NOTE

Prefix Filtering

一些Similarity定义

Jaccard $J(x,y) = |\frac{x \cap y}{x \cup y}|$
Cosine $C(x,y) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} = \frac{\sum_i x_i y_i}{\sqrt{|x|} \cdot \sqrt{|y|}}$
Overlap $O(x,y) = |x \cap y|$

例1 规格化Record

$D_x =$ "$yes\ as\ soon\ as\ possible$"
$D_y =$ "$as\ soon\ as\ possible\ please$"

统计:

Word	yes	as	soon	as1	possible	please
Token	A	B	C	D	E	F
Doc. Freq.	1	2	2	2	2	1

排序:
$x = [A, B, C, D, E]$
$y = [F, B, C, D, E]$

相似度:
$J(x,y) = \frac{4}{6} = 0.67$
$C(x,y) = \frac{4}{\sqrt{5} \cdot \sqrt{5}} = 0.80$

Lemma 1 - Prefix Filtering Principle

> 对于一个已规格化的$Record$集。
> 若$O(x,y) \geq \alpha$,
> 在x的$(|x| - \alpha + 1)$前缀与y的$(|y| - \alpha + 1)$前缀中,
> $\exists$至少一个相同$token$。
> 其中$\alpha = \lceil \frac{t}{1+t}(|x| + |y|) \rceil$

例2

Jaccard similarity threshold $t = 0.8$

> $w = [\underline{C}, D, F]$
> $z = [\underline{G, A}, B, E, F]$
> $y = [\underline{A, B}, C, D, E]$
> $x = [\underline{B, C}, D, E, F]$

> $prefix_x = |x| - \lceil t \cdot |x| \rceil + 1$

1. 对于prefixes中的token建立倒排索引，如：C -> { (w, 0), (x, 1) }
2. 对于记录x，根据其prefixes中的token，即B和C，可于倒排索引中筛选出候选对，可得{ (x, y), (x, w) }
3. 因为$|w| \leq x \cdot t = 4$，可对于(x, w)予以修剪。

例3

Jaccard similarity threshold t = 0.8

> $y = [\underline{A, B}, C, D, E]$
> $x = [\underline{B, C}, D, E, F]$

(x, y)尽管不符合$O(x, y) \geq \alpha = 5$，但因B被选为候选对。
对此可将除前缀外的剩余token纳入考虑。对于(x, y)，可得$1 + min(3, 4) = 4 \leq 5$,则可将此对予以修剪。

Lemma 2 - Prefix Filtering Principle

> 对于一个已规格化的$Record$集。
> $w = x[i]$，则$x_l(w) = x[1..(i-1)]$、$x_r(w) = x[i..|x|]$。
> 若$O(x, y) \geq \alpha$，$\forall w \in x \cup y$，
> $O(x_l(w), y_l(w)) + min(|x_r(w)|, |y_r(w)|) \geq \alpha$

Lemma 3

对于给定的记录x，$(|x| - \lceil \frac{2t}{1+t} \cdot |x| \rceil + 1)$长前缀足以算法1确定精确结果。

Suffix Filtering

例4

...