CASE STUDY

GATHERING A WORLD OF INFORMATION
PUTTING IT IN YOUR HANDS

30 DIGITS WEB EXTRACTOR
PRECISION DATA EXTRACTION
WITH SCALE AND EASE

Linking People to Content at TrustYou

Turning the wealth of information on the Internet into valuable information ready to be analyzed, mashed-up, and repurposed for your needs

Context

Content Revolution

The grandfather of the Internet (ARPANET) began with connecting two institutions in 1969. In ‘88 the Internet entered the commercial world bringing businesses into the Net. The World Wide Web appeared in ‘91 with the first web browser allowing text documents to be easily viewed. Amazon went online in ’95 with its then original review system. By 2006 there was over 146 billion dollars in revenue for products sales online plus 73.5 billion in travel. Today’s buyers want to know what yesterdays customers thought.

Challenge

Acquiring Information that can be Processed

TrustYou has a unique system that can evaluate text and extract out the positive and negative sentiments. It can further categorize them into topics relative to the business need. For instance, it can read a user review on a hotel and tell that the visitor enjoyed his breakfast and thought his bed was too hard resulting in a positive rating for food and drink at the hotel and a negative on room conditions. It does this using the latest semantic and statistical methods in conjunction with top institutions focused on this area.

The challenge though was how to get the vast amount of review data from the Internet. To be of value, they required information on hotels all over the world and their accompanying reviews. This information is scattered across dozens of websites each with potentially millions of pages. These sites are each designed with different structures and styles sometimes even covering areas unrelated to hotels. Once on a page with review data, one still has to deal with a large number of obstacles. Only the review data itself is relevant and each piece of data has to be separately tagged and stored to be used. This data is in no way standardized across those sites with dozens of date formats and no limit to methods of displaying rankings.

TrustYou Profile

TrustYou is the first independent quality search engine. We summarize information from millions of reviews and opinions on the Internet in one single place. Long browsing sessions while looking for a suitable hotel or restaurant are now a thing of the past. TrustYou's proprietary statistical and semantic algorithms extract sentiment information from reviews and comments about hotels or restaurants found on message boards, travel portals, blogs and communities.

The wealth of information generated from millions of reviews is then analyzed and summarized – only the "quality essence" is presented to the user: The most frequently mentioned positive, negative and neutral aspects. Every bit of information can be traced back to its original source and is thus completely verifiable.

TrustYou was founded by search technology veterans in March 2008 and is cooperating with the Center for Information and Language Processing of the Ludwig Maximilians University in Munich ("search guru" Prof. Dr. Franz Guenthner).

1st Place – Most Innovative Travel Start-Up – "Launch Pad 2008" Competition by VIR e. V.

Solution

Combining State of the Art Spidering with Human Pattern Recognition

There are web crawlers, fetches, spiders, connectors and other tools. Some are sold by large corporate companies and others are free to download. They are almost exclusively created to follow every link and strip every piece of text off a page and throw it into a search engine. The Web Extractor was created with a focus on getting only and exactly the information from a site that it needs. This benefits not only applications that require information in fields and columns or structured XML. It also assists search engines by avoiding garbage information and providing valuable meta-data.

The Web Extractor does this through its unique user interface which allows the one configuring to define as precisely or vaguely, as the need may be, which links to follow and which data to retrieve. It guides the user through this step by step to accomplish tasks for a variety of sites that would take a programmer days to write, parse, and format for each different case. The user can do all of this with just a simple knowledge of HTML to recognize links and tags. To define more complex patterns, the user can apply standard regular expressions to virtually any aspect of the configuration accommodating the limitless varieties of website styles and structure.

> "30 Digits has the best spider on the market for extracting quality data from complex sites. They also stay on the edge of web trends assuring the solution continues to work as the web evolves."
>
> Jakob Riegger, CEO
>
> TrustYou GmbH

The Web Extractor has been developed on the foundation of years of experience in the information access industry, web based standards and protocols, and most importantly tried and tested on some of the largest and most complex sites. It scales across processors with threading and across servers with a central administration. It deals with issues as diverse as forms, authentication, cookies, and JavaScript. If you need more precise data; if the spider you are using does not fulfill your requirements; the 30 Digits Web Extractor is the solution.

CASE STUDY

Benefits

Quality Data: the way you want it

30 Digit's Web Extractor retrieves details like the address and number of stars on hundreds of thousands of hotels and the author, text, and scores given on millions of reviews. This is all delivered in clean XML custom for TrustYou's system.

> "Our mission and motto is – *Linking People to Content*. With the Web Extractor we enable our business partners to do just that through supplying the content to power the Internet applications of tomorrow."
>
> Justin Gilbreath, Co-Founder and Managing Director
>
> 30 Digits GmbH

Wealth of the Internet in your hands

- The wealth of review information all over the Internet can now be analyzed for categories and sentiments to help customers make better decisions
- Hotel owners can see how they are viewed and react to the social perception of their products and services

Information delivered in custom format

- Any data on the web page can be turned into a field for analysis
- Data formatted for multiple search engines, custom XML, CSV, TXT, and more
- Definable schedules for crawling

Focus on core business

- No concerns about changing web site standards
- No extra development costs
- Focus on core strengths with more time and money to focus on differentiators

CASE STUDY

Challenges

Getting to the Data

With the variety of structure possible in a website, it is essential to zoom in on the precise information of relevance. Areas which have nothing to do with the topic of interest should be completely avoided to prevent needless crawling and wasted time. Site maps, standard links, and linear paths are not enough to get to data. Forms and searches can be necessary to reach content.

Extracting the Correct Information

Information that is designed for the eye to recognize because of formatting, context, or images can be elusive to machines needing to make the same distinctions consistently without mistake. The patterns behind the diversity have to be understood and definable to be replicable from page to page. Interfaces which are interactive on clicks and other transactions have to be simulated.

Normalizing for Comparison

Diverse data that has no point of reference or scale is useless. Rankings can be represented by numbers, symbols, images, and more. Dates can be in different formats and languages. All of these different kinds of information have to be standardized and conform to have meaning in comparison to one another.

About 30 Digits

3 Founders with over 20 years of experience in Information Management. Intelligent Search Solutions that support the collective knowledge with smart access to data islands is the task of 30 Digits. The goal is direct access to content in connection with one's specific area of interest. The Munich based company delivers products, services, support, and complete solutions out of one hand: for Software Companies, IT-Service Providers, and a growing number of middle and large organizations who wish to turn the information flood into water for the plant that is Knowledge Capital. Nearly all company information is valuable. One just needs to know when and where. Allow your knowledge workers to find their optimal path to the sources of information with 30 Digits. Linking People to Content.

CASE STUDY