Internet Safety Technical Task Force
Technology Submission Template

EthoSafe, Inc.
http://www.ethosafe.com

ABSTRACT

EthoSafe was born out of an idea that the “ethos” of an online, interactive dialogue can be maintained by implementing tagging technology and moderation services that automatically enforce the community’s publication rules. Consumer brands entering the social media market want to protect their brands and their online users from content that falls outside the standards of its dialogue. EthoSafe is developing a platform that focuses on creating a safe environment through content filtering using learning technology, artificial intelligence and human review.

Keywords

Filtering, tagging, learning technology, content moderation

Functional Goals

Please indicate the functional goals of the submitted technology by checking the relevant box(es):

□ Limit harmful contact between adults and minors
□ Limit harmful contact between minors
✓ Limit/prevent minors from accessing inappropriate content on the Internet
□ Limit/prevent minors from creating inappropriate content on the Internet
□ Limit the availability of illegal content on the Internet
□ Prevent minors from accessing particular sites without parental consent
□ Prevent harassment, unwanted solicitation, and bullying of minors on the Internet
□ Other – please specify

PROBLEM INTRODUCTION

EthoSafe provides a turnkey solution for the moderation of user-generated content and social media. We ensure that brands and organizations can take advantage of the benefits of social media while also providing safety for their brand and customers.

Jupiter Research reported in March 2007 that “porn and other offensive imagery, inappropriate language, offensive postings, and spam have created nightmares for brands and ad agencies that have launched social media and campaigns featuring use-generated content.” Of those interviewed, 67% of marketers expressed fear of ability to control content and brand image.

The target market for EthoSafe is any interactive program that invites online dialogue with its customers or constituents. The brand industries most likely to engage in UGC initiatives are in the entertainment, retail, recreation, sports, beverages and media categories. Brands that engage with the under-18 audience are particularly interested in EthoSafe services.

PROPOSED SOLUTION

The EthoSafe platform is a software-as-a-service (SaaS) solution. Client websites integrate with it via a RESTful API. The EthoSafe platform analyzes text, audio, video, or image content to determine whether or not certain tags apply to it. The tags that are assigned to the piece of content are then compared against rules that the client establishes to determine whether or not the content conforms to those rules. A message indicating content acceptance or rejection is then sent back to the client website via HTTP POST.

Before a client begins to send content into the EthoSafe platform for analysis, rules that determine what constitutes content acceptability are configured. EthoSafe has a series of approximately 55 standard tags, to which a client can add custom tags. Of these tags, a client can specify which tags are not acceptable. For instance a client can say that full nudity, hate language, and sexual violence are not allowed.

A client website notifies EthoSafe of new content by POSTing information to a RESTful API. The endpoint of the content notification API is at http://api.ethosafe.com/content/notify. The information that is sent to the API includes the client’s access key, the API version they wish to use, an XML packet describing the content that should be reviewed, and a cryptographic signature used to verify authenticity of the data.

The XML packet that describes the content to be reviewed resembles the following:

```
<contentPackage>
  <identifier>1234567890</identifier>
  <submissionDate>
    2008-01-01T12:00:00Z
  </submissionDate>

  <submittingUser>
    <userId>1</userId>
    <username>testuser</username>
    <emailAddress>test@example.com</emailAddress>
    <ipAddress>127.0.0.1</ipAddress>
  </submittingUser>

  <contentItem mediaType="text">
    <text>This could be an image caption</text>
  </contentItem>

  <contentItem mediaType="image">
    <sourceUrl>
      http://www.example.com/myimage.jpg
    </sourceUrl>
  </contentItem>
</contentPackage>
```

The contentPackage element groups together one or more contentItems for review. The reason for this hierarchy is that a logical unit of content on a client's site may include one or more different types of content, i.e, a blog post that contains an image, or a video that includes a caption. This document uses the word "content" to mean either a contentPackage or contentItem. The identifier element within the contentPackage element is the unique identifier that the client uses to identify the logical unit of content. The EthoSafe platform passes this identifier back to the client with the acceptance or rejection message after the content review is complete. The response sent to the client during a call to the content notification API indicates success or failure of receipt of the content notification request. The response indicating acceptance or rejection of the content is sent asynchronously after the review is complete.

The EthoSafe API passes off the content notification to the central component of the platform, referred to as the Orchestrator. It is responsible for coordinating the content review process and distributed services that implement this process. The majority of the steps the Orchestrator executes are involved in the tagging of a piece of content. The Orchestrator will break a contentPackage apart and analyze each contentItem individually. At the end of the process, the review of all contentItems will be consolidated into a single acceptance or rejection message for the contentPackage.

It is conceptually correct to think of the EthoSafe platform as a tagging engine. Tags are assigned to a piece of content with information describing whether or not the tag applies or does not apply to that piece of content and a confidence interval for the tag. An example tag that is assigned to a piece of content might indicate that there is no partial nudity in the content. EthoSafe has a standard set of tags in the categories of language, nudity, violence, sex, potential harm (e.g. drug references), and spam. In addition to the standard tags, clients have the option of specifying custom tags.

The first step in the review process is that content is imported and, if necessary, transcoded. All video media is transcoded into Flash video and all audio media is transcoded into an MP3.

The second step of the review process is the generation of a unique digital fingerprint for the piece of content being reviewed. The fingerprint is an SHA-512 digest of the text or binary data of the piece of content. The current fingerprint implementation is extremely literal and will not detect such things as cropped images and clipped movies. A more advanced fingerprinting algorithm is being developed.

The digital fingerprint is used in step three. A master index of reviewed content is checked to see if the piece of content in question has been reviewed before. This master index is called the Tag Index (and the "Smart Database" in marketing materials), since it holds all the tags that have previously been generated for a piece of content. At the end of step three, if any tags were retrieved from the index, they are checked against the rules that the client has established to determine whether or not there is sufficient information to accept or reject the piece of content. If there is enough information, then the Orchestrator stops the review process for this piece of content, otherwise is continues.

The Tag Index constitutes a core benefit of the EthoSafe system. Objectionable content, especially multimedia content, tends to get posted to multiple sites. In almost all cases, the EthoSafe platform will recognize the content as having been already reviewed and little analysis will need to be done to the piece of content. Since all content is analyzed against a series of standard tags, the tags generated in a review for one client can be reused during the review for a different client.

The fourth step is analysis by artificial intelligence. The artificial intelligence consists of many small systems, each trained to analyze content for a specific tag. Examples of these systems include a Bayesian classifier that classifies content for hate language or a pattern matcher that checks for profanity. The artificial intelligence step in the review process is the one that consumes the most technical resources. As applicable technologies are made available, they are integrated into this stage. (Currently, the artificial intelligence is more capable of analyzing text than images, audio, and video. As analysis technologies for multimedia become more mature, they will be integrated into the system.) As with step three, the accumulated tags are checked against the rules to determine if the review should continue.

The fifth step in the review process is that a human reviewer looks at the content. This is a necessary step to resolve any ambiguity encountered in the artificial intelligence stage or to accomplish analysis that is not yet possible with artificial intelligence (e.g., analyzing a video for wildlife gore). At the end of the human review stage, the content has been assigned all of the tags necessary for the rules to evaluate an acceptance or rejection message for the content.

The sixth, and penultimate step, is that the Tag Index is updated with any new tags generated during the current review and any pieces of artificial intelligence that is based on learning technology (e.g. Bayesian classifiers) will be trained with the result of the human review.

The final step is that the reviews for each piece of content being reviewed will be logically ANDed together into a single acceptance or rejection message for the contentPackage entire. This message is sent back to the client site.

The message sent back to the client site is sent via HTTP POST. The POST parameters include the identifier initially sent with the call to the content notification API, an indication of acceptance or rejection, and a cryptographic signature used to verify message authenticity.

The amount of time required to review a piece of content varies based on the type of content being reviewed, how deep into the review process the content needs to go before sufficient tags are accumulated, and the utilization of the human review workforce. If a piece of content's publication rules can be satisfied by the Tag Index, the average turnaround time of a piece of content is around 4 seconds. For a piece of text content or an image that needs to be reviewed by a human, the average turnaround is around 20 to 25 seconds. This varies somewhat based on the length of the text and immediate availability of a reviewer. Audio and video content takes longer to review due to the overhead of media transcoding and the limitation that a review takes at least as long as the clip takes to run, due to the necessity that a human needs to watch it at least once.

The EthoSafe system is under active development with new features being released regularly. The next major release will include a customer portal to help give client's better visibility into their content and enhanced QA functionality that takes advantage of periods of underutilization of the human review workforce to re-review content previously reviewed.

Technical Attributes

- Able to analyze text, audio, image, or video content
- Acceptability is determined by a set of client-specified rules
- Integration via a RESTful API

Limitations

- The system does not address such issues as legal compliance of content or copyright infringement.
- The system does review the destination content of a URL placed within a piece of text content.
- The system cannot review Flash content (except video), PDFs, Microsoft Office documents, or anything else that is not plain text

Effectiveness

The effectiveness of the EthoSafe system is measure by spot-checks of reviewed content and by client feedback. The efficacy of the human review workforce is verified by sending test content through the human review step. The tags generated are then compared against known values. Lastly, a more robust QA system is currently in development that uses periods of underutilization of the human review workforce to re-review previously reviewed content.

Implementation Requirements

There are no implementation requirements. The API of the EthoSafe system is technology agnostic. Implementation toolkits, which make integration easier, are provided for .NET, Java, PHP, and Ruby.

The standard implementation does require that the client have control over the codebase of their website. For clients that are hosting a site on a third-party platform the EthoSafe system can retrieve new content via an RSS feed.

Technical Standards

The EthoSafe system uses technical standards heavily, including the following:

- Cryptographic signatures generated using the SHA2 and MD5 algorithms
- XML Schemas
- Standards-compliant HTML and CSS.

Reliance on Use of Law Policy

EthoSafe does not currently rely on the use of law or policy.

Internationalization

EthoSafe is currently built for the US market. Its infrastructure has been prepared for international use, but the workforce has not been put in place and some minor alterations to the technology will be necessary.

EXPERTISE

EthoSafe provides a safety net for companies and organizations that are using social media and its inherent two-way dialogue to protect the brand and its customers from inappropriate content that violates the ethos of the community. Unlike the handful of competitors in this growing industry, EthoSafe understands that the best content moderation combines learning technology and human review to ensure the most complete and accurate content tagging and filtering.

COMPANY OVERVIEW

EthoSafe was started in 2007 by four principals who have each owned and managed internet and/or media companies that created interactive marketing solutions in the B2C and B2B markets as well as non-profits. Using technology to provide wonder as well as efficiencies, the four principals are considered experts in the field based on implementations large and small in a wide variety of industries. Combining business acumen, a deep understanding of the marketing environment and internet-based technology expertise, the founders of EthoSafe bring a strong foundation to this start-up enterprise.

The EthoSafe platform is currently in beta release and in the process of integrating beta customers including direct customers HBO, Warner Brothers and Laughing Cow as well as customers from partners KickApps and EveryZing. The company is currently seeking Series A funding and has engaged with a number of angel groups. This first round of funding is expected to close before the end of 2008.

It is expected that EthoSafe will generate a small revenue stream in 2008, with paid subscriptions beginning in the fall. The financial model of the company includes two rounds of private placement investment by the middle of 2009. The company is expected to be profitable by the end of 2009.

BUSINESS MODEL OVERVIEW

EthoSafe is offered through a Software as a Service (SaaS) model to social media publishers with monthly payment options based on packages of review units (i.e., similar to a web hosting or cell phone usage model). The current pricing model begins at $3,000 a month and increases up to $15,000 a month.

The sales channels will be direct to brand manager and through interactive and ad agencies. EthoSafe will also partner with other Web 2.0 technology players such as white label social network platforms or others enabling user-generated content programs.

Because the EthoSafe package pricing is based on the amount of activity generated on a social media program, the cost for the service is commensurate with the service provided. EthoSafe is able to keep its costs lower than its competitors because of the efficiencies built in to the platform. Smaller companies or non-profits will benefit from the platform's efficiencies at a higher proportional rate than larger companies because content from all EthoSafe subscribers is used to determine the content tagging for all.

MORE INFORMATION

To find additional information about EthoSafe, please go to www.ethosafe.com/isttf. In this directory, you will find the Executive Summary, which is being used for investment conversations; the Datasheet, which is being used for sales conversations; and a copy of the API documentation. Illustrations provided in these documents will benefit your understanding of the EthoSafe solution.

CONTACT INFORMATION

Michelle Chambers, CEO
8 Wallis Court
Lexington, MA 02421
mchambers@ethosafe.com
office: 800 263-4516 ext. 207
mobile: 617 470-0646

CERTIFICATION

I certify that I have read and agree to the terms of the Internet Safety Technical Task Force Intellectual Property Policy.

USE OF THIS DOCUMENT

This document should not contain information that cannot be made available to the public. (See Legal Notice below) This submission will be made available to the Technical Advisory Board, the Task Force, and the Attorneys General. Additionally, after initial review, submissions may be made public and published online for public commentary. Please note that you must be prepared, in any follow-up discussions on your submission with the Task Force, to provide sufficient, non-confidential details and explanation about how your technical solution works and upon what information it relies, in order to allow the Task Force meaningfully to evaluate your solution.

LEGAL NOTICE

The Berkman Center, the Task Force and Task Force members, and the Technical Advisory Board, including its members and affiliates, are under no obligation to maintain the confidentiality of the submitted abstracts or other materials you provide. Please do not submit any information in your technical abstract that is confidential, proprietary or not for public dissemination. Please submit only information that you are willing to have made public. All submissions are subject to the Task Force Intellectual Property Policy: http://cyber.law.harvard.edu/research/isttf/ippolicy. By submitting your abstract or proposal, you certify that you have read and agree to the terms of that Policy.