binary team

نظري نظم الوسائط المتعددة والفائقة | المحاضرة: 7 | د. محمد زهير صندوق | عدد الصفحات: 6

سنتابع في هذه المحاضرة في خوارزميات الضغط:

❖ خوارزمية الترميز الحسابي "Arithmetic Coding":

- تُصنَّف على أنّها طريقة للضغط بدون فقد للبيانات (Loss less).
- تعتمد على استبدال سلسلة من رموز الدخل (محارف النص) برقم وحيد float ندعوه بالرّقم السحري.
- يكون خرج هذه الخوارزمية هو الرّقم السحري (single number) وقيمته أكبر أو تساوي الصفر وأقل من الواحد ؛ أي ضمن المجال [0,1[.
- خوارزمية فعّالة لضغط النصوص حصراً.

خطوات الخوارزمية:

تُقسَم الخوارزمية على مرحلتَين:

- **في المرحلة الأولى:** يتم مسح النص المُراد ضغطه كاملاً (المرور على جميع محارفه) وإنشاء جدول التوزيع الاحتمالي المقابِل لمحارف النص وذلك من خلال الخطوات التالية:

1- نقوم بإحصاء عدد المحارف (بدون تكرار) ؛ أي إحصاء عدد أنواع المحارف في النص ، ويتم إنشاء جدول عدد أسطره يساوي عدد أنواع المحارف.

2- من أجل كل مِحرَف يتم إحصاء عدد مرّات تكراره ؛ أي frequency كل محرف ، وحساب احتمال وروده ، حيث أنّ:

$$\text{احتمال ورود مِحرَف في نص معين} = \frac{\text{عدد مرّات ورود هذا المِحرَف في النص}}{\text{عدد محارف النص}}$$

3- يتم تجزئة الفضاء الاحتمالي ، وكما نعلم فإنّ قِيَم الاحتمالات في الفضاء الاحتمالي تتراوح ضمن المجال [0,1] فيتم تجزئة هذا المجال الذي ندعوه (interval) إلى مجالاتٍ جزئيّة (sub-intervals) وذلك من أجل كل مِحرَف ، ويتم تحديد طول المجال الجزئي لمِحرَف وفقاً لاحتمال ورود ذلك المِحرَف في النص ، وبعد كل اختيار لمجال جزئي لأحد المحارف يتم اعتبار ما تبقّى هو المجال الجديد والذي سيتم تجزئته.

مثال:

ليكن لدينا النّص التالي (وهو النّص الذي تمّ اعتماده عندما نُشرت هذه الخوارزمية):

SWISS MISS

بتطبيق المرحلة الأولى من الخوارزمية نجد:

1- عدد محارف النص السابق بدون تكرار هو 5 محارف ، لذلك سنقوم بإنشاء جدول مكوَّن من 5 أسطر.

2- نبدأ مِحرَف مِحرَف وصولاً لآخر النص:

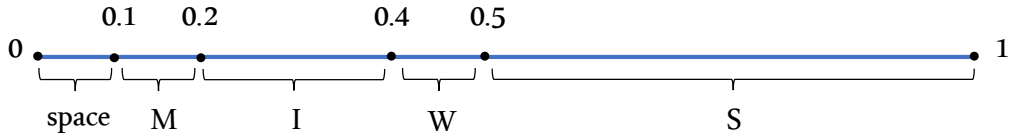
- من أجل أوّل مِحرَف S: عدد مرّات وروده هو 5 مرّات ، فيكون احتمال وروده $= \frac{5}{10} = 0.5$.
- المحرف التالي هو W: عدد مرّات وروده هو مرّة واحدة ، فيكون احتمال وروده $= \frac{1}{10} = 0.1$.

ونكمل من أجل بقيّة المحارف..

3- نقوم بتجزئة المجال [0,1] إلى مجالاتٍ جزئيّة من أجل كل مِحرَف:

- من أجل أوّل مِحرَف S: بما أنّ احتمال وروده 0.5 إذاً طول المجال المُراد اختياره سيكون 0.5 ، وباعتبار أنّه تم تجزئة المجال [0,1] إلى مجالَين [0.5,1] و [0,0.5] واختيار المجال [0.5,1] ليكون القسم الذي يحجزه المِحرَف S من الفضاء الاحتمالي.
- من أجل المِحرَف التالي W: بما أنّ احتمال وروده هو 0.1 إذاً يتم اختيار مجال طوله 0.1 ، فيتم تجزئة المجال [0,0.5] إلى مجالَين [0.4,0.5] و [0,0.4] واختيار المجال [0.4,0.5] ليكون القسم الذي يحجزه المِحرَف W من الفضاء الاحتمالي.

ونتابع من أجل بقية المحارف فنحصل على ما يلي:



ويكون جدول الفضاء الاحتمالي الذي تم إنشاؤه:

Char	Freq.	Prob.	Range
S	5	5/10=0.5	[0.5, 1.0)
W	1	1/10=0.1	[0.4, 0.5)
I	2	2/10=0.2	[0.2, 0.4)
M	1	1/10=0.1	[0.1, 0.2)
space	1	1/10=0.1	[0.0, 0.1)

ملاحظة:

لكل نص يتم إنشاء جدول توزيع احتمالي يخصّه ، وهو <u>جدول غير وحيد</u> ، إذ يمكن تغيير ترتيب المحارف فيه ، كما يمكن تغيير قِيَم حدود المجالات باختيار مجالاتٍ مختلفة (مثلاً: أن يتم البدء باختيار المجال [0.2,0.7] من أجل المِحرَف S) لكنّ قِيَم الاحتمالات تبقى ذاتها ؛ أي أنّ <u>منطقيّة الجدول وحيدة</u> للنص نفسه.

~~~~~~~~~~~~~~~~

- **في المرحلة الثانية**: يتم مسح محارف النص من جديد وبالاعتماد على جدول التوزيع الاحتمالي الناتج عن الخطوة الأولى يتم تطبيق خوارزمية الترميز التالية:

```
low  = 0.0;
high = 1;
while ( ( c = getc( input ) ) != EOF ) {
    range = high - low;
    high = low + range * high_range( c );
    low =  low + range * low_range( c );
}
output ( low );
```

توضيح خطوات الخوارزمية:

1- تبدأ هذه الخوارزمية بتعريف متحولَين وإسناد قيمة ابتدائيّة لكلٍّ منهما: $low = 0,\ high = 1$.

2- من أجل كل مِحرَف من محارف النص يتم الدخول بحلقة $while$ وحساب ما يلي:

- يتم إيجاد طول المجال المُراد تجزئته: $range = high - low$.
- يتم تعديل قيمة $high$ من أجل الاستفادة منها في التكرار التالي ، وذلك بتطبيق المعادلة:

$$high = low + range * high\_range(c)$$

حيث أنّ $high\_range(c)$ يمثّل الحدّ الأعلى للمجال الذي يقوم بحجزه الحرف c ؛ أي سيتم العودة عن طريق هذا التابع إلى جدول التوزيع الاحتمالي لإحضار هذه القيمة.

- يتم تعديل قيمة low من أجل الاستفادة منها في التكرار التالي ، وذلك بتطبيق المعادلة:

$$low = low + range * low\_range(c)$$

حيث أنّ $low\_range(c)$ يمثّل الحدّ الأدنى للمجال الذي يقوم بحجزه الحرف c.

- وبالوصول لنهاية النص تنتهي حلقة $while$ ، ويتم تنظيم جدول عدد أسطره بعدد تكرارات الحلقة ليتم وضع القيم السابقة ضمنه.

3- يكون خرج الخوارزمية هو قيمة low لآخر تكرار للحلقة ؛ أي لآخر سطر في الجدول ، وندعو هذه القيمة بالرّقم السحري.

نتابع على المثال السابق:

- من أجل المِحرَف الأول S:
  - حدود المجال لهذا المِحرَف: $high\_range(S) = 1,\ low\_range(S) = 0.5$.
  - طول المجال: $range = 1 - 0 = 1$ (لأنّه التكرار الأول للحلقة ، فتكون قِيَم $low = 0,\ high = 1$)
~~~~~~~~~~~~~~~~

- تعديل قيمة كل من $low, high$:

$$high = low + range * high\_range(S) = 0 + 1 * 1 = 1$$
$$low = low + range * low\_range(S) = 0 + 1 * 0.5 = 0.5$$

- من أجل المِحرَف التالي W يتم الدخول بالتكرار التالي للحلقة:
  - حدود المجال لهذا المحرف: $high\_range(W) = 0.5,\ low\_range(W) = 0.4$.
  - طول المجال: $range = 1 - 0.5 = 0.5$ (لأنّه يتم الاعتماد على قِيَم $low, high$ من التكرار السابق)
  - تعديل قيمة كل من $low, high$:

$$high = low + range * high\_range(W) = 0.5 + 0.5 * 0.5 = 0.75$$
$$low = low + range * low\_range(W) = 0.5 + 0.5 * 0.4 = 0.7$$

ويتم المتابعة بنفس الطريقة للوصول لنهاية محارف النص ، فينشأ لدينا الجدول التالي:

C	Low_range(c)	High_range(c)	range	L	h
S	0.5	1	1	0.5	1
W	0.4	0.5	0.5	0.7	0.75
I	0.2	0.4	0.05	0.71	0.72
S	0.5	1	0.01	0.715	0.72
S	0.5	1	0.005	0.7175	0.72
	0	0.1	0.0025	0.7175	0.71775
M	0.1	0.2	0.00025	0.717525	0.71755
I	0.2	0.4	2.5E-05	0.71753	0.717535
S	0.5	1	5E-06	0.7175325	0.717535
S	0.5	1	2.5E-06	0.71753375	0.717535

نجد أنّ قيمة الرقم السحري في السطر الأخير من الجدول والتي تمثّل قيمة المتحوّل low هي $low = 0.71753375$.

وبذلك يتم ضغط النص من خلال تخزين جدول التوزيع الاحتمالي بالإضافة لقيمة الرّقم السحري بدلاً من تخزين النص.

ملاحظة:

نجد أنّه من أجل ضغط المثال السابق والاستعاضة عنه بتخزين جدول التوزيع الاحتمالي وقيمة الرّقم السحري فإنّ ذلك يزيد الحجم بدلاً من القيام بعمليّة ضغط! إلّا أنّ هذه الخوارزمية فعّالة جداً لضغط النصوص كبيرة الحجم ، وجدول التوزيع الاحتمالي مهما كان النّص كبيراً سيكون محدوداً لأنّه يخزّن معلومات عن نوع كل محرف ضمن النص.

وعدد أنواع المحارف الأعظمي (عدد المحارف بدون تكرار) فيما لو كان النّص مرمّزاً باستخدام جدول ASCII هو 256 مِحرَفاً ، فيكون العدد الأعظمي لأسطر جدول التوزيع الاحتمالي هو 256 سطراً ، وهذا الحجم ليس كبيراً مقارنةً بحجم نص مؤلَّف من عدّة صفحات.

تذكرة بترميز Unicode:

يتم الترميز على 2byte وبالتالي فإنّ عدد المحارف المرمَّزة في جدول الترميز $65536 = 2^{16}$ وهي مقسَّمة لِـ 23 ألف لترميز مقاطع اللغة الصينية (فهي تحوي على 23 ألف مقطع صوتي) و 3000 من أجل ترميز محارف لغات أخرى (اللغة العربية ، الفرنسية ، ...) وما تبقّى يُستخدم لترميز الرسوم والرموز الرّياضية وغيرها.

إذاً بالاعتماد على ترميز Unicode فإنّ ذلك سيجعل عدد أسطر جدول التوزيع الاحتمالي الأعظمي هو 65536 سطراً ، ومع ذلك يبقى حجم هذا الجدول محدوداً مقارنةً بحجم النص وتبقى خوارزمية الضغط فعّالة.

~~~~~~~~~~~~~~~

ملاحظة خارجية:

السّبب في كَون عدد ترميز محارف Unicode كبيراً من أجل اللغة الصينية هو أنّ هذه اللغة هي لغة مقاطع ، فيتم الترميز على مستوى المقطع وليس على مستوى الحرف الواحد كما في لغتنا.

مثلاً لتخزين النص التالي: "ماو تسي تونغ" فإنّه يُخزَّن على 12 مِحرَف باللغة العربيّة لأنّه يتم ترميز مِحرَف مِحرَف ، أمّا في اللغة الصينية يتم الترميز على مقاطع والنص السابق مكوَّن من ثلاثة مقاطع صوتيّة (ماو ، تسي ، تونغ).

لكن لو تمّ تغيير المقطع "تسي" للمقطع "تسا" فإنّ ذلك يُعتبر رمزاً جديداً في اللغة الصينية وذلك خلافاً للغة العربية.

~~~~~~~~~~~~~~~

عملية فكّ الضغط المطبَّق باستخدام خوارزمية Arithmetic coding (The Decoder):

يتم بالطريقة المُعاكِسة وذلك بالاعتماد على الرّقم السحري وجدول التوزيع الاحتمالي وبتطبيق خوارزمية فكّ الترميز التالية:

```
number = input_code();
for ( ; ; ) {
  symbol = find_symbol_in_range( number );
  putc( symbol );
  range= high_range(symbol)- low_range(symbol);
  number = (number - low_range(symbol))/ range;
}
```

توضيح خطوات هذه الخوارزمية:

1- تبدأ هذه الخوارزمية بتعريف المتحوّل $number$ والذي يتم تهيئته بالرّقم السحري المُخزَّن في عملية ضغط النص.

2- يتم الدخول في حلقة for ، وفي كل تكرار يتم إيجاد ما يلي:

- نوجد المِحرَف المقابِل للرقم $number$ ؛ أي المِحرَف الذي يحوي مجاله على قيمة $number$ ، وذلك بالعودة لجدول التوزيع الاحتمالي من خلال التابع $find\_symbol\_in\_range(number)$.
- نقوم بحساب طول المجال للمِحرَف الذي تم إيجاده ؛ أي حساب قيمة $range$.
- نقوم بتعديل قيمة المتحول $number$ ؛ أي حساب القيمة الجديدة له لاستخدامها في التكرار التالي وذلك وفقاً للمعادلة:

$$number = \frac{number - low\_range(symbol)}{range}$$

- تتوقّف الحلقة عندما يصبح $number = 0$ ، وتكون المحارف النّاتجة هي محارف النص قبل الترميز.

بالعودة إلى المثال السابق وتطبيق خوارزمية فك الترميز:

1- $number = 0.71753375$.

2- نبحث عن المِحرَف الذي ينتمي لمجاله الرّقم السابق من جدول التوزيع الاحتمالي فنجد أنّه المِحرَف S لأنّ $0.7 \in [0.5,1]$، وطول مجاله $range = 0.5$، فنقوم بتسجيل المِحرَف S، ونوجد الرقم الجديد:

$$number = \frac{0.71753375 - 0.5}{0.5} = 0.4350675$$

نكرّر العملية السابقة ونبحث عن المِحرَف الذي يحوي مجاله على قيمة number فنجد أنّه المحرف W لأنّ $0.4 \in [0.4,0.5]$ فنقوم بتسجيل المِحرَف W ونوجد الرقم الجديد كما سبق ، ونكمل حتى يصبح number = 0 ونحصل على الجدول التالي:

input	output	range	number
0,71753375	S	0,5	0,4350675
0,4350675	W	0,1	0,350675
0,350675	I	0,2	0,753375
0,753375	S	0,5	0,50675
0,50675	S	0,5	0,0135
0,0135	space	0,1	0,135
0,135	M	0,1	0,35
0,35	I	0,2	0,75
0,75	S	0,5	0,5
0,5	S	0,5	0

نلاحظ أنّنا حصلنا على النص الأصلي من عمود output ، فتتم بذلك عمليّة فكّ الضغط.

ملاحظات:

- تعاني خوارزمية الضغط المذكورة من مشكلة التدوير للأرقام الناتجة ، كما أنّها لا تطبَّق بشكلٍ كبير لأنّها تضغط النّصوص فقط وقلّما يحدث الضغط لنصٍّ لوحده.
- يوجد مثال آخر لتطبيق خوارزمية الضغط وفكّه على كلمة Arithmetic في الصفحات 12, 13, 14 من الملف الثامن.
- تُعتبر هذه الخوارزمية سؤال أساسي في الامتحان ، فقد يرِد السؤال بأن يُعطى رقم ما مع جدول التوزيع الاحتمالي ويُطلب إيجاد النص (فكّ الترميز) ، أو أن يُعطى النّص ويُطلب إيجاد الرّقم السحري الناتج عن عملية الضغط.

ورد خطأ في المحاضرة 2 – صفحة 4 – ضمن الملاحظة (عندما يتم شراء صبغة فيتم بذلك الحصول على المركبتَين H, S ...) والتصحيح (فيتم بذلك الحصول على المركبتَين H, L).

Word press and preparation:

Salem Kabbani

Reviewed by:

Naya Hafez

انتهى المقرر