Introducion
The number of digital collection is increasing with the development of Internet. A method to automatically extract concepts from digital collection is becoming more and more important.

C-value/NC-value [1] is one of the method to automatically extract multi-words terms from corpus using Automatic Term Recognition(ATR). The objective of this project is to implement a Java program using C-value/NC-value to automatically extract important concepts from Textbooks.

Tagging
Use part-of-speech tagger to tag the corpus.

Before Tagging
Janathan and Tina are building an outdoor Skating rink.

After tagging
Jonathan and Tina are building an outdoor skating rink.

Linguistic filter
Use linguistic filter to extract words and phrase from corpus.

1. Noun+Noun,
2. (Adj|Noun)+Noun
3. ((Adj|Noun)+|((Adj|Noun)*NounPrep)?)

Stoplist
Stoplist is a list of words that should be avoided when extracting phrases from corpus.

NC-value is used to assign weight to context words of candidate terms. It will enhance accuracy of C-value.

C-value

\[
C\text{-value}(a) = \begin{cases} 
\log_2[a] \cdot f(a) & \text{if } a \text{ is not nested,} \\
\log_2[a] \cdot f(a) - \frac{a_{not\,nested}}{P(T_a)} \sum_{b \in T_a} f(b) & \text{otherwise}
\end{cases}
\]

where
- \(a\) is the candidate string,
- \(f(.)\) is its frequency of occurrence in the corpus,
- \(T_a\) is the set of extracted candidate terms that contain \(a\),
- \(P(T_a)\) is the number of these candidate terms.

Conclusion
We know that, the higher C-value candidate term has, the more likely it is an important term in corpus.

We test C-value with a high school math textbook. In the result of our experiment, rate of change has the highest C-value, which followed by instantaneous rate of change, average rate of change, function, instantaneous rate.

With the help of Automatic Term Recognition(ATR), we could help a student improve his/her learning process, especially when the learning is done in English as a second language.

Acknowledgements


Thank you to Professor Diana Inkpen for her supervision and to the UROP project for this opportunity.

Poster made by: Yusi Fan (yfan094@uottawa.ca)