

TSBK08 - Datakompression

Entropi-estimering, Huffman-kodning och LZW-kodning

Fredrik Wallström

Handledare : Harald Nautsch
Examinator : Harald Nautsch

Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår. Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art. Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart. För ytterligare information om Linköping University Electronic Press se förlagets hemsida <http://www.ep.liu.se/>.

Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances. The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility. According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement. For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: <http://www.ep.liu.se/>.

Sammanfattning

Abstract.tex

Författarens tack

Framför ett tack till Harald Nautsch för en väl genomförd kurs,
TSBK08 – Datakompression.

Innehåll

Sammanfattning	iii
Författarens tack	iv
Innehåll	v
1 Introduction	1
1.1 Motivation	1
1.2 Mål	1
1.3 Frågeställningar	1
1.4 Avgränsningar	1
2 Teori	2
3 Method	4
4 Results	5
5 Discussion	6
5.1 Results	6
5.2 Method	6
5.3 Källkritik	7
6 Conclusion	8



1 Introduction

1.1 Motivation

Att komprimera filer är väldigt användbart i dagens samhälle. Dels för att spara filer i ett mindre format, vilket leder till de tar upp mindre plats på lagringsenheten, och för att det går snabbare att överföra informationen mellan olika användare. Det finns flertalet metoder för att komprimera filer och olika metoder passar bättre till vissa filtyper. Den här rapporten kommer att undersöka två olika komprimeringsmetoder, Huffman-kodning och Lempel-Ziv-Welch-kodning (LZW).

1.2 Mål

Målet med denna rapport är att undersöka två olika komprimeringsmetoder, Huffman-kodning och LZW-kodning. Hur fungerar dessa komprimeringsmetoder och vilka olika fördelar och nackdelar finns, samt hur bra de presterar vid komprimering av olika filtyper.

1.3 Frågeställningar

De frågeställningar som denna rapport kommer att behandla listas nedan.

1. Hur stor plats på lagringsenheten sparar vi då en fil komprimeras med Huffman-kodning respektive LZW-kodning.
2. Hur nära kommer vi den optimala takten när vi komprimerar en fil med Huffman-kodning.
3. Vilka filtyper passar bäst att komprimera med Huffman-kodning respektive LZW-kodning.

1.4 Avgränsningar

Det finns flera olika metoder för komprimering av filer. Denna rapport undersöker endast Huffman-kodning och LZW-kodning och tar ej hänsyn till övriga existerande metoder för komprimering.



2 Teori

Nedan presenteras hur Entropi estimering, Huffman-kodning och LZW-kodning fungerar.

Entropi estimering

Det går att mäta hur bra komprimerat en sekvens är, detta gör man genom mäta takten på koden. Takten på koden ges av [nautsch2018].

$$\text{Takt } R = \frac{\text{genomsnittligt antal bitar per kodord}}{\text{genomsnittligt antal symboler per kodord}} \quad \text{Bitar per symbol}$$

Det ger alltså ett tal som säger i snitt hur många bitar vi har kodat varje symbol med.

Entropin för en källa är en teoretisk lägre gräns för hur många bitar vi behöver för att koda varje symbol. Takten kan alltså inte vara under denna gräns. Entropin av en källa ges av [nautsch2018].

$$\text{Entropy } H(X) = - \sum p(X) \log p(X)$$

Där $p(X)$ är sannolikheten för given symbol och \log är logaritm med bas 2. Vi summerar över alla symboler.

Huffman-kodning

Huffman-kodning är en metod för att komprimera filer till ett mindre format. Metoden utvecklades av den dåvarande studenten David A. Huffman år 1952. Idén med metoden är att tilldela varje symbol som ska kodas med X antal olika bitar, 0 eller 1. Exempel kan symbolen A tilldelas bitarna 101011. I metoden kommer de symboler som förekommer flest gånger i sekvensen, det vill säga den symbol som har högst sannolikhet att förekomma, att tilldelas minst antal bitar. Det betyder också att de symboler som förekommer minst gånger i sekvensen kommer tilldelas mest antal bitar. Exempel kan symbolen A tilldelas med bitarna 10 och symbolen B tilldelas bitarna 1001010101. Detta betyder att symbolen A förekommer oftare i sekvensen än symbolen B. Alltså kommer sekvensen att komprimeras på bästa möjliga vis

eftersom A förekommer oftare än B och A:s kodord är kortare än B:s. Redan här inses att med Huffman-kodning behövs information om hur sekvensen ser ut. Det vill säga, man behöver veta sannolikheterna för de olika symbolerna i sekvensen. Nedan beskrivs mer ingående hur komprimering och avkomprimering fungerar i Huffman-kodning [huffmancoding2018].

Komprimering

Att komprimera en sekvens med Huffman-kodning förutsätter att det finns en given sannolikhetsstabell för sekvensen. Det finns alltså en lista som mappar en symbol mot dess sannolikhet, vi kallar här varje symbol och dess sannolikhet för en nod. Metoden fungerar då enligt följande:

1. Ta ut de två symbolerna (noderna) som har lägst sannolikhet i listan.
2. Addera dessa noders sannolikheter och lägg till den som en ny nod i listan, resterande noder behålls.
3. Börja sedan om från ett om det finns fler än en stycken noder kvar i listan.

Med den här metoden byggs en trädstruktur upp, som kan ses i figur blablabla, över hur de olika symbolerna kan tilldelas olika bitar. Lövnoderna i trädet representerar de olika symbolerna. Roten av trädet representeras av en nod med sannolikhet 1 eftersom att summan av alla sannolikheter i en sekvens alltid summeras till 1. För att tilldela en sekvens av bitar till alla symboler traverserar man igenom hela trädet och bildar en lista som mappar respektive symbol sin egna unika sekvens av bitar. Här är det förutbestämt att varje gång man går vänster i trädet adderar man en 0:a till sekvensen av bitar och varje gång man går höger adderar man en 1:a till sekvensen, eller vice versa. figur blabla visar hur ett Huffman-träd kan se ut [huffmancoding2018].

När mappningen mellan symbol och sekvens av bitar existerar kan den givna sekvensen komprimeras. Detta görs genom att läsa av en symbol i i taget ifrån sekvensen som ska komprimeras och sedan matcha den symbolen med symbolerna som finns i listan som mappar symboler mot en sekvens av bitar. Då ges en sekvens av bitar som adderas till den resulterande bitsekvensen. Denna procedur upprepas tills det inte finns fler symboler att läsa av i sekvensen som ska komprimeras.

Avkomprimering

För att avkomprimera en Huffman-kod behövs samma sannolikhetslista som användes vid komprimeringen av en sekvens. Förutsatt att denna lista är känd, kan ett Huffman-träd byggas upp på samma vis som vid komprimeringen. Med Huffman-trädet kan man sedan traversera ifrån roten och ner beroende på vilken bit, 0 eller 1, man stöter på i den sekvens som ska avkodas. När man sedan kommer till en lövnod betyder det att en symbol kan avkodas, denna symbol adderas sedan till den resulterande sekvensen. Proceduren startar sedan om ifrån rotnoden och avslutas då det inte längre finns bitar kvar att läsa av i bitsekvensen [huffmancoding2018].

LZW-kodning

Komprimering

Avkomprimering



3 Method

Nedan går igenom hur entropi estimering, Huffman-kodning och LZW-kodning har genomförts.

Entropi estimering

För att göra en entropi estimering av givna källor, som angavs i laborationens uppgift, skrevs ett program i C++. Programmet utgick ifrån en funktion som beräknade entropin för en given källa enligt formeln som beskrivs i teori avsnittet. Funktionen tar in en sannolikhetsstabell som mappar varje symbol mot dess sannolikhet att den förekommer i källan. Denna lista behövde alltså genereras innan entropin kunde beräknas.

För att generera en sannolikhetsstabell jag estimerade sannolikheterna för varje symbol i källan. Detta gjordes genom att först spara hela filen i en lista, för att sedan iterera över den och addera 1 i en ny lista för varje påträffad symbol i källan. Denna nya lista mappar således symboler mot hur många gånger den förekommer i källan, det vill säga, en frekvenstabell. Denna frekvenstabell användes sedan för att beräkna sannolikhets Tabellen genom att dividera varje frekvens med källans storlek. Den önskade entropin kunde sedan beräknas.

I laborationen har även betingade entropier beräknats för källorna. En betingad entropi betyder att vi beräknar entropi då vi känner till hur källan ser ut K steg tillbaka i tiden. De betingade entropierna som har beräknats är då $K = 1$ och då $K = 2$, detta uttrycks genom följande, $H(X_i | X_{i-1})$ och $H(X_i | X_{i-1}, X_{i-2})$. För att beräkna de betingade entropierna krävdes en sannolikhetsstabell för par av symboler och för tripplar av symboler. Dessa sannolikhetsstabeller togs fram på motsvarande sätt som för en symbol, med undantag att två respektive tre symboler lästes in i taget. Dessa symboler adderades sedan genom att konkatenera dess bit representation, detta skapar en unik ny symbol. Vi har då sannolikhetsstabeller för par och tripplar av symboler. Sedan beräknades den gemensamma entropin, $H(X_1, X_2)$ och $H(X_1, X_2, X_3)$ med samma entropi funktion som användes för singlar av symboler. Efter det användes kedjeregeln, $H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1})$, för att beräkna den betingade entropin.

Huffman-kodning

LZW-kodning



4 Results

This chapter presents the results. Note that the results are presented factually, striving for objectivity as far as possible. The results shall not be analyzed, discussed or evaluated. This is left for the discussion chapter.

In case the method chapter has been divided into subheadings such as pre-study, implementation and evaluation, the result chapter should have the same sub-headings. This gives a clear structure and makes the chapter easier to write.

In case results are presented from a process (e.g. an implementation process), the main decisions made during the process must be clearly presented and justified. Normally, alternative attempts, etc, have already been described in the theory chapter, making it possible to refer to it as part of the justification.



5 Discussion

This chapter contains the following sub-headings.

5.1 Results

Are there anything in the results that stand out and need be analyzed and commented on? How do the results relate to the material covered in the theory chapter? What does the theory imply about the meaning of the results? For example, what does it mean that a certain system got a certain numeric value in a usability evaluation; how good or bad is it? Is there something in the results that is unexpected based on the literature review, or is everything as one would theoretically expect?

5.2 Method

This is where the applied method is discussed and criticized. Taking a self-critical stance to the method used is an important part of the scientific approach.

A study is rarely perfect. There are almost always things one could have done differently if the study could be repeated or with extra resources. Go through the most important limitations with your method and discuss potential consequences for the results. Connect back to the method theory presented in the theory chapter. Refer explicitly to relevant sources.

The discussion shall also demonstrate an awareness of methodological concepts such as replicability, reliability, and validity. The concept of replicability has already been discussed in the Method chapter (3). Reliability is a term for whether one can expect to get the same results if a study is repeated with the same method. A study with a high degree of reliability has a large probability of leading to similar results if repeated. The concept of validity is, somewhat simplified, concerned with whether a performed measurement actually measures what one thinks is being measured. A study with a high degree of validity thus has a high level of credibility. A discussion of these concepts must be transferred to the actual context of the study.

The method discussion shall also contain a paragraph of source criticism. This is where the authors' point of view on the use and selection of sources is described.

In certain contexts it may be the case that the most relevant information for the study is not to be found in scientific literature but rather with individual software developers and open

source projects. It must then be clearly stated that efforts have been made to gain access to this information, e.g. by direct communication with developers and/or through discussion forums, etc. Efforts must also be made to indicate the lack of relevant research literature. The precise manner of such investigations must be clearly specified in a method section. The paragraph on source criticism must critically discuss these approaches.

Usually however, there are always relevant related research. If not about the actual research questions, there is certainly important information about the domain under study.

5.3 Källkritik

Wikipedia ska sagts inte användas som källor i en rapport men inom just datavetenskap är wikipedia oftast en trovärdig källa och har därför använts i denna rapport. En annan anledning beror också på att just Huffman-kodning och LZW-kodning är kända metoder för komprimering av filer och för just fakta angående metoderna tycker jag att wikipedia är en pålitlig källa.



6 Conclusion

This chapter contains a summarization of the purpose and the research questions. To what extent has the aim been achieved, and what are the answers to the research questions?

The consequences for the target audience (and possibly for researchers and practitioners) must also be described. There should be a section on future work where ideas for continued work are described. If the conclusion chapter contains such a section, the ideas described therein must be concrete and well thought through.