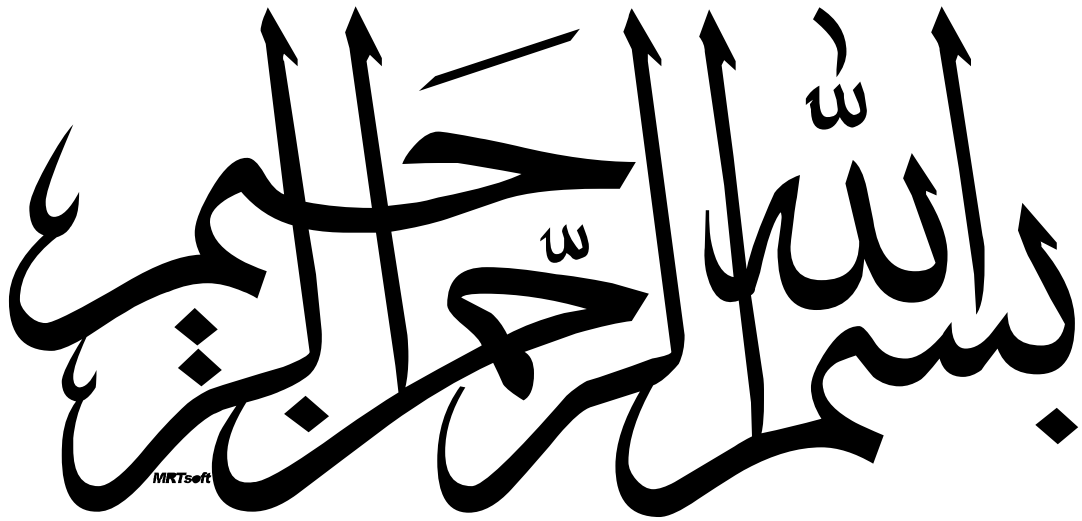

بسم الله الرحمن الرحيم
MRTsoft

دانـــشگاه یـــزد

دانشکده مهندسی برق و کامپیوتر

گروه مهندسی کامپیوتر

پایان‌نامه

برای دریافت درجه کارشناسی ارشد

مهندسی کامپیوتر – هوش مصنوعی

رتبه‌بندی هوشمندگراف وب برای مقابله با صفحات فریب‌آمیز

استاد راهنما

دکتر علی‌محمد زارع‌بیدکی

استاد مشاور

دکتر ولی درهمی

پژوهش و نگارش

معین شهبازی

مهر ماه ۱۳۹۲

تقدیم به پدرم

به خاطر همه خوبی‌ها و مهربانی‌هایش؛

و همچنین تقدیم به مادرم، بهشت خدا در زمین

که همواره با دعاهای خیرش پشت و پناه من بوده و هست؛

.... اگرچه می‌دانم این اوراق بی‌بها ذره‌ای از زحماتشان را جبران نمی‌کند.

تقدیر و تشکر

شکر شایان نثار ایزد منان که توفیق را رفیق راهم ساخت تا این پایان‌نامه را به پایان برسانم. سپس از پدر و مادر که نخستین آموزگارانم بودند تا معلمین، دبیران و استادانم که هر یک به‌زعم خویش بر آموخته‌هایم افزودند، قدردانی می‌نمایم. مراتب سپاس خود را از جناب آقای دکتر علی‌محمد زارع‌بیدکی به‌عنوان استاد راهنما که همواره نگارنده را مورد لطف و محبت خود قرار داده‌اند، ابراز می‌دارم. همچنین از جناب آقای دکتر ولی درهمی، استاد مشاور اینجانب که از رهنمودهای ایشان در این پژوهش بهره برده‌ام، کمال تشکر را دارم.

چکیده

با گسترش روزافزون وب در سال‌های اخیر و کاربردهای آن در همه زمینه‌ها از جمله تجارت الکترونیک، بازرگانی و اجتماعی و همچنین با بالا رفتن پتانسیل سود در وب، اکثر توسعه دهندگان صفحات به فکر بازدیدکننده بیشتر از سایت‌ها هستند. در این میان تشخیص محتوای با کیفیت از تلاش‌های فریب‌کارانه جهت به بازی گرفتن موتورهای جستجو به یکی از چالش‌های اصلی این موتورها تبدیل شده است. تاکنون الگوریتم‌های مختلفی برای مقابله با صفحات فریب‌آمیز ارائه شده‌اند که به دو دسته کلی الگوریتم‌های براساس محتوا و الگوریتم‌های براساس پیوند تقسیم می‌شوند. دسته اول با تجزیه و تحلیل محتوای اسناد توانایی تشخیص و مقابله با صفحات فریب‌آمیز را دارند. همچنین در دسته دوم با تجزیه و تحلیل پیوند و رفتار بین صفحات عمل مقابله و تشخیص انجام می‌شود. مشکل اصلی الگوریتم‌های مبتنی بر محتوا، نادیده گرفتن پیوندهای بین صفحات است. در مقابل در الگوریتم‌های براساس پیوند تشخیص الگوهای غیرطبیعی در میان گراف‌های بزرگ از چالش‌های این الگوریتم‌ها است.

هدف اصلی این پایان‌نامه، ارائه یک روش رتبه‌بندی ترکیبی به منظور مقابله با صفحات فریب‌آمیز است. در این روش با ترکیب الگوریتم‌های مختلف رتبه‌بندی سعی می‌شود رتبه صفحات فریب‌آمیز پایین کشیده شود. با توجه به کارایی پایین روش‌های پایه‌ای موجود در مقابله با این پدیده روش پیشنهادی سعی می‌کند با دخیل کردن یادگیری ماشین در فرآیند رتبه‌بندی بهبود نسبی در کارایی روش مبنا بوجود می‌آید. در این روش سعی می‌شود هر چه بیشتر پویایی گراف وب با استفاده از روش‌های یادگیری مدل شود. فرآیند یادگیری جهت ترکیب رتبه‌های مختلف با استفاده از مشاهدات و دانش قبلی می‌باشد که در این حالت روش پیشنهادی دارای کارایی و دقت بیشتری نسبت به روش موجود می‌باشد.

کلمات کلیدی: صفحات فریب‌آمیز، رتبه‌بندی، یادگیری ماشین، موتور جستجو

فهرست مطالب

فهرست شکل‌ها

فصل اول

مقدمه

در سالیان اخیر استفاده از وب به عنوان منبع اصلی کسب اطلاعات، رشد فزاینده‌ای داشته است. یکی از دلایل اصلی این مهم نیز رفع نیازهای اطلاعاتی[۱] کاربران در حداقل زمان ممکن است. ازجمله ابزارهای بسیار کارآمد در این زمینه موتورهای جستجو[۲] بوده و امروزه نقطه ورود بسیاری از کاربران برای جستجو در وب، موتورهای جستجوی مشهوری مانند گوگل[۳]، یاهو[۴] و بینگ[۵] هستند. بر اساس آمار منتشره، تنها درکشور ایالات متحده آمریکا، موتورهای جستجوی گوگل و یاهو به ترتیب با حدود ۱۲۵ و ۱۱۵ میلیون کاربر، مقام اول و سوم پربازدیدترین وب‌سایتها را در ژوئن ۲۰۰۹ به خود اختصاص داده‌اند [۵]. در فرآیند کاری این موتورها کاربران پرس‌وجوی[۶] خود را در قالب یک یا چند کلمه کلیدی[۷] در موتور جستجو وارد نموده و در پاسخ لیستی از اسناد[۸] (عموماً صفحات وب) که معمولاً شامل کلمات کلیدی مورد نظر بوده و به احتمال زیاد مرتبط[۹] با نیاز اطلاعاتی آنان است را دریافت می‌کنند.

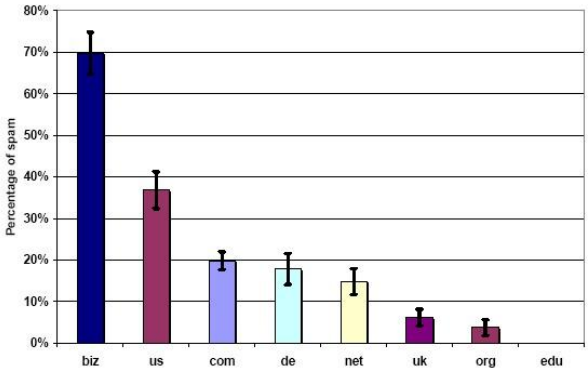
به علت هزینه پایین تولید محتوا در محیط وب، حجم اطلاعات بسیار زیاد بوده و همچنان نیز رو به افزایش است. برای نمونه تعداد صفحات نمایه‌سازی شده توسط موتور جستجوی گوگل، حدود ۳۷ میلیارد صفحه در اکتبر ۲۰۱۰ می‌باشد [۲]. لذا ممکن است در پاسخ به پرس‌وجوی کاربر تعداد بسیار زیادی از اسناد، کاندیدای انتخاب به عنوان سند مرتبط باشند. همچنین کاربران معمولاً به ۱۰ یا ۲۰ نتیجه اول اکتفا نموده [۱] و در صورتی که به هدف خود نرسند، یا پرس-وجوی خود را تغییر داده یا از جستجو منصرف می‌شوند. در نتیجه سازمان‌ها و شرکت‌های بسیاری برای به دست آوردن بازدیدکننده بیشتر و در نتیجه سود بیشتر تلاش می‌کنند موتورهای جستجو را گمراه کنند و رتبه صفحات مورد نظر خود را با روش‌های فریب‌کارانه بالا ببرند. معرفی پدیده صفحات فریب‌آمیز[۱۰] مبحثی است که پیدایش آن همزمان با ظهور اینترنت بوده و در

[1] Information Need
[2] Search Engine
[3] Google.com
[4] Yahoo.com
[5] Bing.com
[6] Query
[7] Keyword
[8] Document
[9] Relevant
[10] Spam

سال‌های اخیر بسیار شایع شده است. اما به تازگی محافل دانشگاهی و علمی مرتبط نیز توجه خود را به این مسئله و روش‌های مقابله با آن معطوف کرده‌اند. دو مشکل اساسی که صفحات وب فریب‌آمیز برای موتور جستجو ایجاد کرده و مانع از کارکرد صحیح آن می‌شود، شامل بر هم زدن تعادل و دقت موتور جستو در بازیابی نتایج و همچنین به هدر دادن منابع موتورهای موتور جستجو در زمان نمایه‌سازی[1] صفحات است.

۱-۱ بازیابی اطلاعات خصمانه[2]

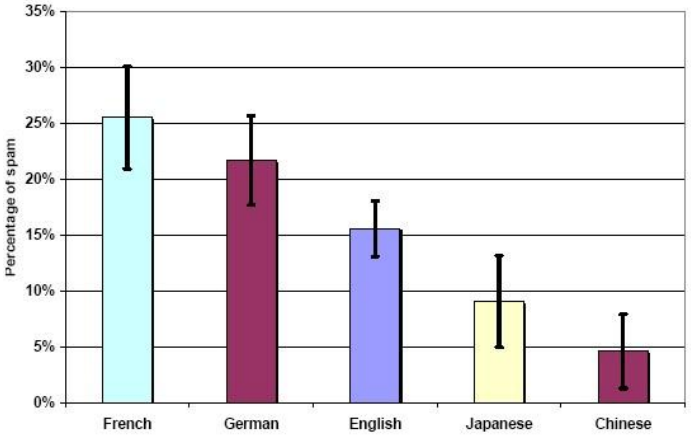
به طور کلی به بازیابی اطلاعات خصمانه شامل وظایف جمع‌آوری، نمایه‌سازی، فیلتر کردن و بازیابی اسناد است با این تفاوت که از میان این اسناد بخشی از آن‌ها با روش‌های فریبکارانه و نامشروع دستکاری شده‌اند. بر طبق پژوهشی [۱] که در سال ۲۰۰۶ بر روی دامنه‌های مختلف انجام شده است حجم صفحات فریب‌آمیز در ۸ دامنه بزرگ اینترنت مورد بررسی قرار گرفت. شکل ۱-۱ نتیجه این آزمایش را نشان می‌دهد.



شکل ۱-۱: حجم صفحات فریب‌آمیز بر روی دامنه‌های مختلف [۱]

همان‌گونه که مشاهده می‌شود در دامنه تجاری biz.* بیشترین درصد صفحات فریب‌آمیز وجود دارد. همچنین در ادامه همان پژوهش درصد وجود صفحات فریب‌آمیز در زبان‌های مختلف نیز بررسی شده که نتیجه آن در شکل ۱-۲ آمده است.

[1] Indexing
[2] Adverserial Information Retrieval



شکل ۱-۲ : درصد وجود صفحات فریب‌آمیز در زبان‌های مختلف [۱]

به منظور مقابله با این پدیده موتورهای جستجو استراتژی‌های مختلفی را به کار می‌گیرند. از جمله می‌توان گفت که این موتورها ابتدا بر اساس معیارهای تعریف شده برای تشخیص صفحات فریب‌آمیز، آنها را شناسایی کرده و سپس رتبه‌بندی[۱] آنها را تنزل می‌دهند و یا این‌که پس از تشخیص، آنها را از لیست نمایه‌ها حذف کرده و دیگر آنها را خزش[۲] نکنند. مرحله اول این فرآیند می‌تواند یک مسئله دسته‌بندی[۳] باشد. ورودی دسته‌بندی‌ها ممکن است ویژگی‌های استخراج شده از محتوای متن، ویژگی‌های پیوندهای فرامتن و همچنین اطلاعات ثبت نامی DNS[۴] و یا بازخورد کاربران باشند.

۱-۲ تعریف مسئله و صفحات فریب‌آمیز

واژه فریب‌آمیز کردن صفحه به ترفندهایی گفته می‌شود که برای دست بردن در الگوریتم‌های رتبه‌بندی موتورهای جستجو و تغییر نتیجه جستجو به سمت صفحات وب هدف استفاده می‌شوند. موتورهای جستجو معمولا صفحات وب را بر اساس دو معیار ارتباط پرس‌وجو با صفحه وب و اهمیت صفحه وب رتبه‌بندی می‌کنند. معیار اول بیان کننده میزان شباهت پرس‌وجو با صفحات وب نمایه‌سازی شده است که در [۳] روش‌هایی برای اندازه‌گیری این شباهت آمده است. همچنین منظور از اهمیت صفحه در واقع میزان اهمیت و مهم بودن آن صفحه بدون در نظر گرفتن پرس‌وجو

[1] Page Ranking
[2] Crawl
[3] Classification
[4] Domain Name System

است. این اهمیت مثلا ممکن است با توجه به تعداد ارجاعاتی که از صفحات دیگر به این صفحه شده است اندازه‌گیری شود.

طبق تعریف دیگری که برای صفحات فریب‌آمیز آمده است هرگونه تلاش و فعالیت غیر نرمال برای برهم زدن روش‌های رتبه‌بندی صفحات در موتور جستجو به منظور بالا بردن رتبه یک یا چند صفحه خاص به صورت غیر منطقی، صفحه فریب‌آمیز نامیده می‌شود [۲]. این مسئله از این جهت قابل اهمیت است که از یک طرف باعث از بین رفتن شهرت موتور جستجو می‌شود و از طرف دیگر نیز باعث ایجاد هزینه‌های اضافی برای دارندگان موتورهای جسجو است. بنابراین روش‌های مقابله با این پدیده می‌تواند بسیار مفید باشد.

قدم اول در مقابله با این پدیده و سازندگان این صفحات شناخت روش‌هایی است که به کار می‌برند. بدین منظور برای آشنایی هر چه بیشتر با روش‌های فریب‌کارانه یک طبقه‌بندی از این روش‌ها تعریف می‌کنیم که در ادامه ارائه می‌شود.

۱-۳ طبقه‌بندی روش‌های فریب‌کارانه

تکنیک‌های فریب‌کارانه برای ایجاد صفحات فریب‌آمیز در دو دسته بزرگ تکنیک‌های تقویتی[1] و تکنیک‌های مخفی‌سازی[2] دسته‌بندی می‌شوند [۴]. در دسته اول از روش‌هایی به منظور رسیدن به یک جایگاه بالا در بحث میزان اهمیت و ارتباط صفحه استفاده می‌شود در حالی که دسته دوم شامل روش‌هایی است که بر روی الگوریتم‌های رتبه‌بندی صفحات تاثیر می‌گذارند. در ادامه به تشریح هر دسته می‌پردازیم.

[1] Boosting Techniques
[2] Hiding Techniques

۱-۳-۱ فریب دادن مبتنی بر واژه[1]

در این روش که زیرمجموعه‌ای از تکنیک‌های تقویتی است محتوای صفحه هدف به گونه‌ای سازمان دهی می‌شود که با درخواست‌های بیشتری منطبق شود.

در اندازه‌گیری میزان اهمیت و مرتبط بودن یک صفحه تنها با اتکا به محتوای آن، قسمت‌های مختلف صفحه که متن در آن‌ها قرار می‌گیرد با وزن‌های مختلفی توسط موتور جستجو ارزیابی می‌شود. به عنوان مثال می‌توان متون قرار گرفته در قسمت‌های بدنه صفحه، عنوان صفحه، عنوان پیوندها و دیگر قسمت‌ها را نام برد که از این میان معمولا عناوین پیوندها بیشتر توسط کاربران دیده می‌شوند و وزن بالاتری نیز خواهند داشت. بنابراین فریب‌گرها[2] با تمرکز روی این موضوع می‌توانند اهمیت یک یا چند صفحه خاص را بالا ببرند.

یکی از اصلی‌ترین الگوریتم‌هایی که مورد توجه سازندگان صفحات فریب‌آمیز مبتنی بر واژه قرار می‌گیرد الگوریتم $TFIDF$ است [۳]. بر طبق این الگوریتم میزان اهمیت صفحه p برای پرس‌وجوی q به صورت زیر محاسبه می‌شود.

$$TFIDF(p,q) = \sum_{t \in p \ and \ t \in q} TF(t) \,.\, IDF(t) \qquad (۱-۱)$$

که در رابطه (۱-۱) $TF(t)$ تعداد رخدادهای واژه tام در صفحه p و $IDF(t)$ برای واژه tام برابر نسبت کل اسناد به تعداد اسنادی است که شامل واژه t بوده‌اند. قسمت $IDF(t)$ از رابطه (۱-۱) به طور معمول خارج از کنترل فریب‌گرها است. اما آنها می‌توانند تنها با تمرکز بر روی قسمت $TF(t)$ میزان اهمیت یک یا چند صفحه را تغییر دهند. این کار معمولا می‌توانند با تکرار بیش از حد طیف وسیعی از کلیدواژه‌های متفاوت و غیر تکراری در بخش‌های مختلف صفحه و یا با تکرار بیش از اندازه یکسری کلیدواژه‌های خاص در صفحه صورت گیرد که در نهایت هر دو روش باعث بالا رفتن میزان اهمیت و ارتباط صفحه برای پرس‌وجوهای عامه‌پسند و خاص می‌شوند.

[1] Term Spamming
[2] Spammer

۱-۳-۲ صفحات فریب‌آمیز مبتنی بر پیوند[1]

یکی دیگر از روش‌های رایج و مطرح در تکنیک‌های تقویتی، صفحات فریب‌آمیز مبتنی بر پیوند است. در این روش موتور جستجو از اطلاعات پیوندهای بین صفحات برای اندازه‌گیری اهمیت و ارتباط صفحه استفاده می‌کند. بنابراین فریب‌گرها نیز با تمرکز بر روی این اطلاعات و با ایجاد کردن انواع ساختارهای پیوند به دنبال بالا بردن میزان اهمیت یک یا چند صفحه خاص هستند. در این روش صفحات موجود در گراف وب را می‌توان در سه دسته بزرگ تقسیم‌بندی کرد. دسته اول شامل صفحات معتبری است که سازندگان آن‌ها هیچگونه اجازه‌ای به حضور فریب‌گرها در آن صفحات را نمی‌دهند و به طور مختصر به این دسته، وب غیر قابل دسترسی[2] گفته می‌شود. در دسته دوم ایجاد و تغییر در صفحات معمولا توسط گروه‌های دیگری انجام می‌شود که با فریب‌گرها رابطه‌ای ندارند اما فریب‌گرها با روش‌های محدودی می‌توانند در این صفحات دستکاری‌هایی را انجام دهند. به عنوان مثال فریب‌گرها می‌توانند با یادداشت گذاشتن در وب سایت‌های اجتماعی و عمومی یا در وبلاگ‌ها پیوندهایی را به سایت‌های خود برقرار کنند. به این دسته از صفحات نیز معمولا وب قابل دسترس[3] می‌گویند دسته سوم از صفحات به طور مستقیم توسط خود فریب‌گرها ایجاد شده و آن‌ها کنترل کاملی بر روی پیوندهای خروجی و محتوای آن را دارند و از این صفحات معمولا برای ایجاد کردن مزرعه‌های پیوند[4] استفاده می‌کنند. این دسته از صفحات وب، صفحات مالک[5] نام دارند.

با توجه به این تقسیم‌بندی، می‌توان روش‌های ایجاد صفحات فریب‌آمیز بر اساس پیوند را در دو دسته ایجاد بر اساس پیوندهای خروجی[6] و همچنین پیوندهای ورودی[7] دسته‌بندی کرد. در دسته اول فریب‌گرها به امید بالا بردن رتبه، صفحه را با طیف وسیعی از پیوندهای خروجی پر

[1] Link Spamming
[2] Inaccessible Pages
[3] Accessible Pages
[4] Link Farms
[5] Own Pages
[6] Outgoing Links
[7] Incoming Links

می‌کنند. این روش که شبیه‌سازی دایرکتوری[1] نام دارد شامل کپی کردن پیوندهای مختلف در صفحه از وب سایت‌های معتبر فهرست راهنمای وب است. در دسته دوم تمرکز بر روی پیوندهای ورودی است چرا که در الگوریتم‌های رتبه‌بندی این پیوندهای تاثیر بیشتری بر روی رتبه صفحه خواهند داشت. بدین منظور فریب‌گرها با روش‌های گوناگونی از جمله ایجاد کردن انواع مزرعه‌های پیوند، خرید دامنه‌های تاریخ دار، تبادل پیوند بین سایت‌های همانند، یاداشت گذاری پیوندها در وبلاگ‌ها و ایجاد کردن ظرف‌های عسل[2] به هدف خود می‌رسند [۴].

الگوریتم‌های هدفی که در این روش بیشتر مورد توجه فریب‌گرها است، شامل الگوریتم‌های رتبه‌بندی مبتنی بر پیوند می‌شوند.

۱-۳-۳ مخفی‌سازی محتوا[3]

از تکنیک‌های دیگر رایج برای فریب‌گرها، دستکاری کردن محتوای صفحات است. بدلیل این‌که این دستکاری‌ها اغلب باعث خراب کردن ظاهر صفحات می‌شوند بنابراین فریب‌گرها روش‌هایی را در پیش گرفته‌اند که از طرفی ظاهر صفحات خوب به نظر برسد و از طرفی نیز اهمیت صفحه در موتورهای جستجو بالا برود.

عبارت‌های فریب‌کارانه را می‌توان طوری در صفحه قرار داد که در هنگام تفسیر توسط مرورگرها دیده نشوند. یکی از این تکنیک‌های رایج استفاده کردن از رنگ‌های یکسان در بدنه صفحه و پس زمینه صفحه است. بدین ترتیب برای انسان هیچ متنی روی صفحه دیده نمی‌شود اما برای موتور جستجو این متون قابل مشاهده است. نمونه‌ای از این تکنیک در زیر آورده شده است.

```
<body bgcolor=”white”>
  <font color=”white”> hidden text here </font>
</body>
```

[1] Directory Cloning
[2] Honey Pot
[3] Content Hiding

از دیگر فریب‌های مخفی‌سازی محتوا می‌توان استفاده از تصاویر تک پیکسلی را به عنوان متن پیوندها نام برد. در این صورت نیز پیوندهایی در صفحه وجود دارد که برای انسان قابل مشاهده نیست. نمونه‌ای از این فریب نیز به صورت زیر است.

```
<a href=”targetspampage.html”><img src=”tinyimg.gif” /></a>
```

۱-۳-۴ ماسک گذاری[1]

در این روش نسخه‌های مختلفی از صفحه برای خزشگرهای موتورجستجو و کاربران به نمایش در می‌آید. اگر فریب‌گرها بتوانند در هنگام بازدید از صفحه، بازدیدکننده آن را تشخیص دهند، می‌توانند خزشگر موتور جستجو را فریب داده و در هنگام مرور صفحه نسخه‌ای غیر از آنچه که به کاربران نمایش داده می‌شود را برای نمایه سازی به خزشگر تحویل دهند. در این صورت فریب‌گرها بدون اینکه مورد ردیابی موتورهای جستجو قرار گیرند به هدف خود رسیده‌اند. برای فریب‌گرها تشخیص دادن خزشگر از دیگر کاربران می‌تواند به دو صورت انجام گیرد. در روش اول آن‌ها می‌توننند با ساختن لیستی از آدرس‌های IP خزشگرها، در هنگام بازدید از صفحه با کاوش در لیست، آن‌ها را تشخیص دهند. در روش دوم در پروتکل HTTP[2]، نام برنامه درخواست کننده صفحه تحت عنوان User Agent مشخص می‌شود بنابراین فریب‌گرها با تغییر این اطلاعات می‌توانند در هنگام درخواست خزشگر یا کاربر نسخه‌های مختلفی از صفحه را ارائه دهند. به عنوان نمونه در زیر این اطلاعات برای یک کاربر که با مرورگر Internet Explorer قصد بازدید از صفحه را دارد آورده شده است.

```
GET /members.html HTTP /1.0
Host: www.ce.yazd.ac.ir
User-Agent: MSIE/4.0 (compatible MSIE 6.0 Windows NT 5.1)
```

[1] Cloaking
[2] Hyper Text Transfer Protocol

به عنوان یک راه‌حل ساده جهت مقابله با این نوع فریب‌ها می‌توان با تغییر دادن فیلد اطلاعاتی مربوطه در خزشگر و معرفی کردن آن شبیه به کاربران از نسخه‌های مختلف صفحات در وب سایت‌های فریب‌آمیز آگاه شد.

۱-۳-۵ تغییر جهت[1]

نمونه‌ای دیگر از مخفی‌سازی محتوا در صفحات، تغییر جهت است. در این فرآیند کاربران هنگام درخواست کردن یک صفحه به صورت خودکار به صفحه‌ای دیگر که یک صفحه فریب‌آمیز است هدایت می‌شوند. در این روش ابتدا فریب‌گرها با ایجاد کردن دامنه‌های معتبر، خود را به عنوان یک سایت معتبر به خزشگر موتور جستجو معرفی می‌کنند. سپس در ادامه محتوای این صفحات را طوری تغییر می‌دهند که کاربران به هنگام درخواست این صفحات، به صورت خودکار به صفحات هدف تغییر مسیر داده شوند. به عنوان نمونه یکی از تکنیک‌هایی که فریب‌گرها جهت تغییر مسیر از آن استفاده می‌کنند، صفر کردن مقدار زمان نوسازی صفحه و تنظیم کردن مقدار آدرس نوسازی با آدرس صفحات هدف است. در زیر نمونه ای از این فریب مشخص شده است.

```
<meta http-equiv=”refresh” content=”0” url=”target.html” >
```

اگرچه این روش به سادگی توسط خزشگر قابل شناسایی است، اما فریب‌گرها معمولا از تکنیک‌های پیچیده‌تری در این خصوص استفاده می‌کنند که قابل پیگیری توسط خزشگر هم نباشد، مانند استفاده کردن از اسکریپت‌هایی که بخشی از صفحه بوده و به صورت زیر تعریف می‌شوند.

```
<script language=”javascript”>location.replace(“target.html”)</script>
```

تا اینجا یک طبقه‌بندی کوچک از روش‌هایی که فریب‌گرها به منظور ایجاد صفحات فریب‌آمیز از آن‌ها استفاده می‌کنند آورده شد در ادامه پژوهش به معرفی چالش‌های پیشرو جهت مقابله با این پدیده می‌پردازیم.

[1] Redirection