection company
• Customers: have documents 1, 2, 3, …. , 10^6
• Web: has pages 1, 2, 3, …. , 10^{11}

• Your job is to find “almost identical documents”
• What do you do ???
# Step 1: Q-grams

<table>
<thead>
<tr>
<th>DocID</th>
<th>Qgram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>To</td>
</tr>
<tr>
<td>1</td>
<td>ob</td>
</tr>
<tr>
<td>1</td>
<td>be</td>
</tr>
<tr>
<td>1</td>
<td>be</td>
</tr>
<tr>
<td>1</td>
<td>e o</td>
</tr>
<tr>
<td>1</td>
<td>or</td>
</tr>
<tr>
<td>1</td>
<td>or</td>
</tr>
<tr>
<td>1</td>
<td>r n</td>
</tr>
<tr>
<td>1</td>
<td>no</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>2</td>
<td>. .</td>
</tr>
</tbody>
</table>

### Web

<table>
<thead>
<tr>
<th>URL</th>
<th>Qgram</th>
</tr>
</thead>
<tbody>
<tr>
<td>abc.com</td>
<td>Oba</td>
</tr>
<tr>
<td>abc.com</td>
<td>Bam</td>
</tr>
<tr>
<td>abc.com</td>
<td>ama</td>
</tr>
<tr>
<td>abc.com</td>
<td>ma</td>
</tr>
<tr>
<td>abc.com</td>
<td>a h</td>
</tr>
<tr>
<td>bcd.com</td>
<td>. .</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
Step 2: Compute $m$ Min-hashes

<table>
<thead>
<tr>
<th>docID</th>
<th>mh1</th>
<th>mh2</th>
<th>mh3</th>
<th>mh4</th>
<th>…</th>
<th>mh500</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>To</td>
<td>que</td>
<td>Ion</td>
<td>or</td>
<td></td>
<td>Not</td>
</tr>
<tr>
<td>2</td>
<td>…</td>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>url</th>
<th>mh1</th>
<th>mh2</th>
<th>mh3</th>
<th>mh4</th>
<th>…</th>
<th>mh500</th>
</tr>
</thead>
<tbody>
<tr>
<td>abc.com</td>
<td>sec</td>
<td>ret</td>
<td>def</td>
<td>ens</td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>bcd.com</td>
<td>…</td>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: $m = \text{a few hundreds}$
Step 3: Compute Signatures

<table>
<thead>
<tr>
<th>docID</th>
<th>h(mh1, ..., mh20)</th>
<th>h(mh21, ..., mh40)</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2345234</td>
<td>3232</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>docSIG</th>
<th>docID</th>
<th>Sig</th>
<th>webSIG</th>
<th>url</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2345234@1</td>
<td></td>
<td>abc.com</td>
<td>676876@1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3232@2</td>
<td></td>
<td>abc.com</td>
<td>3232@2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>...</td>
<td></td>
<td>abc.com</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>452342@25</td>
<td></td>
<td>abc.com</td>
<td>787892@25</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>23423@1</td>
<td></td>
<td>bcd.com</td>
<td>23423@1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

CSE 321 Winter 2010 -- Dan Suciu
Step 4: Find docs with common signatures

Need two indexes:
for every signature $s$, $\text{doc}[s] = \text{set of documents containing } s$
for every webpage $w$, $\text{web}[] = \text{set of webpages containing } s$

For all $s$ in $\text{Sig}$ do
  for $d$ in $\text{doc}[s]$, for $w$ in $\text{web}[s]$ do
    print($d$, $w$)