

2.7 Stemming Algorithms

- _ Goal of stemming was to improve performance and reduce system resources by reducing the number of unique words that a system has to contain.
- _ The stemming process creates one large index for the stem.
- _ 1 Introduction to the Stemming Process
- _ 2 Porter Stemming Algorithm
- _ 3 Dictionary Look-Up Stemmers
- _ 4 Successor Stemmers

2.7a Introduction to the Stemming Process

- _ Stemming algorithms are used to improve the efficiency of the information system and to improve recall.
- _ *Conflation*
 - _ the term used to refer to mapping multiple morphological variants to a single representation (stem).
- _ The principle is that the stem carries the meaning of the concept associated with the word and the affixes.

- _ Languages have precise grammars that define their usage, but also evolve based upon human usage.
- _ Thus exceptions and non-consistent variants are always present in languages that typically require 'exception look-up tables' in addition to the normal reduction rules.

- _ The idea of equating multiple representations of a word as a single stem term is to provide significant compression.
- _ For example, the stem "comput" could associate "computable, computability, computation, computational, computed, computing, computer, computerese, computerize" to one compressed word.
- _ compression of stemming does not significantly reduce storage requirements.
- _ misspellings and proper names reduce the compression even more.
- _ Another major use of stemming is to improve recall.
- _ As long as a semantically consistent, stem can be identified for a set of words.
- _ The process of stemming helps to not to miss relevant items.

- Stemming of the words “calculate, calculates, calculation, calculations, calculating” to a single stem “calculat” assures whichever of those terms is entered by the user, it is translated to the stem and finds all the variants in any items they exist.
- stemming can not improve precision as the precision value is not based on finding all relevant items but just minimizing the retrieval of non-relevant items.

- Stemming can cause problems for Natural Language Processing (NLP) systems by loss of information needed for aggregate levels of natural language processing (discourse analysis).
- The tenses of verbs may be lost in creating a stem, but they are needed to determine if particular concept being indexed occurred in the past or will be occurring in the future.
- Time is one example of the type of relationships that are defined in Natural Language Processing systems.
- Stemming algorithm removes suffixes and prefixes, sometimes recursively, to derive the final stem.

- Other techniques such as table lookup and successor stemming provide alternatives that require additional overheads.
- Successor stemmers determine prefix overlap as the length of a stem is increased.
- Table lookup requires a large data structure.