WILD CARD QUERIES

WILD CARD QUERIES SUCH AS

Mon* OR *mon ARE EASY TO

DO WITH PREFIX SEARCH:

- α* = find all documents containing words starting with α

SIMPLE PREFIX SEARCH

- *β = find all documents containing words ending in β

PREFIX SEARCH, BUT YOU HAVE TO CONSTRUCT A DATA STRUCTURES CONTAINING THE REVERSE OF EACH TERM

BUT WHAT IF WE SEARCH FOR

α * β ?

- Solution 1

PREFIX SEARCH α*, PREFIX SEARCH *β

AND INTERSECT THE RESULTING LIST

EXPENSIVE !!

EVEN WORSE IF WE SEARCH FOR SOMETHING LIKE : α*β AND γ*γ

LOTS OF COMBINATIONS !!

- Solution 2

PERMUTEX INDEX

IT TRANSFORMS THE WILDCARD QUERY SO THAT THE * OCCUR AT THE END.

CONSTRUCTING AND USING THE PERMUTEX INDEX

(1) For every term in our lexicon, index it under all possible rotation of the word with the special char $

EXAMPLE

Hello is indexed by :

Hello$, ello$H, lloHe, loHel, o$Hell, $Hello

NOTE : Permutem index size $\simeq$ 4 * lexicon size

(2) When querying for $\alpha * \beta$, the search is reduced to a prefix search of the rotated query $\beta \$ \alpha *$

How to do it?

- $X \longrightarrow X\$$ CASO BASE
- $X* \longrightarrow X*\$ \longrightarrow *\$X \rightarrow \$X*$
- $*X \rightarrow *X\$ \rightarrow X\$*$
- $*X* \longrightarrow *X*\$ \rightarrow X*\$* \overset{?}{\longrightarrow} X*$
- $X*Y \rightarrow X*Y\$ \longrightarrow *Y\$X \rightarrow Y\$X*$
- $X*Y*Z \rightarrow$ lookup for $Y\$X*$, then match with $Z\$*$