University of Bergen

Master Thesis

Optimal Compact Trie Clustering - A genetic approach

Author:
Snorre Magnus DavÃÿen

Supervisor:
Richard Elling Moe

in the

Department of Information Science and Media Studies

September 2013

“Quote comes here"

Quote Author

UNIVERSITY OF BERGEN

Abstract

Faculty of Social Sciences

Department of Information Science and Media Studies

Masters degree

Optimal Compact Trie Clustering - A genetic approach

by Snorre Magnus DavÃŸen

Abstract comes here!

Acknowledgements

Acknowledgements!

Contents

List of Figures

List of Tables

List of Listings

Chapter 1

Introduction

Introduce the background for the master thesis work. What is information retrieval, why is it used. Shortly talk about clustering and classification and how it is used to help make order in large document collections. What kind of problems might clustering solve?

Introduce the Klimauken project, and the compact trie clustering algorithm.

1.1 Motivation

Academic motivations:

The Compact Trie Clustering algorithm has some benefits (being computationally fast, supporting phrases), but performs poorly on data sets which have not been pre-filtered by say a search engine. How can this master thesis work contribute to the performance/-effectivity of the algorithm. Why is this a interesting area to research.

Other motivations:

Information retrieval an interesting field. Genetic algorithms an interesting means to the optimization of large parameter sets. Computing fitness with genetic algorithm a challenging and exiting task. The opportunity to learn a new programming language (Python).

1.2 Research question

The main research question of this thesis is:

"What are the optimal parameter values for Compact Trie Clustering with regard to the news corpus?"

Explanation of research question here...

A list of subsidiary goals here:

- Develop a method/software for determining optimal parameters for the Compact Trie Clustering algorithm.
- Apply the method/software to determine recommended parameter values for news documents.

Why is this an interesting research question, and how does it relate to the motivation.

Target audience: Researchers and users of the suffix tree/compact trie clustering algorithm in general, and the research group working with Klimauken and other information retrieval research in particular.

What are the limitations imposed upon the research. I.e. scope of research, tests performed, number of corpora used etc.

Chapter 2

Theory

Introduction to chapter. I will rewrite and in general use the theory section from the project proposal.

2.1 Clustering and information retrieval

A general introduction to text document clustering should be introduced here.

2.1.1 Suffix Trees and Suffix Tree Clustering

What is a suffix tree? How is one built?

2.1.2 Compact Trie Clustering

What is the compact trie clustering algorithm...

2.1.3 Performance measures

What kind of performance measures are used for clustering...

2.1.4 Available corpora

Which corpora are used in clustering and/or classification research? Which ones are suited to clustering? Which ones are used in this master thesis research? Explain scope...

2.2 Genetic Algorithms

General overview of a genetic algorithm here.

2.3 Related Work

Introduce related research work in this chapter.

Chapter 3

Methodology

An introduction to the chapter.

3.1 Experimental Evaluation

Will rewrite and use the corresponding section from the project proposal.

3.1.1 Evaluation Measures

Provide info about the two forms of evaluation measures used (one from Improved suffix tree clustering article, and one used by Richard). How do they work, what are the differences...

3.2 Corpora

Write a bit about the corpora used in the research

3.3 Experimental Research

Rewrite and use corresponding section from project proposal. Try to flesh out the methodology a bit (how did I perform the testing, what where the hypotheses etc). Talk about experimental constraints and data used.

Chapter 4

Development and Testing

Chapter introduction. Describe the development process in some detail. I.e. The learning phase (learning Python and familiarizing myself with the algorithm). The second stage (modifying the algorithm slightly, refactoring). Third stage (implementing a genetic algorithm with which to test different parameter sets). Fourth stage (making a distributed version of the algorithm to make testing of larger parameter sets feasible/possible). Fifth stage (final touches on the algorithm, implementing support for converting other corpora to the snippet file format). (Sixth stage) Storing results in a database and developing a method of extracting statistical numbers. Overview of system?

Describe how the algorithm and parameters where tested. I.e. how data and results were gathered. Write about different testing stages. First stage (testing individual parameter options to determine usable value ranges. Are results correct, why?). Testing of non-distributed genetic algorithm (does the small test-case give indication of viability?). Testing distributed algorithm (larger parameter sets tested show better results?). Final testing (testing hypotheses, include two corpora?).

4.1 Stages of development

This section will outline the main sequence in which work on MaDaME was done. The overview of the iterations will explain when the initial development of the different parts were started to give an short explanation of the development process.

4.1.1 Learning stage

Familiarize with algorithm ...

4.1.2 Modification of algorithm

Modified so and so ...

4.1.3 Genetic algorithm

Explain how and why it was developed as it was ...

4.1.4 Distribution of genetic algorithm

Why distribute? How? Possible problems...

4.1.5 Results storage and corpus processing

Storing results, tracking top chromosomes over generations, extracting averages for graphs etc.

4.1.6 Overview of the completed system

Give a short overview of the system ...

4.2 Testing

Give overview of testing process

4.2.1 Value Range Tests

Describe tests to discover reasonable parameter ranges ...

4.2.2 Genetic Algorithm Test

Describe first test with about 200 individuals and 50 generations.

4.2.3 Distributed Genetic Algorithm Test

Describe distributed test with more individuals and more generations

4.2.4 Final testing

Describe final test and how it answers resesearch question

Chapter 5

Analysis and Discussion

Chapter introduction. Should here provide some information about which parts of the work are going to be discussed. Should talk about the test results and how they correspond to the hypotheses. I.e. Does the testing reveal that a better parameter set has been found than the default one. Does this parameter set perform better than the default in different corpora? (Should perhaps test on two additional corpora?). This chapter should also investigate wether the test results are statistically significant (can I say yes or no on the null hypotheses?). Give definitive or estimated answer pending test results.

5.1 Results

Summarize and discuss results.

5.2 Validity and relevance

Show that data gathered are both valid and relevant. I.e. is the method of research rigorous and correct (methods of data gathering and testing). And does the data answer the hypotheses. Also discuss the statistical significance of the data in relation to hypotheses.

5.2.1 Data autenticity

Discuss how the validity of data should not be an issue even though the algorithm is distributed. (I.e. results from clients are validated). Algorithm deterministic...

5.2.2 Effects of two different measurements

Discuss how the varying measurements might affect the results... Does using one measurement over the other invalidate results? Should both be used (one to measure single category documents, the other to measure multiple category documents)?

Chapter 6

Summary and Conclusion

Summarize motivation

Restate research question *"Is it possible to create a tool which allows naïve users to easily add metadata to their Web sites using natural language?"*

6.1 Results

Summarize results

6.2 Future research

What did I not have time to use? What was out of scope for this thesis? What would I like to investigate further.

6.3 Conclusion

Final remarks