Abstract of Thesis presented to UFF as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

Stemming Process Investigation for the Portuguese Language
Reinaldo Viana Alvares
março - 2005

Advisor: Ana Cristina Bicharra Garcia
Department: Computer Science

The information retrieval process is a usual task for the human. However, having a complex automation. This happens because the quality of the results is often related with the degree of the user's satisfaction, a difficult parameter to measure. In general this quality is evaluated being taking into account a group of queries in a text collection, and their relevant answers.

Commonly, two evaluation measures are used in this process: the first is the precision, wich represents the proportion of recovered relevant items from the total of recovered items; and the second is the recall, wich represents the proportion of recovered relevant items from the total of relevant items of the collection.

One of the challenges is to find efficient forms to represent the documents, in order to avoid ambiguity. An alternative to solve this problem consists of obtaining a unique representation for words that appear for a same concept. This task can be defined as stemming.

Many times, the stemming process is dependent to the morphologic structure of the target language. For the Portuguese language, there were found few solutions to assist the demand for these algorithms.

The morphologic complexity of Portuguese language, and the few stemming solutions found for this language, were the motivation for the research shown in this