Mou Hao

Clustering Analysis and Approximate Hierarchy Clustering Algorithm

Master's Thesis in Information Technology

March 14, 2016

University of Jyväskylä

Department of Mathematical Information Technology

Author: Mou Hao

Contact information: Ag C416.1, mouhao1990@outlook.com

Supervisor: Unsupervised work

Title: Clustering Analysis and Approximate Hierarchy Clustering Algorithm

Työn nimi: LaTeX-tutkielmapohjan gradu3 käyttö

Project: Master's Thesis

Study line: All study lines

Page count: 24+2

Abstract: Clustering Analysis and Approximate Hierarchy Clustering Algorithm.

Keywords: LaTeX, gradu3, Master's Theses, Bachelor's Theses, user's guide

Suomenkielinen tiivistelmä: Empty TODO

Avainsanat: LaTeX, gradu3, pro gradu -tutkielmat, kandidaatintutkielmat, käyttöohje

Preface

This is where you can write a preface for your thesis. Most theses don't have prefaces, but if you write one, keep it short (at least one page).

The preface should discuss more the thesis process than the content of the thesis. For example, if there is something out of the ordinary in your choice of a thesis topic or if something out of the ordinary happened during its prepararion, the preface is where you could write about it. It is also customary in a preface to thank by name those persons who helped you with your thesis – at least your supervisor, your spouse and your children, if any. (Your family likely will have helped you by encouraging and supporting you.)

The preface is typically in the first person ("I"). It is also common to sign it.

Jyväskylä, March 14, 2016

The Author

Glossary

TEX	A batch-oriented typesetting system written by Donald Knuth in 1977–1989 (**knuth86:_texbook**).
LATEX	A system, built on top of TEX (**knuth86:_texbook**), for typesetting structured documents (**lamport94:_latex**). Its current version is LATEX 2ε.

List of Figures

List of Tables

Contents

1 Introduction

Introduction (describe the structure of the thesis and point out the contributions)

2 Clustering

Clustering (includes normal and hierarchical clustering)

The clustering methods are mainly divided into two categories viz., partition based clustering and hierarchical clustering, based on the way they produce the results.

2.1 Similarity Measure

Similarity measure, which is the essential part of any clustering algorithm, a brief but concrete discussion about similarity measure is conducted in this section.

2.2 Partial Clustering

Partial clustering: - distance based k-means - density based

2.3 Hierarchical Clustering Algorithm

discuss about the basic idea of hierarchical clustering algorithm. discuss the usage of HCA show the basic single average linkage algorithm by naive code.

Algorithm 1 Primitive AHC algorithm

1: **procedure** PRIMITIVE AHC(s,d)
2: $S_{origin} \leftarrow S$
3: $n \leftarrow |S|$
4: $den \leftarrow []$
5: $size[x] \leftarrow 1$, for all $x \in S$
6: **for** $i \leftarrow 0, \ldots, n-2$ **do**
7: $(I,J) = argmin_{(SxS)\setminus\Delta} d$
8: append (I,J) to den
9: $S \leftarrow S \setminus \{I,J\}$
10: Create a new label L, $L \notin S \cup S_{origin}$
11: Update the matrix containing the distances
$d[L,x] = d[x,L] = \text{FORMULA}(d[I,x], d[b,J],$
$d[I,J], size[I], size[J])$, for all $x \in S$
12: $size[L] \leftarrow size[I] + size[J]$
13: $S \leftarrow S \cup \{L\}$
14: **end for**
15: **return** den
16: **end procedure**

The FORMULA is the updating formula used for the chosen linkage, while d is the distance metric. Note that our notation somewhat freely uses I and J to mean either the label of the cluster or the cluster itself.

3 Approximate Hierarchical Clustering Algorithm

As discussed in the previous chapter, the time complexity of a traditional accurate hierarchical clustering algorithm is $O(n^3)$. However, the datasets size of today's application domain has dramatically increased. Conducting a traditional agglomerative hierarchical clustering on such dataset would not be proper for an expected running time.

To handle the scaling challenge, researchers found two approaches. At fist, researchers are focused on how to find faster algorithms which generate the same hierarchical tree as the original algorithm. The hardworking of the first approach came to its limit, when dataset scale becomes even lager. The later contribution turns to find a approximate hierarchical clustering algorithm, which is not consistency to but closely resembles the exact algorithms.

This chapter will mainly discuss the current research in the approximate hierarchical clustering algorithms.

3.1 Experiment and Comparison Difficulty

Before the discussion of the approximate algorithms, an brief illustration about the difficulty of experiment and comparison of such approximate hierarchical clustering algorithms.

3.2 Algorithm by Patra

Patra et al. **patra2010distance** proposed a method, l-AL for AHC to deal with large dadtasets problem with average linkage. In their work, a set of leaders are proposed to represent the whole datasets, which are then applied the standard average link method. The advantage is that this method works for any distance metric, and reduces the running complexity as it is not requested to store the whole dataset in to memory, only the leaders are retained.

To perform a l-AL algorithm, firstly, a set of leaders should be chose, and a standard average linkage method will be implement inside each group. The average linkage method has been discussed in the previous chapter. The focus below will be mainly about the way they used

to choose leaders from the origin dataset.

Algorithm 2 Leaders Selecting algorithm

```
1: procedure LEADERSELECT(dataset, τ)
2:     leaders ← emptyhashmapofnodeandsetofnodes
3:     for i ← dataset do
4:         IF exists a l in leaders, and ||l − i|| < τ
5:             put current node i into the set in leaders with key i.
6:         ELSE
7:             make i a new leader, put an empty set into leaders with key i
8:     end for
9:     return leaders
10: end procedure
```

This algorithm takes two input, one is the whole dataset which is used in the hierarchical clustering, the other one is a tolerante value τ. The return value is the chosen leaders with their attached nodes.

3.3 Algorithm by Gilpin

3.2 Algorithm by Gilpin et. al

3.4 Twister Tries Approach

3.3 Twister tries + the extension I worked on later

3.5 Comparison

3.4 Some comparison (this can also be done directly in the section, but it might be easier to do it separatly)

Appendix: the article

I will here assume that you know the basics of using the LaTeX system. The original LaTeX

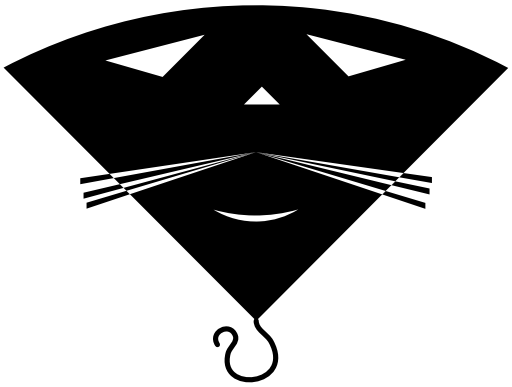

Figure 1. The cover picture of my Finnish-language LaTeX guide (**kaijanaho03:_latex_ams_latex**) is here merely as an example of how to include a picture in a thesis.

book (**lamport94:_latex**) is the official manual. There are also a lot of books in English about using LaTeX. I have also written one in Finnish (**kaijanaho03:_latex_ams_latex**).[1] A good English guide, freely available on the Internet, is *The Not So Short Introduction to LaTeX 2ε* (**oetiker:_not_so_short_introd_latex**). Remember to read the LaTeX source ode of this sample, not just the typeset version (eg. PDF).

Please note that the instructions given in this sample are by no means official. Always follow your supervisor's instructions even if they conflict with what this sample says.

[1] Many TeX and LaTeX books use a cat figure in their cover. The cover picture of my own book was rather abstract; see Figure 1.

4 The structure of the thesis

There should be 5–9 numbered chapters in a thesis, including Introduction and Conclusion. If necessary, you can use sections and subsections to give the thesis a more fine-grained structure.

The chapters that lie between Introduction and Conclusion are sometimes collectively called the *body* of the thesis. It is often said to start with a *theoretical part*, which is then followed either a *main theorem*, a *constructive part* or a *empirical part*.

4.1 The theoretical part

The goal of the theoretical part of a thesis is to develop the theoretical background required in the thesis. The idea is that a reader of the thesis should, based on just the thesis itself, be able to understand all the special concepts and methods used in the thesis. A good thesis also gives well-argued reasons for why exactly these concepts and methods are in use in the thesis (with the main alternatives given in the literature mentioned).

The best way to present and use the theoretical bakcground depends on what the thesis is like. The theoretical part of a mathematico-theoretical work differs considerably fron the theoretical part of a constructive software development work; quite different from both is the theoretical part of a quantitative or qualitative empirical study that is based on the traditions of the behavioral or the social sciences. Reading other theses of the same type, as well as similar published research reports, will give you a good impression of what is required of your own thesis.

4.2 After the theory

The theoretical part is followed by your contribution:

- In a mathematico-theoretical thesis it is usually a sequence of definitions and lemmas of your own devising, which then culminate in the proof of your main theorem.

- In a constructive thesis it is usually a computer program or other artefact that you have made yourself.
- In an empirical thesis it is a set of empirical results obtained by applying a empirical research method.

You should present your contribution with precision, giving reasons for the choices you have made. You should follow the best practices of the research tradition you are using.

5 Using the literature

The theoretical part is almost always based solely on the literature. When discussing your contribution, you may also need to cite the literature.

Remember to avoid plagiarism. If you copy, either verbatim or with slight changes (or, example, in your own translation) text from some source, make it clear to the reader. Mark your quotes (using quotation marks or some other clear manner) and give a precise citation. If you do not quote verbatim, mark any changes you have made. In most situations, however, it is better to use your own words, based on more than one source. Even then, give clear citations.

The gradu3 document class automatically uses the BIBLATEX system (**biblatex-manual**) and it Chicago style (**biblatex-chicago-manual**). You can switch off this automation by using the \documentclass-option manualbib, but that means you have to take care of the bibliography yourself, and the techniques discussed here may not be available. Please note that the Department recommends using a Chicago style for your bibliography.

5.1 Citations

You can cite sources in two ways. First, you can use the citation as a noun: **aho-compilers** briefly discuss the use of graph coloring in the register allocation phase of a compiler. In this case, use the \textcite command. Second, you can use a citation as a parenthetical, which is not read aloud: Graph coloring is one possibile way to allocate registers (**aho-compilers**). Use the \parencite command for this.

Both commands (\textcite and \parencite) take three parameters, two of which are optional. The first (optional) parameter is a pre-note, the second (optional) parameter is a post-note, and the third (mandatory) parameter is the citation key (**biblatex-manual**). The citation in the preceding sentence was made using the following command:

```
\parencite[see][Section~3.7]{biblatex-manual}
```

If you give these commands just one optional argument (that is, one enclosed in square brackets), it will be interpreted as a post-note. If you want to give only a pre-note, leave the post-note empty (**biblatex-manual**):

```
\parencite[see][]{biblatex-manual}
```

It is also possible to cite multiple sources in the same citation (**biblatex-manualbiblatex-chicago-manua**
Use the command \parencites for this. For each citation, give it the same parameters as you would give a single \parencite command. It is good practice (but often not necessary) to end the command in a \relax, so that no surprises ensue.

```
\parencites%
  [see][Section~3.7]{biblatex-manual}%
  [regarding citations in general, see also][Section~5.3.2]%
    {biblatex-chicago-manual}%
\relax.
```

If you break the command into multiple lines, use the comment sign to end each line, to prevent spurious spaces.

5.2 The bibliography database

You should add all the sources you want to cite in a separate bibliography database written on the BIBTEX format. You can use many bibliographical tools in creating and maintaining it, but it is perfectly possible to write it by hand. The name of your bibliography database must be given as an argument to the \addbibresource command.

The database in BIBTEX format is a text file following special formatting rules. It consists of records, each of which starts with an @ sign, which is then followed by the type of the record. The rest of the record goes inside curly braces. For example, the compilers book cited earlier (**aho-compilers**) can be represented as the following record:

```
@Book{aho-compilers,
  author =       {Alfred V. Aho and Monica S. Lam and Ravi Sethi and
                  Jeffrey D. Ullman},
  title =        {Compilers},
```

```
  subtitle =     {Principles, Techniques, \& Tools},
  publisher =    {Pearson Addison Wesley},
  year =         2007,
  address =      {Boston},
  edition =      2
}
```

The type of this record is “book”. The first word inside the curly braces is the citation key, which is used in the \textcite and \parencite commands. It is followed by a comma and a set of named fields like “author”, “title”, “subtitle” and “publisher”. The content of the field is written inside curly braces, although numerical data can be written without them.

The names of the authors are written mainly in the conventional way. An alternative is to invert it, giving the surname first, followed by a comma and the first name (“Aho, Alfred V.”), and in some special cases this is mandatory.[1] If there are multiple authors, their names must be separated by an “and”. If you do not list all authors, put “and others” after the last listed name.

If the author of some source is an organization, its name must be written within another set of curly braces (**unicode620**):

```
@Book{unicode620,
  author =       {{Unicode Consortium}},
  title =        {The Unicode Standard, Version 6.2.0},
  year =         {2012},
  url =          {http://www.unicode.org/versions/Unicode6.2.0/},
  urldate =      {2013-01-29}
}
```

If a source, for some reson, has no named author, leave the “author” field out ntirely. In that case, the citation uses the source’s title (**presidential-novel**):

```
@Book{presidential-novel,
  title =        {O},
  subtitle =     {A Presidential Novel},
```

[1] For example, if the author has a double surname without a hyphen separating them; as one example, the name of Simon Peyton Jones should be written in the database as “Peyton Jones, Simon”.

```
  publisher =     {Simon \& Schuster},
  year =          {2011},
}
```

A journal article (**strachey-fundamentals**) is given a record like the following:

```
@Article{strachey-fundamentals,
  author =        {Christopher Strachey},
  title =         {Fundamental Concepts in Programming Languages},
  journal =       {Higher-Order and Symbolic Computation},
  year =          2000,
  volume =        13,
  number =        {1--2},
  pages =         {11--49},
  doi =           {10.1023/A:1010000313106}
}
```

Note especially the field “doi”, in which you can write the Digital Object Idenifier (DOI) of the article. It is usually a better choice than any URL, as the DOI is a permanent identifier for the article. Most DOIs are also convertible to URLs by prepending `http://dx.doi.org/`.

If the DOI of an online source is not known (or there is none at all), you can use the “url” field. In that case, you should also give the date on which you read the source, in the field “urldate” (using the international standard format YYYY–MM–DD). You should choose the address with great care, so that it is as precise as possible and remains valid as long as possible. If the page has a specially indicated permanent link (or permalink), use it.

When citing a WWW page that is not a book or an article or any other formal publication, you can use the “online” record type (**debian-social-contract**):

```
@Online{debian-social-contract,
  title =         {Debian Social Contract},
  year =          {2004},
  url =           {http://www.debian.org/social_contract.en.html},
  urldate =       {2013-01-29}
}
```

Some sources are edited collections of independent articles. In that case, you should generally cite a specific article in it (**prechelt-credibility**) instead of the full collection. Even then, you should add both the collection and the cited article as their own records, and use a "crossref" field in the article record to refer to the collection:[2]

```
@Collection{making-software,
  editor =       {Andy Oram and Greg Wilson},
  title =        {Making Software},
  subtitle =     {What Really Works, and Why We Believe It},
  publisher =    {O'Reilly},
  year =         2011
}
@InCollection{prechelt-credibility,
  author =       {Lutz Prechelt and Marian Petre},
  title =        {Credibility, or Why Should I Insist on Being
                  Convinced},
  crossref =     {making-software},
  pages =        {17--34}
}
```

Note that a collection has an "editor" instead of an "author".

For more information about the structure of a bibliography databasem see the BIBTEX manual (**bibtexing**), the BIBLATEX manual (**biblatex-manual**), and the BIBLATEX-Chicago manual (**biblatex-chicago-manual**). There are also more examples in the source code of this document.

5.3 The bibliography

The bibliography database is converted into the bibliography by using the utility program biber. It is fairly new, and is often missing from machines whose TEX installation is not up to date. Of the ssh-accessible Linux servers of the University, only charra.it.jyu.fi has it at

[2]It is permissible to combine the article and the collection into one InCollection record, for example if one cites only one article in the collection. In that case, the title of the collection goes in a "booktitle" field, and no "crossref" field is used.

this time. It is installable in Ubuntu since version 12.10 (Quantal Quetzal) and in Debian since version 7 (Wheezy). For Windows, use the 32-bit MikTEX package miktex-biber-bin.[3]

On the command line, biber is simple to use. Once LATEX(or pdfLATEX) has been run once, invoke biber with the document name (without the .tex part) as its argument. After that, run LATEX (or pdfLATEX) at least once, until the latest run does not request another run. For example:

```
$ pdflatex malliopas
[...]
Package biblatex Warning: Please (re)run Biber on the file:
(biblatex)                malliopas
(biblatex)                and rerun LaTeX afterwards.
[..]
Output written on malliopas.pdf (18 pages, 96855 bytes).
Transcript written on malliopas.log.
$ biber malliopas
INFO - This is Biber 0.9.9
[...]
INFO - Output to malliopas.bbl
$ pdflatex malliopas
[...]
LaTeX Warning: Label(s) may have changed. Rerun to get cross-references right.
[...]
Output written on malliopas.pdf (21 pages, 107373 bytes).
Transcript written on malliopas.log.
$ pdflatex malliopas
[...]
Output written on malliopas.pdf (21 pages, 107509 bytes).
Transcript written on malliopas.log.
```

5.4 Known problems

The BIBLATEX version 2.6 (released April 30, 2013) has a bug causing the following error message:

[3]Last I looked, there was no 64-bit package of biber for MikTEX.

```
Runaway argument?
{bibliography = {{Kirjallisuusluettelo}{Kirjallisuus}}, references = \ETC.
! Paragraph ended before \DeclareBibliographyStrings was complete.
```

This bug was fixed in the following version, 2.7 (released July 7, 2013). If upgrading is not an option, there is a simple fix. Look in the file `.../biblatex/lbx/finnish.lbx` for the line

```
editorsan        = {{toimittaneet ja selityksin varustaneet,% FIXME: unsure
```

Edit the line to look like this:

```
editorsan        = {{toimittaneet ja selityksin varustaneet}% FIXME: unsure
```

(Replace the comma with a closing curly brace.)

6 Special properties of the document class

Generally, gradu3 behaves like the report document class that is shipped with LaTeX. There are, however, some differences:

- You do not need to load the packages inputenc, fontenc, and babel.
  - You must indicate the character set you are using by giving it as an option to the \documentclass command. Nowadays utf8 is generally a good choice, although some situations may require using latin1 or latin9.
  - If your thesis is written in English, indicate this using the option english to the \documentclass command. (The default is Finnish.)
- If you are writing a Bachelor's Thesis, use the option bachelor to the \documentclass command.
- Specify the metadata of your thesis using the commands given in Table 1. They must be given before the \maketitle command.
- If you want, you can write a preface after the \maketitle command. Use the \preface to start it.
- After the preface, if any, you may write a list of terms by using the thetermlist environment. Inside it, you can use the \item[*term*] command to indicate which term you are defining.
- After \maketitle, preface (if any), and term list (if any), use the \mainmatter command. It will automatically generate the tables of contents, figures, and tables that are needful.
- The commands \subsubsection, \paragraph ja \subparagraph are not supported.
- Appendices are not \chapters, they are \sections.
- The peceding chapter discussed how to cite sources and geneate a bibliography.

Command	Meaning
\title	The title of the thesis (do not use the \thanks command)
\translatedtitle	The Finnish title of an English-language thesis, the English title of a Finnish-language thesis
\studyline	Study line (optional if using the bachelor option)
\tiivistelma	Abstract in Finnish
\abstract	Abstract in English
\avainsanat	Keywords in Finnish
\keywords	Keywords in English
\author	Author's name (if multiple authors, give each their own command – the \and command is not supported)
\contactinformation	The contact information of the author
\supervisor	The supervisor of the thesis (if multiple supervisor, give each their own command; optional if using the bachelor option)

Table 1. Commands for declaring metadata

7 Conclusion

The last chapter of a thesis is the Conclusion (some authors use Conculsions, instead). Keep it short, and discuss what one can conclude about the thesis statement or research question given in the Introduction, in light of all that has been written in the thesis. The Conclusion is also the place to discuss any limitations and weaknesses of the thesis (especially those that cast doubt on the reliabliity of the results given in the thesis), if they have not been already discussed, for example in a Discussion chapter. It is also customary to state, what further research might be beneficial in light of this thesis.

If the Conclusion threatens to become too long, it is a good idea to split the interpretation of the results into its own chapter, often called Discussion, making Conclusion short and sweet.

After Conclusion, there is the bibliography, indicated by the \printbibliography command, followed by appendices, if any.

Appendices

A Moving from gradu2 to gradu3

Moving an incomplete thesis from gradu2 to gradu4 is not particularly difficult. The first thing to do is to change gradu2 into gradu3 in the \documentclass command. Most of the options given to it must be removed, as they are not supported. A "kandi" option is changed into "bachelor"; any "english" option is retained, and so is "utf8", "latin1", or "latin9".

Table 2 lists the command name changes that are needed. A dash indicates that there is no corresponding command. Note especially the new commands.

gradu2	gradu3
—	\maketitle
—	\supervisor
\acmccs	—
\aine	\subject
\copyrightowner	—
\fulltitle	—
\laitos	\department
\license	—
\linja	\studyline
\paikka	—
\setauthor	\author
\termlist	thetermlist environment
\tyyppi	\type
\yhteystiedot	\contactinformation
\yliopisto	\university
\ysa	—

Table 2. Command changes from gradu2 to gradu3

The most effort is likely needed to converting citations and the bibliography.

B Rarely needed features

In addition to features already mentioned, gradu3 offers the following additional features:

- The standard options “draft” and “final” work.
- The option “finnish” works (but is not needed, as it is the default).
- You can change the University of the thesis by using the \university command.
- You can change the Department of the thesis by using the \department command.
- You can change the formal subject of the thesis by using the \subject command. In English theses, the subject should be prefixed by “in” (for example, “in Information Technology”); in Finnish theses, use a capital initial letter and the genitive form (“Tietotekniikan”).
- You can change the type of the thesis by using the \type command.
- You can set the date of the thesis by using the \setdate command. Give it three parameters (day of month, month, and year) in numerical form.
- The chapterquote environment can be used to give an epigraph to a chapter. There is one mandatory parameter (the attribution of the epigraph).
- The command \graduclsdate prints the release date of the current version of gradu3, and the command \graduclsversion prints its version number.