ISSN 2170-0796

ضغط وتشفير البيانات بإستخدام ترميز هافمان

سليمان يعقوب الفرا (Sliman Jakub El-Fara)
توماش ادريانوفسكي (Tomasz Adrianowski)
كارول جوريتسكي (Karol Górecki)
سيباستيان فلوريك (Sebastian Florek)
Lodz University of Technology, Poland
بولندا

ملخص

ترميز هافمان يعتبر واحداً من أبسط الطرق لضغط المعلومات بمختلف أنواعها دون أن يؤدي ذلك إلى أي فقدان للبيانات المضغوطة. وبرغم بساطته فإنه يستخدم في العديد من أنظمة الضغط بكلا نوعيه الفاقد وغير الفاقد للبيانات كمرحلة من مراحل الضغط. من الأمثلة على استخدامه: نظام ضغط المعلومات الصوتية (MP3) وكذلك نظام ضغط الصور (JPEG).

وقد توصل العلماء مؤخراً إلى أن ذات الترميز يمكن أن يستخدم لتشفير المعلومات أيضاً، مما زاد من اهتمام متخصصي الحماية والتشفير به في الآونة الأخيرة وبدءهم باستخدامه في أنظمة المقاييس الحيوية، حيث يستخدم المقياس الحيوي كبصمة الإصبع في إنتاج مفتاح تشفير يستخدم فيما بعد في عملية التشفير. وسنحاول في هذا المقال العلمي دراسة كلا العمليتين: الضغط والتشفير في ترميز هافمان بشيءٍ من التفصيل.

1. تعريف

ترميز هافمان (بالإنجليزية: Huffman coding) هو ترميز متغير الطول يستخدم في تشفير وضغط البيانات، يتناسب فيه طول ترميز كل رمز (حرف) مع احتمال ظهوره.

يعتبر ترميز هافمان من أنواع الضغط غير الفاقد للبيانات ويمكن أن يستخدم أيضاً كمرحلة تشفير أخيرة في طرق الضغط المختلفة بكلا نوعيها الفاقدة (MP3 وJPEG) وغير الفاقدة للبيانات.

2. ضغط البيانات باستخدام ترميز هافمان

لنفترض انه لدينا معلومة نصية تم تسجيلها باستخدام عدد ن من الحروف (الرموز). بناءاً على هذه المعلومة يمكننا حساب احتمال ظهور كل حرف من هذه الحروف في المعلومة النصية (كلما تكرر ظهور الحرف في المعلومة النصية، كلما كان احتماله أكبر).احتمال ظهور الحرف = عدد مرات ظهور الحرف في المعلومة النصية / العدد الكلي للحروف في المعلومة النصية .عند تحويل المعلومة النصية إلى معلومة رقمية (ثنائية) يتم تعيين عدد ثنائي معين لكل حرف (رمز) وارد في المعلومة النصية.تشفير هافمان يعتمد على تمثيل كل حرف من الحروف الواردة في المعلومة النصية على شكل عدد ثنائي مسجل باستخدام عدد من الخانات الثنائية (بت) بحيث يتناسب عدد الخانات المستخدمة لتسجيل الرقم الثنائي عكسياً مع احتمال ظهور الحرف.تطبيق هذه القاعدة يؤدي إلى تقليل عدد الخانات الثنائية المستخدمة لتسجيل المعلومة الرقمية وبالتالي تقليل حجمها (ضغطها) حيث أننا نقوم بتسجيل الحروف المكررة بشكلٍ كبير في النص باستخدام أصغر الأرقام الثنائية (0، 1، 01، 10، 11 ...) في حال أن الحروف الأندر استخداماً في النص تسجل على شكل أرقام ثنائية أكبر، مما يجعل الحجم الكلي للمعلومة الرقمية أصغر مما لو تم تسجيلها باستخدام ترميز ثابت الطول (تخصيص عدد ثابت من الخانات لكل حرف).

ISSN 2170-0796

حتى تتضح عملية الضغط التي تحصل في ترميز هافمان دعنا ننظر إلى المثال التالي: لنفترض أننا نريد ضغط المعلومة النصية التي تتكون من كلمة واحدة (**الاستعلامات**) باستخدام ترميز هافمان.

الخطوة الأولى: تحديد عدد الحروف المستخدمة لتسجيل المعلومة النصية.في المثال السابق تتكون كلمة **الاستعلامات** من الأحرف التالية: ر = { ا، ل، ت، س، ع، م }

الخطوة الثانية: تحديد احتمال ظهور كل حرف من الحروف في المعلومة النصية. نقوم بحساب تكرار كل حرف من الحروف الواردة في النص، وبناءاً عليه نحسب احتمال ظهور حروف المعلومة النصية ونحصل على الجدول الموضح أدناه:

الحرف	ا	ل	ت	س	ع	م
عدد مرات التكرار	4	2	2	1	1	1
احتمال الظهور	0,36	0,18	0,18	0,09	0,09	0,09

جدول 1: يوضح احتمال ظهور كل حرف من الحروف المستخدمة لتسجيل كلمة: الاستعلامات

من الجدول السابق نلاحظ أن الأحرف أ، ل، ت من الأحرف الأكثر استخداماً في المعلومة النصية وبالتالي الأجزاء الرقمية التي ستمثل هذه الحروف في المعلومة الرقمية ستكون الأصغر من حيث الطول.

الخطوة الثالثة: تحديد الأجزاء الرقمية التي ستمثل كل حرف من حروف المعلومة النصية. يتم إجراء هذه الخطوة بحيث يتناسب عدد خانات العدد الثنائي الممثل للحرف عكسياً مع احتمال ظهور الحرف. ما يعني أنه كلما كان احتمال ظهور الحرف أكبر، كلما استخدمنا عدد أقل من الخانات الثنائية لتسجيل العدد الثنائي الذي سيمثله.لإجراء الخطوة الثالثة سنقوم ببناء شجرة هافمان، حيث سنقوم بناءاً عليها بتحديد العدد الثنائي الذي يمثل كل حرف من حروف المعلومة الرقمية.

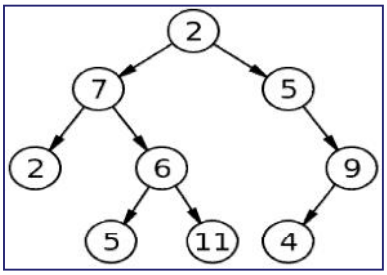
شجرة هافمان تعتبر مثالاً على الأشجار الثنائية. وسنقوم بالتعرف على مفهوم الشجرة الثنائية قبل الانتقال إلى بناء شجرة هافمان.

الشجرة الثنائية هي شجرة بنية معلومات تتكون من عدد من الرؤوس. لكل رأس في الشجرة الثنائية رأسين من الأبناء على الأكثر يتم تمييزهم بـ"الابن الأيسر" و"الابن الأيمن". يوجد في الشجرة ما يسمى بـ"الجذر"، وهو سلف كل الرؤوس، وتسمى الرؤوس التي لا تمتلك أبناءً بـ "الأوراق".

الصورة رقم 1 توضح شجرة ثنائية مكونة من 9 رؤوس (تم تعريف كل رأس برقم معين). جذر هذه الشجرة هو الرأس المعرف برقم 2 الواقع أعلى الشجرة، وله ابنين: الابن الأيمن 5، والابن الأيسر 7. عند الانتقال إلى الإبن الأيمن 5 سنجد أن له ابن واحد فقط، وهو الرأس رقم 9.

الرأس رقم 9أيضاً له ابن واحد، وهو الابن الأيسر رقم 4.

الرؤوس رقم 4، 11، 5 و 2 تعتبر أوراق الشجرة حيث أنها لا تمتلك أبناء.



صورة 1: شجرة ثنائية بحجم 9 وعمق 3, مع جذر قيمته 2.

طريقة بناء شجرة هافمان:

نقوم بإنشاء سلسلة الأشجار الثنائية بحيث تتكون كل شجرة من رأس واحد يحتوي على حرف من حروف المعلومة النصية واحتمال ظهوره.

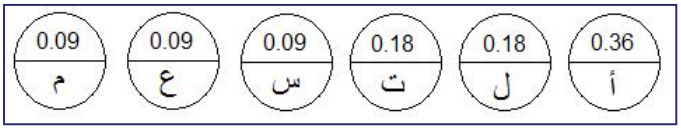
نقوم بإزالة شجرتين من السلسلة بحيث تكون الاحتمالات المسجلة في جذريهما الأصغر من بين كل الأشجار.نقوم بإنشاء شجرة جديدة تتكون من جذر يحتوي على مجموع احتمالات جذري الشجرتين اللاتي قمنا بإزالتهما في الخطوة السابقة. ونقوم بإضافة الأشجار الثنائية المكونة للمثال المذكور مسبقاً.

ISSN 2170-0796

هاتين الشجرتين إلى جذر الشجرة الجديدة.
نقوم بتكرار الخطوات (2، 3) حتى نحصل على شجرة ثنائية واحدة. الشجرة التي حصلنا عليها تسمى شجرة هافمان.
سنحاول الآن بناء شجرة هافمان للمثال السابق.

الخطوة الأولى:

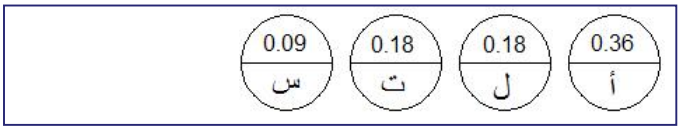
المعلومة النصية (**الاستعلامات**) بنيت باستخدام 6 حروف، وبالتالي سنحصل في البداية على 6 أشجار ثنائية، بحيث تحتوي كل شجرة على رأس واحد يحتوي على الحرف الذي يمثله واحتمال ظهوره في النص.
حرف الألف تكرر في كلمة (الاستعلامات) أربعة مرات، بينما عدد الحروف الكلي المستخدم في الكلمة هو: 11، وبناءاً عليه فإن: احتمال ظهور حرف الألف = 4 / 11 ≈ 0.36 (بعد التقريب). نعيد نفس الحسابات لبقية الحروف. الصورة رقم 2 توضح الأشجار الثنائية المكونة للمثال المذكور مسبقاً.



صورة 2: خوارزمية بناء شجرة هافمان، الخطوة الأولى

الخطوة الثانية:

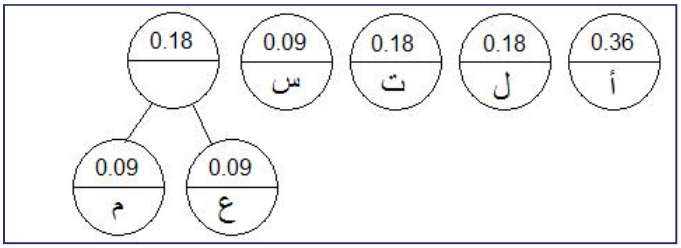
في المثال السابق نجد أن احتمال ظهور الأحرف: س، ع، م هو الأصغر، حيث يساوي بالتقريب 0,09. وبإمكاننا اختيار أي حرفين من هذه الحروف لإزالة الأشجار المكونة له.سنقوم باختيار آخر شجرتين من السلسلة للإزالة كما هو موضح بالصورة رقم 3.



صورة 3: خوارزمية بناء شجرة هافمان، الخطوة الثانية

الخطوة الثالثة:

نقوم الآن بإنشاء شجرة جديدة بحيث يحتوي جذرها على مجموع احتمالي جذري الشجرتين المحذوفتين في الخطوة السابقة.احتمال جذر الشجرة الجديدة = 0,09 + 0,09 = 0,18.

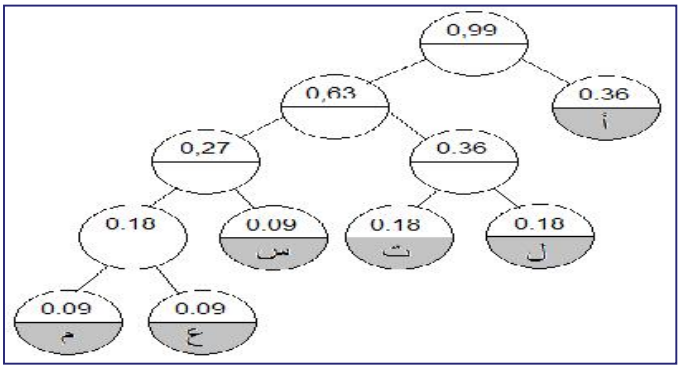


صورة 4: خوارزمية بناء شجرة هافمان، الخطوة الثالثة

الصورة رقم 3 توضح الشجرة الجديدة التي قمنا بإنشائها. لاحظ أن جذر الشجرة الخامسة لا يمثل أي حرف حيث أنه يمثل مجموع احتمالات حرفي ع و م.

الخطوة الرابعة:

الآن نقوم بتكرار الخطوات (2، 3) حتى نحصل على شجرة ثنائية واحدة.بعد تكرار هذه الخطوة 4 مرات سنحصل على شجرة هافمان التالية:نلاحظ أن شجرة هافمان الناتجة تتكون من 6 رؤوس ورقية تمثل الأحرف (الرؤوس المظللة)، بالإضافة إلى 5رؤوس داخلية يحتوي كل منها على مجموع احتمالات أبناءه. نلاحظ أيضاً أن مجموع جميع الاحتمالات المسجل في الجذر يساوي 0,99 (في المثال السابق لم نحصل على العدد 1



صورة 5: شجرة هافمان

ISSN 2170-0796

بسبب تقريب احتمالات ظهور كل حرف من حروف المعلومة النصية) لكن ذلك لا يؤثر على النتيجة النهائية.

الآن سنقوم باستخدام شجرة هافمان لتحديد الأعداد الثنائية التي تمثل كل حرف من الحروف.

آلية تحديد الأعداد الثنائية لكل حرف:

1. نقوم بتمثيل كل رابط يربط بين الرأس وابنه بالطريقة التالية:

– الرابط الذي يربط بين الرأس وابنه الأيمن يمثل بالرقم الثنائي 0

– الرابط الذي يربط بين الرأس وابنه الأيسر يمثل بالرقم الثنائي 1

2. لتحديد القيمة الثنائية لكل حرف نقوم بالانتقال من جذر شجرة هافمان إلى الحرف المراد إيجاد القيمة الثنائية له.

سنقوم الآن بتطبيق هذه الطريقة للمثال السابق.

الخطوة الأولى: بعد تطبيق الخطوة الأولى نحصل على شجرة هافمان كما هو في الصورة رقم 6:

الخطوة الثانية: نقوم الآن بتحديد القيمة الثنائية لكل حرف عن طريق الانتقال من جذر شجرة هافمان إلى الحرف المراد إيجاد القيمة الثنائية له مسجلين الأرقام الثنائية للروابط التي نمر بها.

للوصول إلى الحرف "أ" ننتقل عبر رابط واحد ممثل بالرقم الثنائي 0. إذن العدد الثنائي الذي سيمثل الحرف "أ" هو 0.

للوصول إلى الحرف "ل" ننتقل عبر ثلاث روابط، احدهما ممثل بالرقم الثنائي 1 والآخرين ممثلين بالرقم الثنائي 0. إذن العدد الثنائي الذي سيمثل الحرف "ل" هو 100. للوصول إلى الحرف "ت" نمر بالروابط: 1، 0، 1. إذن العدد الثنائي الذي يمثل الحرف "ت" هو 101.وهكذا...الجدول التالي يوضح الأعداد الثنائية الممثلة لكل حرف من حروف المعلومة النصية:

نلاحظ من الجدول أعلاه أن الحروف المتكررة بكثرة في كلمة "الاستعلامات" تم تمثيلها بعدد ثنائي صغير.

لتحويل المعلومة النصية المتمثلة في كلمة "الاستعلامات" إلى معلومة رقمية نقوم باستبدال كل حرف من حروف الكلمة بالعدد الثنائي الذي يمثله:

الاستعلامات = ا ل ا س ت ع ل ا م ا ت =
1111 1110 0 100 101 110 =
.01001011101101111

نلاحظ أن المعلومة الرقمية (01001011101101111) تتكون من 18 خانة ثنائية.

الحرف	أ	ل	ت	س	ع	م
العدد الثنائي	0	100	101	110	1110	1111

جدول 2: يمثل الأعداد الثنائية التي تمثل كل حرف من حروف المعلومة النصية في ترميز هافمان

3. الفرق بين الترميز ثابت الطول، وترميز هافمان من حيث حجم المعلومة الرقمية الناتجة

سنحاول الآن مقارنة طريقة الترميز ثابت الطول مع ترميز هافمان.

أ. الترميز ثابت الطول

لو أردنا تسجيل المعلومة النصية المتمثلة في كلمة "الاستعلامات" على شكل معلومة رقمية باستخدام طريقة الترميز ثابت الطول فإننا سنقوم بتسجيل كل حرف باستخدام نفس العدد من الخانات الثنائية.

كلمة "الاستعلامات" تتكون من 6 حروف. إذن سنحتاج إلى 3 خانات على الأقل لتسجيل 6 قيم مختلفة تمثل كل واحدة منها حرفاً من الحروف الستة، وستأخذ قيماً من 0 إلى 5.

الجدول رقم 3 يوضح الأعداد الثنائية التي ستمثل كل حرف من الحروف:

الحرف	أ	ل	ت	س	ع	م
العدد الثنائي	000	001	010	011	100	101

جدول 3: يمثل الأعداد الثنائية التي تمثل كل حرف من حروف المعلومة النصية في الترميز ثابت الطول

عند تحويل كلمة "الاستعلامات" إلى معلومة رقمية سنحصل على المعلومة الرقمية:
000001000011010100001000101000010 وهي معلومة مسجلة باستخدام 33 خانة ثنائية (بواقع 3 خانات لكل حرف).

ب. ترميز هافمان

عند استخدام ترميز هافمان حصلنا على معلومة مسجلة على

ISSN 2170-0796

18 خانة، وحيث أن العدد الكلي لحروف كلمة "الاستعلامات" يساوي 11 حرفاً فإن متوسط عدد الخانات المخصصة لكل حرف = 18 / 11 ≈ 1.6 خانة لكل حرف.

نلاحظ أن طريقة **هافمان** ساعدتنا على تقليل حجم المعلومة الرقمية إلى النصف تقريباً، ومقدار الضغط الكبير الذي حصلنا عليه يعود إلى التنوع في تكرار الحروف المختلفة في كلمة "الاستعلامات".

4. استخدام ترميز هافمان لتشفير البيانات

باستخدام ترميز هافمان يمكننا أيضاً تشفير البيانات، حيث تتم عملية تشفير البيانات عن طريق استبدال أبناء الرؤوس الداخلية لشجرة هافمان.

لتحديد طريقة الاستبدال نقوم بتحديد ما يسمى بالمفتاح، وهو عدد ثنائي مسجل باستخدام عدد من الخانات مساوي لعدد الرؤوس الداخلية لشجرة هافمان.

كل خانة ثنائية في المفتاح تحدد فيما إذا كان الرأس التي تمثله في شجرة هافمان قد خضع لعملية استبدال أو لا. القيمة الثنائية 0 تعني عدم وجود استبدال، في حال أن القيمة الثنائية 1 تعني أن الاستبدال قد حصل.

المفتاح 00010 على سبيل المثال يعني أن أبناء الرأس الداخلي رقم 4 في شجرة هافمان قد استبدلت. وحتى يتمكن المستقبل من فك تشفير المعلومة الرقمية وتحويلها إلى معلومة نصية لابد من أن يحصل على هذا المفتاح.

من الجدير بالذكر أن فعالية التشفير في ترميز هافمان تزداد بإزدياد عدد خانات المفتاح، ما يعني أنه كلما كان لشجرة هافمان رؤوس داخلية أكثر، كلما كان التشفير أكثر فعالية.

المراجع

1-A Method for the Construction of Minimum-Redundancy Codes, David A. Huffman, 1952.

2-Data encryption using event-related brain signals, K.V.R. Ravi, R. Palaniappan, C. Eswaran and S. Phon-Amnuaisuk, 2007.

3-Huffman Coding, Steven Pigeon, Universit´e de Montr´eal.

4-Data reduction by Huffman coding and encryption by insertion of shuffled cyclic redundancy code, Nilkesh Patra and Sila Siba Sankar, Department of Electronics & Communication Engineering, National Institute of Technology, Rourkela, 2007.

ISSN 2170-0796