Aparna Ladekar, Archana Mujumdar, Prajakta Nipane, Sonam Titar, Guide: Mrs. Kavitha S.
/ International Journal of Engineering Research and Applications (IJERA)
ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.1551-1555

Automatic Text Summarization Using:Fuzzy GA-GP

Aparna Ladekar, Archana Mujumdar, Prajakta Nipane, Sonam Titar, Guide: Mrs. Kavitha S.

ABSTRACT

The rapid growth of the data in the Internet has overloaded the user with enormous amounts of information which is more difficult to access huge volumes of documents. "Automatic Text Summarization" technique is an important activity in the analysis of high volume documents. The proposed system generates a summary for a given input document based on identification and extraction of important sentences in the documents by reducing the redundancy of data. A novel technique is proposed for summarizing text using a combination of Genetic Algorithms (GA) and Genetic Programming (GP) to optimize rule sets and membership functions of fuzzy systems.

KEYWORDS: 'Text summarization',' genetic algorithm',' genetic programming',' fuzzy system'

1. INTRODUCTION

1.1 Automatic text summarization:

Automatic summarization is the creation of a shortened version of a text by a computer program. The product of this procedure still contains the most important points of the original text. The phenomenon of information overload has meant that access to coherent and correctly-developed summaries is vital. As access to data has increased so has interest in automatic summarization. An example of the use of summarization technology is search engines such as Google.

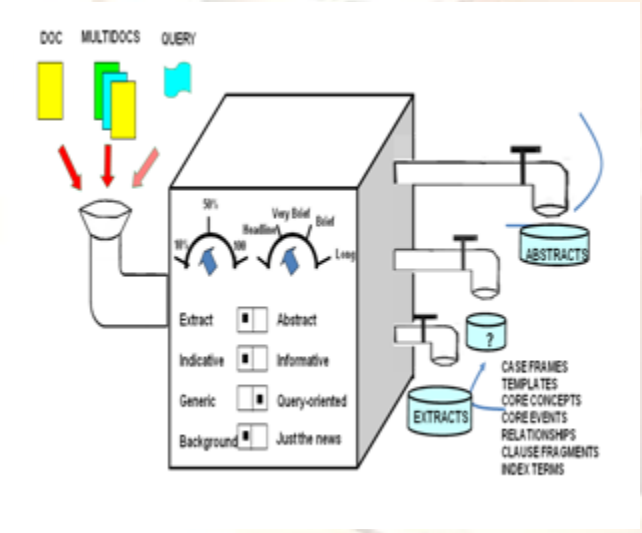

Fig 1.2 A summarization machine

Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate.

Our project is based on genetic programming and genetic algorithm with fuzzy set of rules.

1.2 Genetic Programming (GP)

Genetic Programming (GP) is an evolutionary algorithm that evolves computer programs and predicts mathematical models from experimental data. It uses fixed length character strings to represent computer programs which are later expressed as expression trees. The basic purpose is optimization. The main operators used in evolutionary algorithms are regeneration crossover and mutation.

1.3 Genetic Algorithm (GA)

Genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. It uses techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

1.4 Fuzzy Systems

A fuzzy system is an alternative to traditional notions of set membership and logic. Their application is at the leading edge of Artificial Intelligence.

Fig 2.3 fuzzy system

- Fuzzy Systems include Fuzzy Logic and Fuzzy Set Theory.
- Knowledge exists in two distinct forms:

-The Objective knowledge exists in mathematical form is used in engineering problems; - The Subjective knowledge that exists in linguistic form, usually impossible to quantify.

Fuzzy Logic can coordinate these two forms of knowledge in a logical way. The applications of Fuzzy Systems are many like Information retrieval systems, Navigation system, and Robot vision.

The Fuzzy Set Theory - membership function, operations, properties and the relations have been described in previous lectures.

2. PROPOSED METHODOLOGY:

Summarization can be done with the help of the following processes:

(a) Selection of what is 'important'

(b) Omission of what is 'unimportant'

(c) Generalization from the particular and specific

(d) Identification of general (global) structures

Methodologically, summaries can build up from the details (of a micro-level) by generalization, selection, and omission, or they can work by extraction from within global frameworks (macro-level).

Summaries differ according to the emphases placed on each of these processes. In evaluative and selective summaries the first two processes (a) and (b) dominate. The dominant process in informative summaries is (c) 'generalization'.

2.1 SCORING ALGORITHM:

Our program essentially works on the following logics:

2.1.1 WORD SCORING:

1. **Stop Words:** These are some insignificant words that are so commonly used in the English language that no text can be created without them. They therefore provide no real idea about the textual theme, and have therefore, been neglected while scoring sentences.

E.g. I, a, an, of, am, the, et cetera.

2. **Cue Words:** These are words usually used in concluding sentences of a text, making sentences containing them crucial for any given summary. Cue Words provide closure to a given matter, and have therefore, been given prime importance while scoring sentences.

E.g. Thus, hence, summary, conclusion, et cetera.

3. **Basic Dictionary Words:** 850 words of the English language have been defined as the most frequently used words that add meaning to a sentence. These words form the backbone of our algorithm, and have been vital in the creation of a sensible summary. We have hence, given these words moderate importance while scoring sentences.

4. **Proper Nouns:** Proper Nouns in most cases form the central theme of a given text. Albeit, the identification of proper nouns without the use of linguistic methods was difficult, we have been successful in identifying them in most cases. Proper Nouns provide semantics to the summary, and have therefore been given high importance while scoring sentences.

5. **Keywords:** The user has been given an option to get a summary generated which contains a particular word, the keyword. Though this is greatly limited by the absence of NLP, we have tried our best to produce results.

6. **Word Frequency:** Once basic scores have been allotted to words, their final score is calculated on the basis of their frequency of occurrence in the document. Words in the text which are repeated more frequently than others contain a more profound impression of the context, and have therefore been given a higher importance.

2.1.2 SENTENCE SCORING:

1. **Primary Score:** Using the above methods, a final word score is calculated, and the sum of word scores gives a sentence score. This gives long sentence a clear advantage over their smaller counterparts, which might not necessarily be of lesser importance.

2. **Final Score:** By multiplying the score so obtained by the ratio "average length / current length" the above drawback can be nullified to a large extent, and a final sentence score is obtained.

The most noteworthy aspect has been the successful merger of frequency based and definition based categorization of words into one efficient algorithm to generate an as complete as possible summary for a given sensible text.

3. Proposed Algorithm:

- Initially, a parser is designed that extracts the desired features. This program parses the text into its sentences and identifies the following nonstructural features for each sentence as the input of fuzzy inference system:

  1- The number of title words in the sentence,

  2- Whether it is the first sentence in the paragraph,

  3- Whether it is the last sentence in the paragraph,

  4- The number of words in the sentence,

  5- The number of thematic words in the sentence, and

  6- The number of emphasize words.

- The features extracted in previous section are used as inputs to the fuzzy inference system. We partition these inputs to several fuzzy sets whose membership functions cover all the universe of discourse.
- IF (sentence-location is first) and (Number-of-title-words is very much) and (Sentence-length is not short) and (Number-of thematic- words is many) THEN (Sentence is important).
- Check also with genetic algorithm and genetic programming.
- Choose the best possible solution for it.

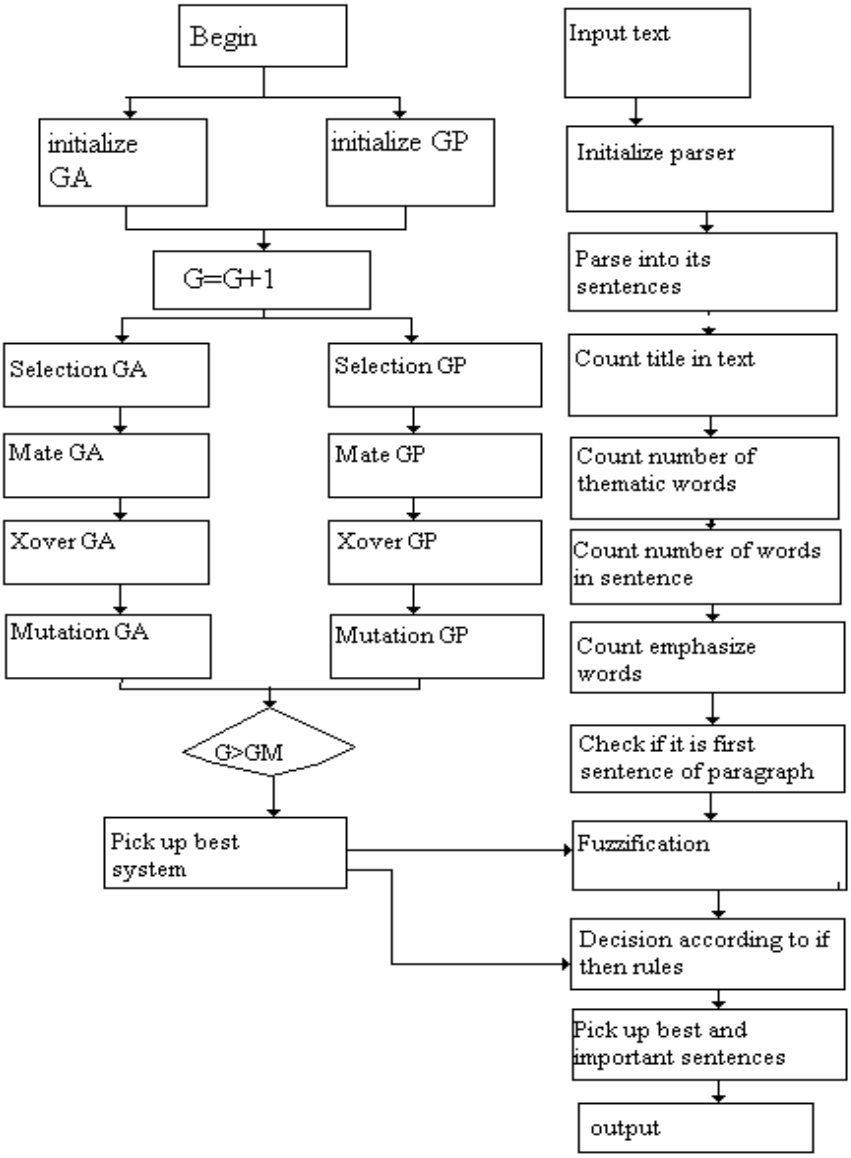

Fig 4 Our proposed algorithm

4. Mathematical Model:

Mathematical models are the functions that we use in our project.

Given are the functions.

Bell membership functions which are specified by three parameters {a, b, c} as in (3). The parameters c and a, determine the center and width of the MF, and then use b to control the slopes at the crossover points.

Bell(x, a, b, c) = 1/1+ ((x-c)/a) raised to 2b

Fuzzy IF_THEN rules

IF (sentence-location is first) and (Number-of-title-words is very much) and (Sentence-length is not short) and (Number-of thematic- words is many) THEN (Sentence is important)

Fitness function

1. Maximization of thematic words in the summary θ s

Per original article θ o:

Θs/θo = (thematic words ratio)

2. Maximization of Emphasized words in the summary Es

Per the original article Eo.

Es/Eo= (Emphasize words ratio)

3. Maximization of frequency of the title words in the summary Ts in compare with the original article To

Ts/To = (Title words ratio)

4. Minimization of Overlap between words in the summary sentences which summary should give us fresh information about the original article. It is calculated with that is the ratio of the frequency of the same words in the summary Σ (Ψ) per all words in the summary Ks.

Σ (Ψ)/Ks = (Overlap ratio)

5. Minimization of the length of the summary as a ratio of the words in the summary Ks per words in the original text Ko.

Ks/Ko = (length ratio)

So fitness of membership functions and rule set individuals are computed as follows:

$$f = \alpha \times \frac{\theta_s}{\theta_o} + \beta \times \frac{1}{\frac{K_s}{K_o}} + \gamma \times \frac{E_s}{E_o} + \eta \times \frac{T_s}{T_o} + \delta \times \frac{1}{\frac{\Sigma(\Psi)}{K_s}}$$

5. CONCLUSION

A novel approach is proposed that extracts sentences based on an evolutionary fuzzy inference engine.

The evolutionary algorithm uses GA and GP in concert.

The genetic algorithm is used to optimize the membership functions and genetic programming is used to optimize the rule sets.

6. REFERENCES

- Khosrow Kaikhah, "Automatic Text Summarization with NNs,", Second IEEE International Conference On Intelligent Systems, June 2004 PP 40-44.
- From Wikipedia, the free encyclopedia.
- Yatsko V. A., Vishnyakov T. N. A method for evaluating modern systems of automatic text summarization. In: Automatic Documentation and Mathematical Linguistics. - 2007. - V. 41. - No 3. - P. 93-103.
- M.-R. Akbarzadeh-T., I. Mosavat, and S. Abbasi, "Friendship Modeling for Cooperative Co-Evolutionary Fuzzy Systems: A Hybrid GAGP Algorithm," Proceedings of the 22nd International Conference of North American Fuzzy Information Processing Society, pp.61-66, Chicago, Illinois, 2003.
- http://nymag.com/daily/politics/autosummary/
- http://forums.devshed.com/software-design-43/auto-summarization-tool-help-needed-508297.html
- http://www.aboutus.org/AutoSummary.com
- Wikipedia has provided the 850 Basic English Words (Appendix: Basic English words list) Most techniques struck while we tried to create summaries manually.
- "Neural Network, Fuzzy Logic, and Genetic Algorithms - Synthesis and Applications", by S. Rajasekaran and G.A. Vijayalaksmi Pai, (2005), Prentice Hall, Chapter 7, page 187-221.
- Related documents from open source, mainly internet. An exhaustive list is being prepared for inclusion at a later date.
- 2006 IEEE International Conference on Fuzzy Systems Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.