2nd Conference on Computer, IT, Electrical and Electronic Engineering 2012
دومین همایش ملی مهندسی کامپیوتر، برق و فناوری اطلاعات
دانشگاه آزاد اسلامی واحد خمین

ارائه یک روش جدید برای واترمارکینگ متون فارسی با استفاده از ویژگی حروف

وحید یزدانی[۱]، محمد علی دوستاری[۲]
[۱]دانشجوی کارشناسی ارشد رشته فن آوری اطلاعات دانشگاه شاهد، v.yazdani@shahed.ac.ir
[۲]استادیار دانشگاه شاهد، doostari@shahed.ac.ir

چکیده - یکی از مسائل مهم در مبادله پیام‌های اینترنتی، مسئله امنیت تبادل اطلاعات است. واترمارکینگ یکی از مهمترین روش‌های مخفی‌سازی اطلاعات است که برای محافظت از حق مالکیت از محتوی چند رسانه‌ها استفاده می‌شود. در این میان در زمینه واترمارکینگ متون فارسی یا عربی با استفاده از خصوصیات حروف به کاررفته در متن، به نسبت متون انگلیسی و چینی تحقیقات کمتری انجام شده است.در این مقاله روش جدیدی برای واترمارکینگ متون فارسی پیشنهاد می‌شود که از حرف "ک" و "ی" برای مخفی‌سازی اطلاعات استفاده کرده و دارای ظرفیت و امنیت بهتری نسبت به روش‌های مشابه است.

کلید واژه- متون فارسی، مخفی‌سازی اطلاعات، پنهان‌نگاری، واترمارکینگ.

۱- مقدمه

استفاده از کامپیوتر، شبکه‌های کامپیوتری، اینترنت و انتقال انواع اطلاعات و رسانه دیجیتال اعم از صوت، تصویر، ویدئو و نرم افزار از طریق این شبکه‌های اطلاعاتی، امری غیر قابل اجتناب در دنیای دیجیتال امروزیست. بنابراین تصدیق هویت رسانه، امنیت اطلاعات در برابر جعل و کپی برداری غیرمجاز، انتقال محرمانه اطلاعات و ... همگی از مواردی به شمار می‌روند که در دنیای دیجیتال امروزی از اهمیت بالایی برخوردارند[۱].

در این میان، روش‌های پنهان‌نگاری و واترمارکینگ به عنوان مهمترین روش‌های مخفی‌سازی اطلاعات محسوب می‌شوند. واترمارکینگ می‌تواند به عنوان یکی از روش‌های پنهان‌نگاری دیده شود که تمرکز بیشتر بر روی استحکام و کمتری بر روی امنیت دارد. امنیتی که پنهان نگاری و واترمارکینگ تامین می-کنند، با امنیت تامین شده توسط رمزنگاری متفاوت است. امنیت واترمارکینگ توسط عدم آگاهی دیگران از وجود اطلاعات مخفی شده در رسانه، تامین می‌شود ولی در رمزنگاری اطلاع از این که در یک رسانه، اطلاعاتی مخفی است، وجود داشته، و کلیه افرادی که دارای کلید باشند قادر به رمزگشایی اطلاعات محرمانه می‌باشند. بر خلاف واترمارکینگ در رمزنگاری هیچ محافظتی از محتوی پیام پس از رمزگشایی وجود ندارد. نکته دیگر تفاوت در فرآیند شکستن سیستم‌های رمزنگاری و واترمارکینگ است. شکستن سیستم‌های رمزنگاری زمانی اتفاق می‌افتد که حمله کننده بتواند پیام مخفی را بخواند ولی در واترمارکینگ این اتفاق زمانی روی می‌دهد که :

۱-حمله‌کننده حضور واترمارکینگ را تشخیص دهد

۲-حمله‌کننده قادر به خواندن، اصلاح یا حذف پیام مخفی باشد.

در ارزیابی الگوریتم‌هایی که برای ایجاد و کشف واترمارکینگ استفاده می‌شوند، پارامترهای زیر متداول است:

۱- استحکام: مربوط به نحوه جاسازی اطلاعات بوده، به طوری که در برابر حملات پایدار باشد. این پارامتر هرچه بالاتر باشد نشانه آن است که میزان تخریب و اثرپذیری بیت‌های واترمارک در اثر تغییرات در رسانه میزبان کمتر است.

۲- ظرفیت: این پارامتر مشخص می‌کند که حداکثر حجم ممکن که رسانه برای درج بیت‌های واترمارک قادر است فراهم کند، چقدر می‌باشد.

۳- امنیت: تعیین می‌کند که رسانه میزبان تا چه میزان وجود بیت‌های واترمارک را می‌تواند مخفی نگه داشته و باعث جلوگیری از دستکاری، جعل و حذف پیام شود[۲].

در حقیقت واترمارکینگ یک روش برای محافظت از حق مالکیت از محتوی چندرسانه‌ها می‌باشد. این چند رسانه‌ها، رسانه‌های مختلفی از جمله متن، تصویر، ویدئو، صدا و اشیاء گرافیکی را شامل شده که هرکدام ویژگی‌های مخصوصی دارند. ازاین ویژگی‌ها برای مخفی‌سازی اطلاعات، دردرون‌شان استفاده می‌شود. از این رو لازم است که الگوریتم‌های واترمارکینگ مختلفی متناسب با نوع اطلاعات توسعه داده شود[۳].

در میان رسانه‌ها، اسناد متنی خواص مختص به خود را دارند. طبیعت باینری، ساختار پارگراف، خط، کلمه، تفکیک مشخص بین پیش‌زمینه و پس‌زمینه که از جمله آن‌هاست. الگوریتم‌های واترمارکینگ نیز از این ویژگی‌ها برای مخفی‌سازی اطلاعات استفاده می‌کنند[۴]. برخلاف واترمارکینگ متنی، روش‌های بسیار زیاد و متنوعی برای واترمارکینگ دیگر انواع رسانه وجود دارد. این در حالی است که در زمینه واترمارکینگ متون فارسی با استفاده از خصوصیات حروف به‌کار رفته، همانند متون چینی و انگلیسی تحقیقات چندانی صورت نگرفته است[۱]. به همین دلیل در این مقاله روش جدیدی برای واترمارکینگ متون فارسی پیشنهاد می‌شود که از از حرف "ک" و "ی" برای مخفی‌سازی اطلاعات استفاده کرده و دارای ظرفیت و امنیت بهتری نسبت به روش‌های مشابه است. مابقی مقاله به صورت زیر سازماندهی شده است:

دربخش دوم، کارهای انجام شده در زمینه واترمارکینگ فارسی و عربی بررسی خواهد شد. بخش سوم مربوط به معرفی الگوریتم واترمارکینگ پیشنهادی بوده و بخش آخر شامل جمع‌بندی روش پیشنهادی خواهد بود.

۲- بررسی کارهای انجام شده در زمینه واترمارکینگ متون فارسی و عربی

متون زبان فارسی بر خلاف متون انگلیسی از راست به چپ نوشته می‌شوند. زبان فارسی دارای ۳۲ حرف بوده که ۱۸ حرف از ۳۲ حرف آن دارای ۱ تا ۳ نقطه می‌باشند. نقاط ممکن است در بالای حرف مانند "ت" و یا در پایین حرف مانند "پ" و یا در میان حرف مانند "چ" قرار بگیرند در حالیکه در زبان انگلیسی فقط دو حرف "i" و "j" دارای نقطه می‌باشند. بر خلاف زبان انگلیسی حروف در کلمات فارسی به هم پیوسته بوده و با توجه به مکان قرارگیری حرف در کلمه ممکن است شکل‌های متفاوتی داشته باشند. به عنوان مثال حرف "ع" با توجه به مکان قرارگیری در کلمه به صورت‌های "ـع"، " ـعـ"، "عـ " و "ع" نوشته می‌شود[۵].

با توجه به خصوصیات ذکر شده درباره حروف فارسی کارهای کمی در زمینه واترمارکینگ متون فارسی انجام شده است. از جمله کارهای انجام شده در زمینه واترمارکینگ متون فارسی و عربی، می‌توان به موارد زیر اشاره کرد:

۲-۱- روش نقطه

از آنجایی که تعداد حروف دارای نقطه در متون فارسی و عربی زیاد است، این روش از نقاط حروف برای درج اطلاعات محرمانه بهره می‌گیرد. بدین صورت که برای درج بیت صفر مکان نقطه در حروف دارای نقطه تغییری نکرده و برای درج بیت یک، مکان نقطه در حروف دارای نقطه کمی به سمت بالا انتقال می‌یابد. شکل ۱ نحوه تغییر مکان نقطه در حرف "ن" را نشان می‌دهد.

شکل۱: نمونه ای از واترمارکینگ با جابجایی نقاط [۷]

از ویژگی‌های این روش می‌توان به ظرفیت بالا و استحکام پایین اشاره کرد. همچنین در این روش با تغییر اندازه و نوع قلم و نوشتن مجدد متن پوششی اطلاعات محرمانه از بین خواهد رفت[۷].

۲-۲- روش استفاده از حروف مشابه با کدهای متفاوت

حروف فارسی و عربی در چهار حرف اختلاف دارند. مابقی حروف تقریبا شبیه به هم بوده و تنها با توجه به مکان قرارگیری‌شان در کلمه تفاوت جزئی خواهند داشت، که از جمله آن می‌توان به حروف "ک" و "ی" فارسی با "ک" و "ی" عربی اشاره کرد. جدول ۱ تفاوت شکل ظاهری این دو حرف را در آغاز، وسط، انتها و جدا از کلمه در دو زبان فارسی و عربی نشان می-دهد.

در این روش برای درج بیت صفر از شکل عربی حروف "ک" و "ی"، و برای درج بیت یک از شکل فارسی حروف "ک" و "ی" استفاده می‌شود. در حالت‌هایی که شکل دو حروف در دو زبان فارسی و عربی تفاوتی ندارد(مانند حرف "ی" در ابتدای کلمات)، اطلاعاتی درج نخواهد شد.

جدول۱ : تفاوت شکل ظاهری حروف "ک " و "ی" [۶]

حرف "ک"	حرف"ی"	زبان	وضعیت کلمه
کـ	یـ	فارسی	آغاز
ك	يـ	عربی	
ـکـ	ـیـ	فارسی	وسط
ـك	ـيـ	عربی	
ـک	ی	فارسی	انتها
ـك	ي	عربی	
ک	ی	فارسی	جدا
ك	ي	عربی	

ظرفیت این روش پایین بوده ولی دارای شفافیت ادارکی بالایی است. همچنین این روش در برابر انتقال به محیط‌های دیجیتالی مانند Word، Pdf، HTML مستحکم بوده ولی برا اثر نوشتن مجدد متن پوششی اطلاعات محرمانه از بین خواهد رفت[۶].

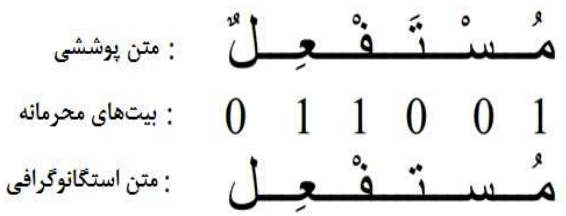
۲-۳- روش استفاده از کلمه "لا"

در این روش از یک کلمه یکسان که دارای دو شکل ظاهری متفاوت بوده و در سیستم استاندارد یونیکد، دو کد متفاوت دارد، برای درج بیت‌های واترمارک استفاده می‌شود. از جمله این کلمات، کلمه "لا" می‌باشد. در کلمه" لا" برای درج بیت یک از "لا " و برای درج بیت صفر از شکل عادی، یعنی "لا" استفاده خواهد شد.کلمه "لا" دارای کد FEFB و کلمه "لا " دارای کد ۰٦٤۰ در سیستم استاندارد یونیکد است.

هرچند ظرفیت این روش پایین بوده ولی این روش محدود به اسناد الکترونیکی نبوده و بر اثر چاپ، اطلاعات محرمانه پایدار باقی خواهد ماند. دراین روش، نوشتن مجدد متن پوششی باعث ازبین رفتن اطلاعات محرمانه خواهد شد[۵].

۲-۴- روش استفاده از علائم موجود بر حروف

در این روش جهت مخفی‌سازی داده بر روی متن پوشش ابتدا باید مطمئن شد که کلیه علائم در پوشش متنی حاضر هستند. سپس به ترتیب، کلیه بیت‌هایی که قرار است مخفی شوند، متناظر با یک علامت قرار می‌گیرند. حال اگر مقدار بیتی که قرار است مخفی شود یک بود، علامت متناظر با آن بیت بدون هیچ تغییری در متن حاضر خواهد بود. ولی اگر مقدار بیتی که قرار است مخفی شود صفر بود، علامت متناظر با آن بیت از متن حذف خواهد شد. شکل ۲ نمونه‌ای از این نوع واترمارکینگ را نشان می‌دهد.



شکل۲: نمونه ای از واترمارکینگ با استفاده از علائم [۱]

از مزایای این روش علاوه بر ظرفیت بالا به کارگیری آسان آن بوده و در برابر حملاتی نظیر OCR و چاپ مقاوم است. ولی این روش فاقد استحکام است و از این رو بهتر است برای واترمارکینگ ضعیف یا تمامیت داده استفاده شود[۱].

۲-۵- روش استفاده ازکاراکتر کشیده

کاراکتر کشیده عربی در نوع الکترونیکی تنها برای مرتب-سازی و فرمت‌بندی به عنوان یک کاراکتر عادی در نظر گرفته می شود. نکته قابل ذکر در استفاده از کاراکتر کشیده این است که تمام حروف به دلیل موقعیت‌شان در کلمات و ماهیت نوشتاری زبان عربی قابل گسترش نیستند. همچنین کاراکتر کشیده نمی‌تواند پس از حرف پایانی کلمه یا قبل از شروع اولین حرف کلمه اضافه شود. کاراکتر کشیده را می‌توان به بعد یا قبل

از حروف که قابلیت اضافه شدن کاراکتر کشیده را دارند، اضافه کرد.

در این روش برای درج بیت صفر کاراکتر کشیده به قبل یا بعد از حرف فاقد نقطه و برای درج بیت یک کاراکتر کشیده به قبل یا بعد از حرف دارای نقطه اضافه که قابلیت اضافه شدن کاراکتر کشیده را دارند، تعبیه خواهد شد. شکل ۳ یک نمونه از جزئیات فرآیند این نوع از واترمارکینگ در حالت درج کاراکتر کشیده به بعد از حروف را نشان می‌دهد.

جهت افزایش و گمراهی بیشتر، می‌توان از هر دو حالت درج کاراکتر کشیده به قبل یا بعد از حروف در یک متن پوششی بهره برد. بدین‌گونه که برای بعضی از خطوط یا پارگراف مشخص در متن، از حالت درج کاراکتر کشیده به قبل از حروف، و برای مابقی خطوط یا پارگراف از حالت درج کاراکتر کشیده به بعد از حروف استفاده کرد.

۱۱۰۰۱۰	بیت‌های محرمانه
من حسن اسلام المرء ترکه مالا یعنیه	:متن پوششی
مـن حسن اسلام المرء تـرکـه مـالا یـعـنـیه ۰ ۱ ۰ ۰ ۱ ۱	: متن خروجی

شکل۳ : نمونه‌ای از واترمارکینگ با استفاده از حروف کشیده [۲]

در مقایسه این روش با روش نقطه، این نکته قابل ذکر است که در این روش، مخفی کردن یک بیت درمتن، معادل درج یک کاراکتر کشیده می‌باشد. در صورتی‌که روش نقطه فاقد چنین افزایش اندازه‌ای است. در نتیجه دارای ظرفیت کمتری نسبت به روش نقطه می‌باشد. در مقابل این روش از امنیت و استحکام بالاتری نسبت به روش نقطه برخوردار بوده و اطلاعات محرمانه در برابر تغییر نوع و اندازه قلم نیز، پایدار خواهد ماند[۲].

۳- روش پیشنهادی

همان‌طوری که قبلا ذکر شد، زبان فارسی و عربی در چهار حرف باهم اختلاف داشته، و مابقی حروف تقریبا شبیه به هم بوده و تنها با توجه به مکان قرارگیری‌شان در کلمه تفاوت جزئی خواهند داشت. روش استفاده از حروف مشابه با کدهای متفاوت که در در بخش ۲-۲ معرفی گردید، از این اختلاف ظاهری حروف برای تعبیه بیت‌های محرمانه استفاده کرده است. مشکل این روش، ظرفیت و امنیت پایین آن می‌باشد. لذا در این بخش روش جدیدی معرفی می‌شود، که مشابه با روش استفاده از حروف مشابه با کدهای متفاوت بوده، ولی علاوه بر مزایای آن روش، دارای ظرفیت و امنیت بالاتری است.

در کلیه روش‌های قبلی ابتدا اطلاعاتی که قرار است مخفی شود، به صورت رشته‌بیت‌هایی از صفر و یک تبدیل شده، که نشان‌دهنده کد ۱۶ بیتی برای هر کاراکتر است (با توجه به طرح کدگذاری UTF-8 که جهت نمایش کاراکتر عربی یا فارسی از ۱۶ بیت استفاده می‌کند). اِشکال مشترک در روش‌های قبلی این است که درهمه این روش‌ها عمل درج بیت‌های محرمانه بدون هیچگونه بهینه‌سازی صورت می‌گیرد. همانطوری که قبلاً ذکر شد، زبان فارسی دارای ۳۲ حرف بوده، که می‌توان هر حرف را با ۶ بیت نمایش داد.

در حقیقت بخش بهینه‌سازی در ارتباط با پیامی است که قرار است مخفی شود. در این روش از یک جدول نگاشت استفاده می‌شود. در این جدول به هر حرف ۶ بیت اختصاص داده می‌شود. اولین حرف از جدول نگاشت دارای مقدار ۰۰۰۰۰۰ بوده و برای مابقی حروف و اَشکال آن‌ها، به این مقدار یک واحد به ترتیب حروف الفبا اضافه خواهد شد. در نتیجه در این روش با صرفه‌جویی در۱۰ بیت، ظرفیت به‌طور چشمگیری زیاد خواهد شد.

بعد از عمل بهینه‌سازی، برای درج بیت صفر از شکل عربی حروف "ک" و "ی"، و برای درج بیت یک از شکل فارسی حروف "ک" و "ی" استفاده می‌شود. شکل ۴ و جدول ۲ مثالی از این روش را نشان می‌دهد. اولین حرف در شی محرمانه، حرف "ب" بوده که دارای کد ۰۰۰۰۰۱ در جدول نگاشت می‌باشد. برای تعبیه دو صفر اول از بیت محرمانه، حرف "ک" و "ی" در کلمه اول به صورت عربی نوشته می‌شود. کلمات سوم، هفتم و هشتم به ترتیب بیت‌های سوم، چهارم و پنجم از بیت محرمانه را نگه می‌دارند. سرانجام کلمه پانزدهم بیت یک را در خود تعبیه می‌کند.

بد	پیام محرمانه
یکی از مسائلی که در مبادله پیام‌های اینترنتی، مهم است، مسئله امنیت تبادل اطلاعات می‌باشد. واترمارکینگ یکی از مهمترین روش‌های مخفی‌سازی اطلاعات	متن پوششی

می‌باشد که برای محافظت از حق مالکیت از محتوی چند رسانه‌ها استفاده می‌شود.	
یکی از مسائلی که در مبادله پیام‌های اینترنتی، مهم است، مسئله امنیت تبادل اطلاعات می‌باشد. واترمارکینگ یکی از مهمترین روش‌های مخفی‌سازی اطلاعات می‌باشد که برای محافظت از حق مالکیت از محتوی چند رسانه‌ها استفاده می‌شود.	متن خروجی

شکل۴ : نمونه‌ای از واترمارکینگ روش پیشنهادی

جدول ۲ : نمونه‌ای از جدول نگاشت

د	ب	حرف
۰۰۱۰۱۰	۰۰۰۰۰۱	کد

یکی از نگرانی‌هایی که در استفاده از این روش وجود دارد، هنگامی است که مقدار اطلاعاتی که قرار است مخفی شود، بسیار کمتر از متن پوششی بوده و برای درج اطلاعات محرمانه چند خط ابتدایی متن کافی باشد. بنابراین حروف "ک" و "ی" فقط در چند سطر ابتدایی دچار تغییر شده و در مابقی خطوط بدون هیچ تغییری در مکان خود می‌مانند. این باعث ایجاد تردید برای یک استراق‌سمع‌کننده که ممکن است وجود یک پیام مخفی را در متن استنباط کند، بشود. این مشکل با ایجاد یک کاراکتر خاص به نام "کاراکتر پایانی" برطرف خواهد شد. این کاراکتر دارای کد ۱۱۱۱۱۱ بوده و بعد از درج آخرین بیت از پیام محرمانه تعبیه خواهد شد. بعد از کاراکتر پایانی به طور تصادفی حروف "ک" و "ی" به صورت عربی یا فارسی نوشته خواهند شد.

آخرین اصلاح در این روش برای بهینه‌سازی، استفاده از جدول نگاشت پویا است. برای این منظور یک بررسی آماری روی اطلاعات محرمانه قبل از درج صورت می‌گیرد، تا فقط برای حروف‌های تشکیل دهنده پیام، جدول نگاشت را تشکیل داد. به عنوان مثال اگر در یک پیام محرمانه فقط چهار حرف حضور دارند، با دو بیت می‌توان اطلاعات محرمانه را درج کرد. نکته‌ای که در این حالت وجود دارد، این است که تنها کسانی که به جدول پویا دسترسی دارند، قادر به استخراج اطلاعات هستند. همچنین سربار تشکیل جدول پویا و به روز رسانی آن برای هر پیام زیاد می‌باشد، ولی به ظرفیت بیشتر و امنیت بالا به دلیل گمراهی که ایجاد می‌شود، کمک خواهد کرد.

۴- نتیجه‌گیری

در این مقاله روش جدیدی برای واترمارکینگ متون فارسی ارائه شده است که با استفاده از دو حرف "ک" و "ی" و تفاوت در شکل ظاهری‌شان در دو زبان فارسی و عربی اقدام به درج اطلاعات محرمانه می‌کند.

تفاوت این روش با سایر روش مشابه را می‌توان در ظرفیت و امنیت بالای آن دانست. در روش پیشنهادی با استفاده جدول نگاشت طول بیت برای نمایش هر حرف کوتاه‌تر شده، و در نتیجه محیط میزبان می‌تواند اطلاعات بیشتری را در خود جای دهد. دراین روش، از آنجایی که از جدول نگاشت (به خصوص جدول نگاشت پویا) استفاده می شود، و برای هر حرف به جای استفاده از ۱۶ بیت به ۶ بیت یا کمتر نیاز است، حتی با استخراج بیت‌ها، تشخیص اینکه کدام بیت‌ها مربوط به یک حرف می‌باشد، بسیار سخت است. در نتیجه امنیت این روش نیز بالاست. همچنین این روش در برابر انتقال به محیط‌های دیجیتالی مقاوم بوده، و با تغییر اندازه قلم، اطلاعات محرمانه از بین نخواهد رفت، ولی بر اثر نوشتن مجدد متن پوششی، و تغییر نوع قلم، اطلاعات محرمانه از بین خواهد رفت.

مراجع

[1] Bensaad, M.L., Yagoubi, M.B., "*High capacity diacritics-based method for information hiding in Arabic text*", International Conference on Innovations in Information Technology (IIT), 2011, vol.11, no.12, pp.433-436, 16 June 2011.

[2] Adnan Gutub, Lahouari Ghouti, Alaaeldin Amin, Talal Alkharobi and Mohammad K.Ibrahim, *"Utilizing Extension Character 'Kashida' With Pointed Letters For Arabic Text Digital Watermarking"* International Conference on Security and Cryptography–(SECRYPT),Barcelona, Spain, 28 – 31 July 2007.

[3] Jaseena K.U., Anita John, *"Text Watermarking using Combined Image and Text for Authentication and Protection"*, International Journal of Computer Applications (0975 – 8887) Volume 20– No.4, April 2011.

[4] Young-Won Kim, Kyung-Ae Moon; Il-Seok Oh, *"A text watermarking algorithm based on word classification and inter-word space statistics"*, Proceedings Seventh International Conference on Document Analysis and Recognition- ICDAR, pp. 775-779, 3-6 Aug. 2003.

[5] M. Shirali-Shahreza, *"A New Persian/Arabic Text Steganography Using "La" Word"*, in Proceedings of the International Joint Conference on Computer, Information, and Systems Sciences, and Engineering (CISSE 2007), Bridgeport, CT, USA, Vol. 2, pp. 339–342, 2007.

[6] M. Hassan Shirali-Shahreza, Mohammad Shirali-Shahreza, *"Arabic/Persian Text Steganography Utilizing Similar Letters With Different Codes"*, The Arabian Journal for Science and Engineering, Volume 35, December 9, 2009.

[7] M.H. Shirali-Shahreza and M. Shirali-Shahreza, *"A New Approach to Persian/Arabic Text Steganography"*, Proceedings of the 5th

IEEE/ACIS International Conference on Computer and Information Science (ICIS 2006), Honolulu, HI, USA, pp. 310-315, July 10-12, 2006.