Mollom

Technical whitepaper

The web is changing. User contribution is now what makes or breaks a site. Allowing users to react, participate and contribute while still keeping your site under control can be a huge challenge. Mollom is a web service (Software as a Service) that helps you identify content quality and, more importantly, helps you stop comment and contact form spam, and cap protect registration forms.

Mollom analyzes the quality of content posted to websites and tries to determine whether this content is unwanted or not. Websites that allow visitors to contribute or post comments are constantly being flooded with inappropriate, distracting or even illegal commercial messages, many of which are uploaded by automatic "spambots." Mollom screens all contributions before they are posted to participating websites. We use Machine Learning techniques, Language Analysis and a reputation system to ease moderation and improve the overall quality of your site's content.

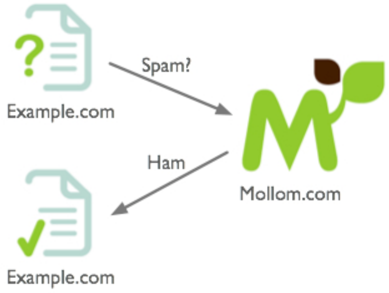
Mollom also provides a centralized CAPTCHA service that allows to protect e.g. user registration forms using both image and audio CAPTCHAs. Mollom constantly monitors and tweaks its CAPTCHAs so they are still easily solvable by humans, but cannot be solved by automated scripts.

This document elaborates on the technical aspects of the Mollom services, but does not go into technical detail on how to implement the service's open API. For this we refer to the developer API documentation which can be found on our website.

1. Mollom products and services

1.1. Text filtering and content analysis service

The Mollom filtering service is a hosted web service that analyzes the quality of content posted to websites. This includes comments, contact form messages, blogs, forum posts, etc. Mollom specifically tries to determine whether this content is unwanted - i.e. "spam" - or desirable - i.e. "ham." Websites that allow visitors to contribute or post comments are often being flooded with inappropriate, distracting or even illegal commercial messages, many of which are uploaded by automatic "spambots". Mollom's text filtering and content analysis service screens all contributions before they are posted to participating websites.



Websites using Mollom send data they want checked to mollom.com, and Mollom replies with either a spam or ham classification. If Mollom is not certain, it will return "unsure", typically prompting websites to ask Mollom's CAPTCHA service for an audio or visual CAPTCHA challenge to present to the user.

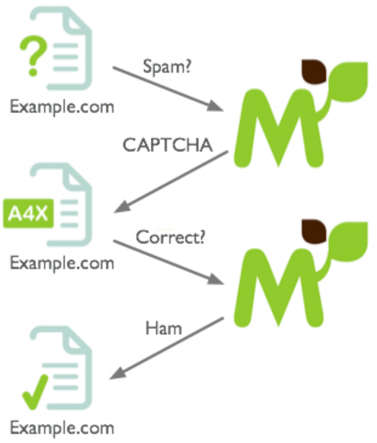
The fact that Mollom can reply "unsure" makes Mollom unique compared to other services. Thanks to the "unsure" reply and the CAPTCHA challenges, Mollom avoids incorrectly classifying legitimate contributions as spam. The strategy of combining text classification with occasional CAPTCHAs has two important benefits:

Mollom statistics

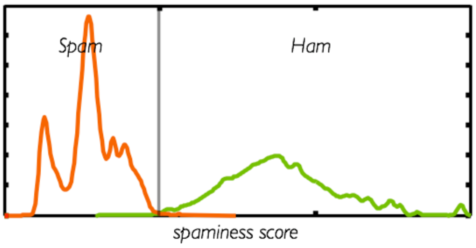
As of October 2008, Mollom protects over 6,000 websites and is used by companies such as Sony BMG (more than 80 of their web sites), Acquia, Adobe, IDG, Fast Company, Now Public, LinuxJournal, Jupitermedia, The New York Observer, and many more. Mollom has blocked over 10,000,000 spam comments since its start, has an average spam-stopping accuracy of approximately 99,83%.

1. It effectively eliminates the need to moderate messages that Mollom decides to block since they are unlikely to have come from legitimate users.

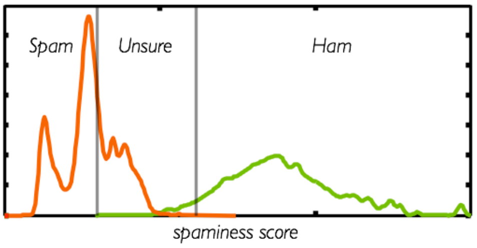
2. Because CAPTCHA rarely challenges legitimate content (currently only approximately 4% of human users are challenged with a CAPTCHA), it makes your site a lot more accessible while still greatly improving its overall quality.



Spam fighting tools compute a score based on words and links present in the content under investigation. This 'spaminess' score indicates how likely it is that a post is spam or not. Conventional spam fighting tools return a 'ham' result when it seems _likely_ that a post is ham rather than spam, given its spaminess score. This decision line is shown in the graph below. Here, the green line denotes known ham messages, while the red line denotes known spam messages. So if a message is analyzed, and its spaminess score is to the right of the decision boundary, it is considered to be ham.



What is the problem with this approach? Not all content is correctly classified. This may appear to be only a tiny fraction on the plot, but when millions of messages are being processed all the time, we are talking about 1,000's of misclassified messages every day. Some posts that are actually spam land on the right side, the ham side, of the decision boundary where they don't belong. This spam is not recognized by the system and is allowed onto your site. On the other hand, some legitimate messages fall into the spam bucket and will be blocked from your website. Neither of these are desirable outcomes. To counteract this, a conventional spam blocking system dumps all the messages in the spam category into a moderation queue. The site moderator has to periodically go through all of the message to manually moderate the few ham messages that were incorrectly classified as spam. This is tedious work.



Here is how Mollom solves these problems. Instead of two classes, we define three: 'spam', 'unsure' and 'ham'. Mollom returns 'spam' only if it is 100% sure that the post is spam and these posts are discarded. If Mollom is quite certain (more certain than using the old technique) that a post is ham, it is accepted. But what about the rest?

We define a gray zone, an area of uncertainty, and here is where the CAPTCHAs come in. When Mollom is unsure about a submission, the user is asked to respond to a CAPTCHA. If the response is correct, and thus the submitter is human, the content will be accepted. Otherwise the post will be rejected. Only a tiny fraction of real human-submitted content falls into our 'unsure' zone and triggers a CAPTCHA (currently, only approximately 4% of human submissions). To the very largest extent, CAPTCHAs are not shown to humans at all, they are shown to the "spambots!"

Mollom supports a wide range of languages, including non-Latin languages such as Arabic and Asian.

Top 5 benefits:

1. Get rid of website spam
2. Higher quality content
3. Moderation made easier
4. Keep your site accessible
5. Multilingual support

Top 5 features:

1. Next generation filtering and text analysis technology
2. CAPTCHA service
3. Open API
4. Highly available platform
5. Detailed statistics

1.2. CAPTCHA service

Mollom also offers CAPTCHA services which allow the protection of forms that do not contain content, such as user registration forms. Automated scripts are known to create hundreds of fake users on websites, even circumventing email-based user authentication. These fake accounts are used to spam, abuse or hack the website. These forms can be secured using CAPTCHAs.

CAPTCHAs are created to differentiate real human users from automated scripts, however some CAPTCHAs can be solved by advanced computer algorithms. Another common way to circumvent CAPTCHAs is the outsourcing of CAPTCHA solving to human agents in developing countries. By using the Mollom CAPTCHA web service, we can constantly monitor the quality of the CAPTCHAs we are generating over all the sites that use Mollom and instantly adapt them if we see they have been hacked. We also use a reputation-based system to monitor the people solving CAPTCHAs, allowing us to block the outsourced human CAPTCHA solvers.

Web site accessibility is a very big issue. Adding image CAPTCHAs to web sites makes them inaccessible to visually impaired people using screen readers or braille screens. Mollom allows users to request audio CAPTCHAs, which can be solved by visually impaired users. But more importantly, only a very small percentage of users will ever see (or hear) a CAPTCHA! Using intelligent filtering techniques, we present CAPTCHAs only to users posting suspicious content. By using Mollom, your website becomes both more accessible and better-protected.

CAPTCHAs tend to be ugly and spoil the nice design of your site. We thought this shouldn't be the case and provide visually much more appealing CAPTCHAs while still being very hard to solve using automated techniques. Mollom constantly monitors its CAPTCHA to make sure humans can always easily solve them, while automated scripts are not able to crack them.

2. Mollom hosted platform

2.1. Network learning

A key Mollom feature is that all participating websites can send feedback to Mollom, explicitly marking comments as spam, profanity or low-quality. Mollom combines the information sent in by all participating websites and learns from it, preventing future abuse.

2.2. Centralized reputation management

The decision if content is spam or not is not solely based on the Language Analysis and Machine Learning based content analysis. Mollom internally builds a reputation of all users submitting content to any of the sites protected by Mollom. This cross-site reputation allows for a highly effective discrimination between spam and non-spam. Because the reputation is build over all the sites protected by Mollom we can very quickly detect spammers, and completely block them from posting any spam content on the protected sites.

Note that content is only considered "spam" (and thus blocked without further checks) if both the reputation system and the content analysis agree that the user and the content can be considered spam. This way, it become practically impossible that Mollom wrongly blocks real users trying to submit content, even if they truly are having a conversation about e.g. Viagra.

The reputation system also supports OpenID, a decentralized, single sign-on system, which allows users to log in to different websites using only one set of credentials. Mollom

CAPTCHA

CAPTCHA stands for "Completely Automated Public Turing test to tell Computers and Humans Apart." It is a type of challenge-response test used to determine whether a user is human or just an automated script. This is done by asking a user to solve a challenge that is hard for computers, but relatively easy for human beings builds a reputation for each of our users' OpenIDs, allowing the inheritance of user reputations across sites. The reputation is based on the quality of the user's content, the amount of CAPTCHAs solved by the user, and how he participates on the protected sites.

2.3. Open API and supported platforms

Mollom uses an open API. Anyone can create a Mollom plug-in for their favorite content management system, even if it is proprietary. Common CMSes that are currently supported are Drupal, Wordpress, Joomla, SilverStripe and Radiant. Open-source libraries for PHP5, Python, .NET, Ruby and Java are provided for easy integration of Mollom in other CMSes. The Mollom Client API documentation is available at http://mollom.com/api.

2.4. Service level agreement

Mollom provides paying customers with a "Standard Service Level Agreement" that guarantees the availability of working Mollom servers. Mollom users with a free Mollom subscription do not receive any service level agreements. A "Premium Service Level Agreement" with support guarantees (i.e. phone support, support response times) and additional service availability guarantees can be negotiated separately.

The Mollom infrastructure was built from the ground up to be extremely scalable and robust. Mollom has servers in multiple independent data centers around the world providing a high degree of redundancy. Mollom provides a "client-side load balancing mechanism"; when one of the Mollom servers is down or overloaded, the customer's application can move on to the next Mollom server. This mechanism allows for a high-availability as it is unlikely that different servers in different data centers around the globe would simultaneously be unavailable. Mollom servers are permanently monitored, and Mollom has technicians on call that can respond to any emergency 24/7 should any issue arise.

In the very unlikely event that the Mollom servers are unavailable, the customer's application can provide a configurable fallback strategy. You can choose to blindly accept all submissions without spam checking (to optimize for data loss avoidance), or you can choose to block all submissions until the Mollom server problems are resolved (to optimize for spam hitting your site). Most of the third-party plugins available on the Mollom website provide such fallback strategy.

3. Pricing

The basic Mollom services are free to sites with limited post volumes. It provides all the features of our Mollom Plus product, but is limited in the number of forms it will protect each day and limited in access to our high-availability back-end infrastructure.

Mollom Plus is our subscription-only, high-availability spam filtering and CAPTCHA service. Mollom Plus not only lets you process up to 10,000 legitimate posts or 10,000 correct CAPTCHAs each day, the Mollom Plus clients have access to a very robust high-availability backend infrastructure not available to our Mollom Free clients. If your site has a large daily volume of posts or you need access to Mollom Plus' high-availability network of servers, Mollom Plus is your best choice. Or, if you hate spam as much as we do and believe that Mollom's services have reduced the amount of time you're spending policing your content, consider a Mollom Plus subscription to help support our efforts to fight spam.

We strongly respect the needs of enterprise users: security, robustness, high quality of service and high volume. For you, we run dedicated Mollom servers or provide volume-based pricing that scales well beyond the volume limits of a Mollom Plus account. Furthermore, we recognize that some people might want to tailor Mollom to the specific needs of their company or products. If you want to integrate Mollom into your products, explore joint go-to-market strategies, or if you want to explore a reseller relationship, make sure to contact us. We'll do all we can to help you!

Mollom bvba
Langveld 2
2600 Berchem
Belgium
www.mollom.com
info@mollom.com