

فشرده سازی اطلاعات

کپی برداری بدون ذکر نام منبع مجاز نیست

parsi e-book

اکثر فایل های موجود بر روی اینترنت با استفاده از نرم افزارهایی

نظیر WinZip فشرده و بر روی سرویس دهندگان FTP مستقر

تا کاربران بتوانند با سرعت مناسب اقدام به دریافت آنها نمایند.

فایل های فشرده ZIP یکی از متداولترین و سهل الوصول ترین

نوع فایل های فشرده می باشند. با فشرده نمودن فایل ها امکان

ارسال سریعتر آنها بر روی اینترنت خصوصاً در مواردیکه سرعت

خط ارتباطی کاربران بالا نباشد ، فراهم می گردد. پس از دریافت

فایل های فشرده با استفاده از نرم افزارهای مربوطه نظیر

parsi e-book
WWW.PARSIBOOK.4T.COM

WinZip می بایست آنها را به حالت اولیه تبدیل (از حالت فشرده

خارج گردند) کرد.

هدف از فشرده نمودن فایل ها کاهش ظرفیت فایل ها بوده و

در زمان استفاده از فایل می بایست مجدداً فایل به حالت اولیه

برگردانده شود. در فرآیند فوق **بیت** هائی از فایل با استفاده از

الگوریتم هائی خاص ، از فایل حذف و زمینه کاهش ظرفیت فایل

فراهم خواهد شد. در زمان استفاده از فایل با استفاده از الگوریتم

فشرده سازی عملیات معکوس انجام و فایل به حالت اولیه خود

برگردانده خواهد شد. در ادامه به برخی از روش های فشرده

سازی اطلاعات اشاره خواهد شد.

یافتن افزونگی در فایل

اکثر فایل های کامپیوتری (با محتویات متفاوت) دارای افزونگی

اطلاعات می باشند. این نوع فایل ها دارای اطلاعات تکراری زیادی

می باشند. برنامه های فشرده سازی اطلاعات ، اطلاعات تکراری

موجود در فایل ها را بر اساس الگوریتم های مربوطه حذف می

نمایند. پس از تشخیص اطلاعات تکراری ، صرفاً اطلاعات تکراری

یک بار در فایل تکرار و و در سایر موارد، از مکانیزمهای خاصی

برای عدم تکرار استفاده می گردد.

جمله زیر از ۱۷ کلمه ، ۶۱ حرف ، ۱۶ فضای خالی ، یک نقطه و یک

dash ، تشکیل شده است :

"Ask not what your country can do for you --

ask what you can do for your country."

parsi e-book
WWW.PARSIBOOK.4T.COM

اگر هر يك از حروف ، فضاى خالى و حروف خاص ، يك واحد از

حافظه را اشغال نمايند ، مجموعاً ۷۹ واحد از حافظه توسط عبارت

فوق استفاده خواهد گرديد (۷۹ = ۱ + ۱۶ + ۶۱) . بمنظور

كاهش ظرفيت فايل مى بايست افزونگى اطلاعات در فايل را

بررسى كرد. با مشاهده و بررسى عبارت فوق ، نتايج زير بدست

مى آيد :

- كلمه "ask" ، دو مرتبه تكرر شده است .

- كلمه "what" ، دو مرتبه تكرر شده است .

- كلمه "your" ، دو مرتبه تكرر شده است .

- كلمه "country" ، دو مرتبه تكرر شده است .

- كلمه "can" ، دو مرتبه تكرر شده است .

- كلمه "do" ، دو مرتبه تكرر شده است .

• کلمه "for" ، دو مرتبه تکرار شده است .

• کلمه "you" ، دو مرتبه تکرار شده است .

با عدم لحاظ نمودن حروف بزرگ و کوچک در عبارت فوق ،

مشاهده می گردد که نیمی از اطلاعات موجود در عبارت فوق ،

زائد و تکراری می باشند. با دقت در عبارت فوق و نحوه افزونگی

اطلاعات مشاهده می گردد که با دارا بودن نه کلمه

for, do, ask, not, what, your, country, can و you می توان

پالایشی مناسبی از عبارت فوق را انجام و در صورت لزوم و با

استفاده از نه کلمه فوق ، مجدداً عبارت اولیه را ایجاد نمود. در

این راستا و بمنظور ایجاد عبارت فوق کافی است به کلمات موجود

در بخش اول (نصف عبارت) اشاره و جایگاه و تعداد تکرار هر

یک از آنها را در بخش دوم مشخص نمود. در ادامه نحوه فشرده

سازی اطلاعات و بازسازی مجدد آنها بررسی می گردد.

کپی برداری بدون ذکر نام منبع مجاز نیست

فشرده سازی اطلاعات

parsi e-book

اکثر برنامه های فشرده سازی از مدل های متفاوت الگوریتم

مبتنی بر دیکشنری ایجاد شده توسط "Ziv و Lempel" ،

بمنظور کاهش ظرفیت فایل ها ، استفاده می نمایند. منظور از

دیکشنری در الگوریتم فوق ، روش های کاتولوگ نمودن بخش

هائی از داده است . سیستم استفاده شده برای سازماندهی

دیکشنری متفاوت و در ساده ترین حالت می تواند شامل یک

لیست عددی باشد. با مراجعه مجدد به عبارت اشاره شده در

بخش قبل ، کلمات تکراری را انتخاب و آنها را در لیست مرتب

parsi e-book
WWW.PARSIBOOK.AT.COM

شده ای بصورت زیر ایند کس می نمائیم . پس از ایجاد لیست فوق

، می توان در مواردیکه از کلمات در عبارت استفاده می شود ، از

اعداد نسبت داده شده و متناظر با آنها استفاده کرد.

دیکشنری ایجاد شده برای عبارت اشاره شده در بخش قبل

بصورت زیر است :

۱. ask

۲. what

۳. your

۴. country

۵. can

۶. do

۷. for

۸. you

با توجه به دیکشنری ایجاد شده ، عبارت مورد نظر بصورت زیر

خوانده خواهد شد :

"1 not 2 3 4 5 6 7 8 -- 1 2

8 5 6 7 3 4"

برای بازسازی مجدد عبارت فوق ، لازم است الگوی معادل آن را

با توجه به دیکشنری استخراج و در محل مربوطه قرار داد. برنامه

هائی نظیر WinZip از فرآیندهای مشابه برای بازسازی مجدد

یک فایل و برگرداندن آن به شکل اولیه استفاده می نمایند.

در فرآیند فشرده سازی عبارت اشاره شده در بخش قبل به

شکل جدید آن (مطابق جدول بالا) چه میزان ظرفیت فایل

کاهش پیدا کرده است ؟ مطمئناً عبارت فشرده شده ظرفیت

کمتری نسبت به عبارت اولیه خواهد داشت . در این زمینه لازم

است به این نکته مهم اشاره گردد که دیکشنری ایجاد شده نیز می

بایست به همراه فایل ذخیره گردد. در مثال فوق ، عبارت اولیه برای

ذخیره سازی به ۷۹ واحد حافظه نیاز داشت . عبارت فشرده شده

(به همراه فضای خالی) ، ۳۷ واحد و دیکشنری (کلمات و اعداد) ،

نیز ۳۷ واحد حافظه را اشغال خواهند کرد. بدین ترتیب ظرفیت

فایل فشرده به ۷۴ واحد حافظه خواهد رسید . با توجه به اطلاعات

فوق مشاهده می گردد که عملاً در رابطه با فشرده سازی عبارت

فوق به موفقیت های بزرگی نائل نشده ایم . در این زمینه لازم

است به این نکته اشاره گردد که در مثال فوق ، صرفاً یک " جمله

" فشرده شده است . فرض کنید جمله فوق بخشی از یک

سخنرانی یک ساعته باشد ، بدیهی است که در سخنرانی فوق

احتمال تکرار کلمات فوق بسیار زیاد خواهد بود . با ایجاد سیستم

دیکشنری ، زمینه استفاده از آن در بخش های بعدی سخنرانی

نیز وجود داشته و در ادامه قطعاً میزان فشرده سازی جملات

موجود در متن سخنرانی نتایج مطلوبتری را بدنبال خواهد داشت

پارسی
بدون ذکر نام منبع مجاز نیست
parsī e-book

جستجو برای الگوها

در مثال ارائه شده ، تمام کلمات تکراری انتخاب و در دیکشنری

قرار گرفتند. در روش فوق ، ساده ترین مدل برای ایجاد

دیکشنری استفاده شده است . برنامه های فشرده سازی از مدل

های کاملاً متفاوت دیگر در این زمینه استفاده می نمایند. برنامه

های فوق نسبت به کلمات متمایز، از یکدیگر شناخت لازم را

نداشته و در این راستا صرفاً بدنبال "الگو" خواهند بود. این نوع

پارسی
WWW.PARSIBOOK.COM
parsī e-book

برنامه ها بمنظور کاهش ظرفیت فایل ها ، با دقت الگوها را

انتخاب و آنها را در دیکشنری مستقر می نمایند. در صورتیکه از

دیدگاه فوق فرآیند فشرده سازی دنبال گردد ، در نهایت با یک

دیکشنری کاملاً متفاوت با آن چیزی که قبلاً ایجاد شده بود ،

مواجه خواهیم بود.

اگر یک برنامه فشرده سازی عبارت معروف اشاره شده در

بخش قبل را بمنظور یافتن افزونگی ، پیمایش نماید ، پس از دنبال

نمودن بخشی از عبارت (your ask not what) ، الگویی جدید را

تشخیص خواهد داد. الگوی فوق حرف "t" بوده که بدنبال آن یک

فضای خالی نیز قرار دارد. (در کلمات "not" و "what") . در

صورتیکه برنامه فشرده سازی الگوی فوق را در دیکشنری مستقر

نماید ، می بایست یک عدد "۱" را در هر زمان که با حرف "t" و

یک فضای خالی بدنبال آن برخورد می نماید ، در دیکشنری ثبت

نماید. با ادامه پیمایش عبارت فوق توسط برنامه فشرده سازی ،

مشاهده می گردد که الگوی تشخیص داده شده (حرف t و فضای

خالی بدنبال آن) به میزان قابل ملاحظه ای در عبارت تکرار نشده

و برای ثبت در دیکشنری واجد شرایط مناسب نخواهد بود ، بدین

تزیب الگوی تشخیص داده شده نادیده گرفته شده و عملیات

یافتن الگوئی دیگر ، دنبال خواهد گردید.

در ادامه برنامه فشرده سازی متوجه الگوی "Ou" می گردد ،

الگوی فوق در کلمات "your" و "country" ، تکرار شده است . در

صورتیکه عبارت مورد نظر یک فایل طولانی بود ، ثبت و نوشتن

الگوی فوق در دیکشنری می توانست به میزان قابل توجه ای از

ظرفیت فایل را کاهش دهد. "Ou" ، یکی از ترکیبات متداول

استفاده شده در زبان انگلیسی است . معیار برنامه فشرده سازی

عبارتی است که در حال پیمایش آن است . در ادامه پیمایش

عبارت فوق ، یک الگوی مناسبتر تشخیص داده خواهد شد.

الگوهای فوق "your" و "country" بوده که هر یک بدفعات تکرار

شده اند. تکرار هر یک از کلمات فوق در عبارت معادل ترکیب

کلمات "your country" است . در چنین حالتی برنامه قشرده

سازی entry موجود در دیکشنری برای الگوی "ou" را با الگوی

"your country" جایگزین می نماید. عبارت ترکیبی "can do

"for" ، نیز در عبارت اصلی تکرار شده است . (یک مرتبه پس از

"your" و یک مرتبه پس از "you") . بدین ترتیب الگوی "can do

"for you" نیز تکراری خواهد بود. بنابراین می توان در عوض

نوشتن ۱۵ حرف (بهمراه قضای حالی) ، از یک عدد استفاده کرد.

در صورت استفاده از الگوی "your country" ، برای ۱۳ حرف از

یک عدد معادل استفاده می گردد ، بدیهی است که الگوی فوق

نادیده گرفته شده در عوض الگوی "r country" و الگوی

جدید "can do fo you" ، در دیکشنری ثبت می گردند. برنامه

فشرده سازی فرآیند فوق را دنبال و پس از یافتن یک الگو ،

محاسبات مربوطه را انجام و الگوی واجدالشرایط را در دیکشنری

ثبت خواهد کرد. مهمترین ویژگی "الگوریتم مبتنی بر دیکشنری" ،

قابلیت تغییر الگوها در زمان فرآیند فشرده سازی است .

با توجه به الگوهائی تشخیص داده شده ، دیکشنری مربوطه بشکل

زیر خواهد بود . در دیکشنری زیر الگوهای تشخیص داده شده

ثبت و برای فضای خالی از کاراکتر " " استفاده شده است .

۱. ask__

۲. what__

۳. you

۴. r__country

۵. __can__do__for__you

با توجه به دیکشنری فوق ، عبارت اشاره شده در بخش قبل

بصورت زیر فشرده می گردد.

--1not__2345__

__12354"

عبارت فوق ۱۸ و دیکشنری ۴۱ ، واحد حافظه را اشغال خواهند

کرد. بدین ترتیب فایل حاوی عبارت اولیه فوق از ۷۹ واحد حافظه

به ۵۹ واحد حافظه کاهش پیدا کرده است. روش استفاده شده

بمنظور فشردن سازی عبارت فوق یکی از امکانات موجود بوده و

می توان در این راستا از روش های دیگر نیز استفاده کرد.

تا چه میزان می توان اطلاعات را فشرده کرد ؟

میزان (نسبت) کاهش ظرفیت یک فایل ، به عوامل متعددی

نظیر : نوع فایل ، اندازه فایل و روش فشردن سازی بستگی دارد.

در اکثر زبانهای طبیعی ، حروف و کلمات الگوهای مناسبی را

بصورت جداگانه و یا ترکیبی ایجاد می نمایند. بدین ترتیب فشردن

سازی فایل های متنی نتایج بسیار مطلوبی را بدنبال خواهد داشت

. فایل های متنی اغلب پس از فشردن سازی به میزان پنجاه درصد

و یا بیشتر ، کاهش ظرفیت را خواهند داشت . اکثر زبانهای برنامه

نویسی (مصنوعی) نیز بدلیل استفاده از مجموعه ای از دستورات

که بصورت تکراری استفاده می شوند ، دارای افزونگی اطلاعات

بوده و پس از فشرده سازی نتایج رضایتبخشی را بدنبال خواهد

داشت . فایل هایی که دارای حجم بالائی از اطلاعات منحصر بفرد

بوده (نظیر فایل های گرافیک و یا فایل های mp3) ، بدلیل عدم

وجود الگوهای تکرار شونده ، بخوبی فشرده نخواهند گردید.

در صورتیکه فایلی دارای تعداد زیادی الگوی تکرار شونده باشد ،

میزان افزونگی اطلاعات موجود در فایل بطرز محسوسی ظرفیت

فایل را افزایش خواهد داد. بدین ترتیب در زمان فشرده سازی

این نوع از فایل ها با توجه به وجود الگوهای تکرار شونده ،

ظرفیت فایل در حد قابل قبولی کاهش پیدا خواهد کرد .

parsi e-book
WWW.PARSIBOOK.4T.COM

میزان فشرده سازی اطلاعات، به الگوریتم استفاده شده توسط

برنامه فشرده سازی نیز بستگی دارد. بدیهی است استفاده از یک

الگوریتم با کار آئی بالا، نتایج مثبتی را در رابطه با فشرده سازی

به ارمغان خواهد آورد



parsi e-book
WWW.PARSIBOOK.4T.COM