پنجمین کنفرانس انجمن رمز ایران
انجمن رمز ایران
Iranian Society of Cryptology
۲۴ الی ۲۶ مهر ماه ۱۳۸۶ دانشگاه علم و صنعت ایران

نهان نگاری اطلاعات در فایلهای XML

فرزین یغمایی
دانشکده مهندسی
دانشگاه سمنان
fyaghmaee@semnan.ac.ir

چکیده: مفهوم نهان نگاری[1] از جمله مفاهیمی است که با گسترش تبادلات دیجیتال و سهولت انتقال اطلاعات اهمیت ویژه ای پیدا کرده است. به طور کلی از نهان نگاری جهت ذخیره سازی اطلاعاتی به صورت پنهان در اطلاعات میزبان[2] برای اثبات مالکیت و یا فهم تبدیلات و تحریفات در اطلاعات میزبان و یا ارسال اطلاعات به صورت مخفی برای گیرنده خاص استفاده می شود. فایلهای چند رسانه ای(اعم از صوتی و تصویری) به واسطه حجم زیاد افزونگی[3] اطلاعات، از قابلیت خوبی جهت نهان نگاری برخوردارند. اما در مقابل فایلهای متنی از ظرفیت بسیار کمی برای پنهان سازی بهره مندند. با توجه به اهمیت فایلهای XML در اینترنت و مقبولیت عمومی این ساختار ذخیره سازی اطلاعات، در این مقاله به بررسی ساختار فایلهای XML پرداخته و سه روش جهت نهان نگاری در آنها ارائه داده ایم که از ظرفیت خوبی در مقایسه با روشهای مرسوم در فایلهای متنی برخوردارند.

واژه های کلیدی: نهان نگاری، فایلهای متنی، نهان نگاری در فایلیهای XML .

۱- مقدمه

نمایش و انتقال اطلاعات به صورت دیجیتال، عمل انتقال و تکثیر اطلاعات را تسریع و تسهیل کرده و این سهولت به نوبه خود منجر به توزیع غیرقانونی و یا تحریف و تغییر اطلاعات شده است. این مساله سبب شده تا گرایشات فراوانی به سمت تکنیکهای مخفی سازی اطلاعات ایجاد شود تا از طریق این اطلاعات پنهانی بتوان تغییرات و تحریفات را تشخیص داد.

به طور کلی دو کاربرد عام برای نهان نگاری اطلاعات می توان تصور کرد:

- اطلاعات اضافه و پنهان شده در اطلاعات اصلی حاوی اطلاعات مربوط به اثبات مالکیت است تا بتوان از طریق اطلاعات پنهان شده پی به مالک اصلی اطلاعات برد و یا آنکه بتوان از طریق این اطلاعات فهمید آیا اطلاعات اصلی در کانال ارتباطی موجب تغییر و تحریف شده است یا خیر؟

[1] Data hiding
[2] host
[3] Redundancy

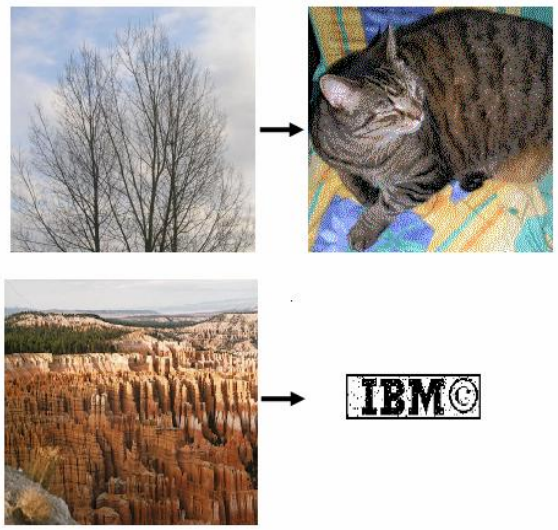
بدیهی است در این حالت اطلاعات مخفی شده باید در مقابل تبدیلات و حملات احتمالی به اطلاعات اصلی پایدار باقی بماند تا بتوان میزان تغییر و یا مالک اصلی اطلاعات را از روی آن تشخیص داد. انواع روشهای الگوگذاری[4] نامرئی از این رده محسوب می شوند.

- اطلاعات مخفی شده ارتباطی به اطلاعات میزبان نداشته و به تنهایی حاوی اطلاعات مورد نظر بین فرستنده و گیرنده است. به عبارت بهتر بر خلاف حالت قبل که اطلاعات مخفی شده معمولاً حجم کمی داشته و جهت حفاظت اطلاعات میزبان به کار می رود، در این روش اطلاعات ارزشمند همان اطلاعات پنهان شده است که به علت اهمیت موضوع از اطلاعات میزبان جهت پنهان سازی آن استفاده شده است. به این ترتیب در این حالت اطلاعات ارزشمند از نظر گیرنده، همان اطلاعات مخفی شده است و نه اطلاعات میزبان. به همین علت در حالت دوم پایداری در مقابل تغییرات چندان مطرح نیست چرا که در این حالت از اطلاعات مخفی شده تنها فرستنده و گیرنده آگاه هستند.

شکل ۱ مقایسه بین این دو کاربرد را هنگامی که اطلاعات در قالب تصویر هستند نشان می دهد. در تصویر (۱-بالا) تصویر یک گربه با حجم نسبتا زیاد در تصویر درخت به صورت پنهان وارد شده است. اما در شکل (۱-پایین) یک آرم تجاری با حجم نسبتا پایین در تصویر قرار دارد. هر یک از دو حالت فوق در کاربردهای خاصی مورد استفاده قرار میگیرند. استفاده از نهان نگاری جهت ارسال اطلاعات به صورت مخفی(حالت دوم) در مباحث پردازش تصاویر به Steganography مشهور است.

به این ترتیب روشهای نهان نگاری براساس حجم اطلاعات مخفی شده، میزان پایداری اطلاعات مقابل تبدیلات، نوع کاربرد و همچنین نوع اطلاعات میزبان (نظیر صوت، تصویر، متن و ..) به انواع مختلفی تقسیم می شوند[9,8,4,1] . اما به عنوان یک اصل کلی هر چه میزان اطلاعاتی که مخفی می شود بیشتر باشد پایداری این اطلاعات در مقابل تبدیلات و تهدیدات کمتر است [9,8]. ادامه مقاله به شکل زیر سازماندهی شده است:

در بخش دوم به مقوله نهان نگاری در متن پرداخته و روشهای معمول نهان نگاری در اطلاعات متنی را بررسی می کنیم. در بخش سوم به معرفی فرمت فایلهای XML پرداخته و در نهایت در بخش چهارم روشهای پیشنهادی خود در زمینه نهان نگاری در فایلهای XML پرداخته و در بخش پنجم به بررسی نتایج می پردازیم.



شکل ۱- مقایسه steganography(تصویر بالا) و الگوگذاری (تصویر پایین) و در فایلهای تصویری

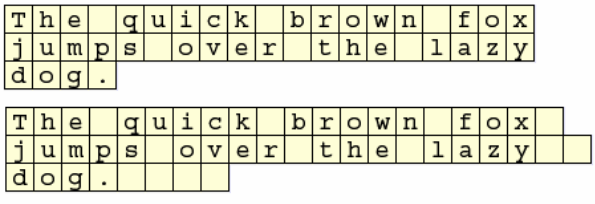
۲-نهان نگاری در متن

اطلاعات متنی از جمله مشکلترین انواع اطلاعات جهت نهان نگاری هستند[9]. به عنوان مثال تغییر ۱۰ پیکسل در یک تصویر ۵۱۲×۵۱۲ به هیچ عنوان از دید چشم محسوس نمی باشد در حالی که اضافه کردن یک نقطه یا یک حرف در متن می تواند معنای جمله یا متن را به کل تغییر دهد. این مسأله به علت عدم وجود افزونگی اطلاعات در فایلهای متنی است که در فایلهای تصویری و صوتی به شدت یافت می شود. دقت کنید منظور ما از فایلهای متنی، فایلهای با فرمت txt. می باشد و نه تصاویر حاوی متن که به document image معروفند و نهان نگاری آنها در مبحث پردازش تصویرمطرح می شود[3,7,5]. در زمینه نهان نگاری اطلاعات متنی سه روش عمده وجود دارد که به اختصار به آنها اشاره می کنیم.

[4] Watermarking

تغییر در جاهای خالی

معمولاً چشم انسان حساسیت چندانی به جاهای خالی در متن ندارد و تعداد آنها را می توان به طرز نامحسوسی کم یا زیاد کرد. در عین حال تغییر تعداد جاهای خالی بر روی مفموم جمله نیز تأثیر چندانی ندارد. یکی از روشهای مرسوم در این زمینه اضافه کردن کارکتر جای خالی در انتهای جملات است. شکل ۲ نمونه ساده ای از تغییرات غیرمحسوس در فضای خالی را نمایش میدهد



شکل۲- نهان نگاری با تغییر در فضاهای خالی [9]

همانگونه که ملاحظه می کنید در شکل تغییر یافته پایین، تعدادی جای خالی بین کلمات در انتهای جملات اضافه شده است که تعداد آنها معرف Ø یا 1 بودن بیت مزبور است و نصف تعداد کارکترهای خالی آخر هر جمله نیز نمایشگر تعداد بیتهای مخفی شده در این جمله است. (مثلا شکل فوق ۴/۲ یعنی دو بیت در جمله ذخیره شده است.) در حالت کلی این روش از نظر حجم ذخیره سازی اطلاعات قابل قبول است ولی باید ویرایشگرها نسبت به جای خالی در انتهای جملات حساس بوده و آنها را اتوماتیک حذف نکنند. مشکل عمده این روش عدم انتقال اطلاعات در صورت چاپ و یا گرفتن تصویر از متن است که عملاً جای خالی انتهای جملات دیگر قابل شمارش نیست.

در این رده روشهای دیگری نیز مطرح است که می توان در [9] مشاهده کرد. به طور کلی عمده روشهای مبتنی بر جای خالی فقط در صورتی که اطلاعات به صورت متن اسکی (ASCII)[5] ذخیره شود کاربرد دارند و در صورت تغییر وضعیت (مثلاً چاپ و یا تبدیل متن به تصویر) کارآیی خود را از دست می دهند.

روشهای مبتنی برگرامر:

در این دسته از روشها اساس نهان نگاری بر مبنای حالات متنوع نگارشی در دستور زبان است. به عنوان نمونه در زبان انگلیسی دو جمله زیر هر دو صحیح می باشد.

bread, butter and milk
bread, butter, and milk

به تفاوت دو جمله دقت کنید. در یکی قبل از and کاما آمده و در دیگری نیامده است. همین نکته ساده می تواند برای ذخیره اطلاعات باینری به کار رود. به این گونه که اگر قبل از and کارکتر کاما باشد نمایانگر بیت صفر پنهان شده است و عدم وجود کارکتر کاما قبل از and نمایانگر بیت یک باشد. این روشها نسبت به تغییر فرم از اسکی به شکل دیگر پایدار است اما با توجه به محدودیت چنین خواصی در یک زبان، معمولاً حجم اطلاعات ذخیره شده در این روش محدود می باشد.

روشهای معنایی:[6]

رده سوم الگوریتها در زمینه نهان نگاری در فایلهای متنی، روش های مبتنی بر معنا می باشد. یکی از ایده های ساده در این روش استفاده از کلمات معادل در متن با دو مفهوم مختلف در نهان نگاری است مثلاً کلمات *"باهوش"* و *"زیرک"* یا کلمات *"جدید"* و *"نو"* معنای یکسانی دارند اما می توان با فرضی قراردادی استفاده از کلمه "باهوش" را معدل مخفی کردن بیت صفر و استفاده از کلمه "زیرک" را به معنای مخفی کردن بیت یک دانست. یکی از مزایای این روش وجود بیش از دو کلمه برای یک مفهوم است که در این حالت مثلاً با داشتن ۴ کلمه *"کهنه"*، *"مستعمل"*، *"قدیمی"* و *"فرسوده"* و استفاده مشخص از هر کدام دو بیت از اطلاعات را ذخیره کرد. شکل زیر این موضوع را نشان می دهد.

بیتهای مخفی شده	کلمه
00	کهنه
01	مستعمل
10	فرسوده
11	قدیمی

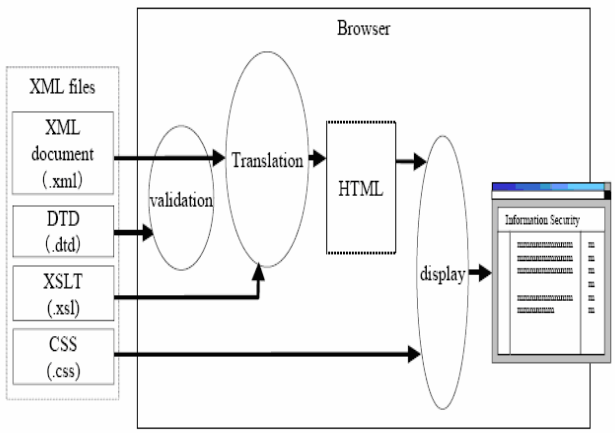
[5] ASCII(American Standard code for Information Interchange)

[6] Semantic methods

از جمله مشکلات این روش آن است که هر چند برخی کلمات معنای مشابهی دارند اما در جملات مختلف کاملاً یکسان به کار نمیروند. مثلاً شما میگویید به *"مدرسه جدیدی رفته اید."* اما نمی گویید به *"مدرسه نوی رفته ایدا!"*.

۳- آشنایی با XML

با گسترش شبکه های کامپیوتری و اینترنت، استفاده از XML[7] به عنوان یک زبان عمومی و کاراً جهت تبادل اطلاعات شدت بیشتری یافته است. به طور کلی استفاده از متون ساخت یافته[8] نظیر فایلهای (XML,PDF) که علاوه بر متن، نحوه نمایش آن را نیزمشخص می سازند به واسطه کیفیت بهتر نمایش و قابلیت های انعطاف بالاتر از اهمیت بیشتری نسبت به متن ساده برخوردار است. در فایلهای XML. محتوای متن، ساختار و نحوه نمایش (Style) به صورت جداگانه در فایلهای با قالبهای مختلف ارزیابی می شوند. محتوای اصلی که حاوی دنباله های کنترلی (tag) و متن اصلی است در فایلی با پسوند XML، و ساختار نمایش در دنباله های کنترلی در فایل DTD[9] و فرمهای نمایش در فایل CSS.[10] ذخیره شده است[11]. معمولاً از فایلهای XSL[11] نیز برای تغیر فرمت XML به HTML برای سازگاری با پوپشگرهای[12] وب استفاده می شود. شکل زیر اجزای یک فایل XML را نمایش می دهد.



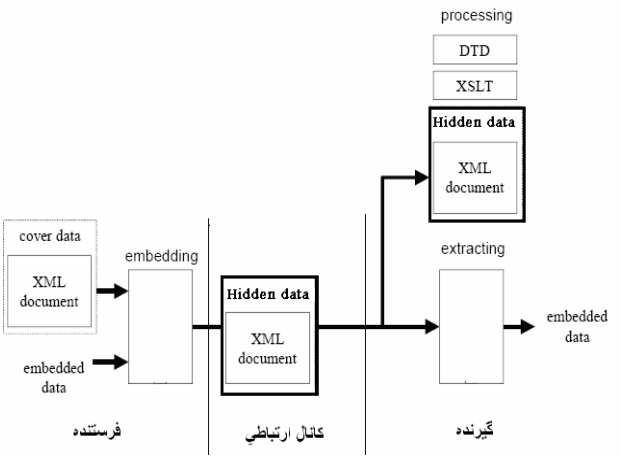
شکل ۳- اجزای اصلی ساختار XML

[7] eXtended Markup language
[8] Structured
[9] Document Type Definition
[10] Cascading Style Sheets
[11] eXtensible Style sheet Language
[12] browser

۴-نهان نگاری در فایل XML

با توجه به اهمیت و گسترش استفاده از فایلهای متنی ساختیافته، نیاز به روشهای نهان نگاری در این فایلها نیز ضروری می باشد که چندان به آن پرداخته نشده است [8]. در این زمینه می توان به کارهای محدودی در زمینه فایلهای PDF و Postscript و Html اشاره کرد[10,6]. در این بخش به معرفی روشهایی جهت مخفی سازی اطلاعات در فایلهای XML می پردازیم. همانگونه که در مطالب قبل گفته شد فایل حاوی اطلاعات همان فایل با پسوند XML. بوده و بقیه فایلها معمولاً حاوی اطلاعات ساختاری و قالبهای نمایشی هستند که می توانند بین گیرنده و فرستنده مشترک فرض شده و از ارسال آنها به گیرنده جلوگیری شود(نظیر DTD). به همین علت این فایلها قابلیت چندانی برای مخفی سازی اطلاعات نداشته و عملاً اطلاعات اضافی باید در فایلهای اصلی XML پنهان شوند. شکل ۴ زیر نمای کلی این مخفی سازی را نشان میدهد.

در این قسمت به شرح روشهای پیشنهادی خود که با توجه به خصایص فایلهای XML طراحی شده است، میپردازیم



شکل ۴- نمای کلی نهان نگاری در فایلهای XML

روش ۱: اجزای کنترلی خالی

بنابر استاندارد روش XML برای ایجاد اجزای خالی می توان از دو شکل دستور زیر استفاده کرد.

< img > </ img> (hidden) →0
< img /> (hidden) →1

با سوییچ کردن بین این دو وضعیت می توان بیتهای صفر یا یک را مخفی کرد. مثلاً در متن زیر دنباله 01010 مخفی شده است.

```
<img src=”pic1.jpg”></img>
<img src=”pic2.jpg”/>
<img src=”pic3.jpg”/></img>
<img src=”pic4.jpg”/>
<img src=”pic5.jpg”></img>
```

روش ۲: جای خالی در دنباله های کنترلی

ایده این روش که در فایلهای متنی ساده نیز وجود دارد استفاده از کارکترهای جای خالی در مواقعی است که در مفهوم جمله چندان تأثیری ندارند. در XML وجود یک جاهای خالی در ابتدای یک دنباله کنترلی تأثیر ندارد.به این ترتیب از کد گذاری زیر برای مخفی سازی 0 و 1 استفاده می کنیم.

```
<tag>, </tag>, or <tag/>          (hidden) →0
<tag  >, </tag  >, or <tag  />    (hidden) →1
```

متن زیر اطلاعات 101100 010011 را ذخیره کرده است.

```
<user ><name>Farzin</name  ><id >01</id></user>
<user><name >Thirdparty</name><id>02</id ></user >
```

روش ۳: جابجایی در ترتیب دنباله کنترلی:

یک روش دیگر برای پنهان کردن اطلاعات، جابجایی در ترتیب دنباله کنترلی عناصر است به صورتی که نحوه نمایش از دید کاربر تغییر نکند. البته این روش برای صفحات XML با اطلاعات خاص و ساخت یافته مناسب است. مثلاً در دو دستور زیر جای متغیر Date نسبت به Month مشخص می کند که منظور ذخیره سازی اطلاعات است صفر است یا یک؟

```
<event month=”MONTH”
date=”DATE”>EVENT</event>          (hidden) →0
<event date=”DATE”
month=”MONTH”>EVENT</event>        (hidden) →1
```

در متن زیر دنباله 01 ذخیره شده است.

```
<event month="MEHR" date="1">Ceremony day</event>
<event date="15" month="AZAR">Birthday </event>
```

ضمناً با توجه به آنکه روشهای فوق از یکدیگر مستقل هستند، امکان ترکیب روشهای فوق و بالا بردن ظرفیت نیز وجود دارد.

۵-مقایسه ظرفیت در روشهای فوق

برای ارزیابی میزان تقریبی نهان نگاری در هر یک از سه روش فوق، ما به طور تصادفی ۵۰۰ صفحه XML با حجم تقریبی بین یک یا دو صفحه را انتخاب کرده و میزان ظرفیت پنهان سازی در هر یک از سه روش ذکر شده را مقایسه کرده ایم. جدول ۱ مقایسه مقدار متوسط بیت ذخیره شده در هریک از روشهای فوق است.

جدول۱-مقایسه ظرفیت در سه روش پیشنهادی

نوع روش	روش اول	روش دوم	روش سوم
متوسط بیت پنهان شده در صفحه	**۴۵**	**۹۴**	**۲۰**

جدول فوق نشان می دهد روشهای اول و دوم از نظر ظرفیت برتری قابل ملاحظه ای نسبت به روش سوم دارند. به نظر می رسد از روش سوم با توجه به شرایط خاص آن، می توان برای الگوگذاری جهت اثبات مالکیت استفاده کرد.

۶- نتیجه گیری

در این مقاله پس از مرور کوتاهی بر مفاهیم نهان نگاری به معرفی ساختار کلی فایلهای XML پرداخته و سه روش جهت پنهان سازی اطلاعات بر اساس مشخصات این زبان معرفی کردیم. به کمک روشهای فوق و یا ترکیبی از آنها می توان تقریبا در هر صفحه وب حدود ۱۵ کارکتر را ذخیره کرد که در مقایسه با روشهای متنی بسیار قابل توجه است. کارهای بعدی در این زمینه می تواند در زمینه یافتن خصایص مناسب در دیگر در فایلهای متنی ساختیافته باشد.

۷- مراجع

1. H. Berghel, "Hiding Data, Forensics, and Anti-Forensics" , COMMUNICATIONS OF THE ACM Journal, April 2007/Vol. 50, No. 4
2. S. Zhong1, X. Cheng1,2, T. Chen, "Data Hiding in a Kind of PDF Texts for Secret", Communication, International Journal of Network Security, Vol.4, No.1, PP.17–26, Jan. 2007

3. A.Mikkilineni, P.Chiang, E.J. Delp," Data Hiding Capacity and Embedding Techniques for Printed Text Documents", Proceedings of the IS&T's NIP22, International Conference on Digital Printing Technologies, Denver, CO, September 17, 2006, pp. 444-447
4. C.Roberto, G. Ryouske,"Data Hiding in Identification and Offset IP fields", Fifth International. Symposium on Advanced Distributed Systems,IEEE ISSADS 2005
5. A.Mikkilineni, G. Ali, J. P. Allebach, E. J. Delp, "Signature-embedding in printed documents for security and forensic applications", Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VI, Volume 5306, San Jose, CA, January 2004, pp. 455-466
6. M.Levene, P.Wood, "XML Structure Compression", Technical Report BBKCS-02-05, School of Computer Science and Information Systems, Birkbeck College, University of London, 2002
7. Q. Wong, N. Memon, "Data Hiding in Binary Text Documents", Security and Watermarking of Multimedia Contents, San Jose, CA, February 2001
8. F. Petitcolas, R.Anderson, M.Kuhn , "Information Hiding,A Survey", , Proceedings of the IEEE, special issue on protection of multimedia content, July 1999.
9. W. Bender, "Techniques for data hiding", IBM SYSTEMS JOURNAL, VOL 35, NO 3&4, 1996
10. M. Shirali-Shahreza, "A New Method for Steganography in HTML Files," Proceedings of the International Joint Conference on Computer, Information, and Systems Sciences, and Engineering (CISSE 2005), Bridgeport,December 2005