

[TechNewsWorld](#) > [Computing](#) > [Data Management](#) | [Read Next Article in Data Management](#) January 28, 2008 12:48:17 PM

Please note that this material is copyright protected. It is illegal to display or reproduce this article without permission for any commercial purpose, including use as marketing or public relations literature. To obtain reprints of this article for authorized use, please call a sales representative at (818) 461-9700 or visit <http://www.ectnews.com/about/reprints/>.

## EXPERT ADVICE

## The Theory and Practice of Secure Data Mining



By David Walling  
TechNewsWorld  
12/21/07 4:00 AM PT


[Back to Online Version](#)  
[E-Mail Article](#)  
[Digg It](#)  
[Reprints](#)

**Data mining isn't always about structured data. Text mining -- or text data mining -- is about comprehending natural language and extracting high quality information from it. Natural languages have structure, too. These structures are generally more complex than a schema, especially one designed for data mining.**

As you read a sentence, its meaning may be clear even before you reach its end. This illustrates our topic. Our minds process text sequentially. As we read, the context presented to us by an author develops in our minds. What precedes clarifies what follows, and vice-versa.

This phenomenon is a result of efficiency. It's how language works. Reducing the number of symbols we use simplifies communication in one sense; but it also forces us to adopt complications like words and grammar. Few of us write with hieroglyphs anymore.


Consequently, we render our thoughts in the form of longer streams of consciousness, like this paragraph. The more reading we do, the better we are at predicting what's looming ahead. Yet, we still must let our author's picture become complete in our mind before we are sure we "get" the meaning.

In data mining , the problem of contextual meaning is lessened when data is structured as it is in a [database](#), where the meaning of a value is implied by its location. Grammar isn't required with a structure like this. The meaning of a series of digits in a phone number column, for example, can be taken for granted. Knowing the meaning of a value allows us to apply rules to the value. We can readily see when data is malformed.

In other words, data cleansing, which is crucial to data mining, becomes possible only once we know what variations are allowable based on context.

### Text Data Mining

Unfortunately, data mining isn't always about structured data. Text mining -- or text data mining -- is about comprehending natural language and extracting high quality information from it. Natural languages have structure, too. These structures are generally more complex than a schema, especially one designed for data mining. Because of these inherent complexities, entire technologies have arisen to extract, parse and analyze text. At the same time, an increasing amount of stored data is becoming subject to privacy measures, especially encryption.

Clearly, data mining operations must access the plain expression of meaning, or plaintext, in order to mine it for useful information. In the case of structured data, the unit subject to encryption may be a relatively small set of symbols, perhaps only a field or row. When data is structured, encryption can be efficiently applied in various dimensions. It should be noted that when text is encrypted, the strength of the encryption might depend on the amount of data being encrypted at one time. For example, AES (advanced encryption standard) and Triple-DES (data encryption standard) symmetric ciphers often use cipher-block chaining techniques to strengthen overall [security](#)  by feeding the output of encrypting one block of data into the next encryption operation, etc., making cryptanalysis more difficult.

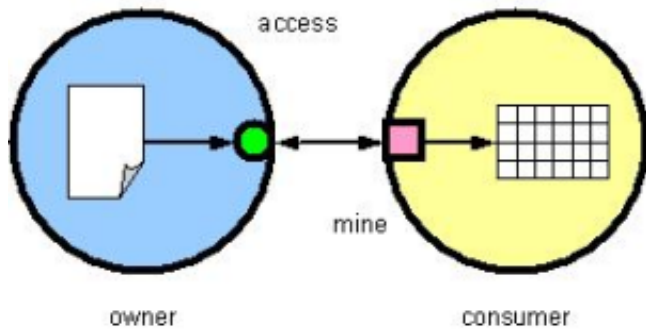
The desired result of encryption is a large mass of bits that provide no contextual reference for the underlying plaintext, which poses a challenge to text data mining. Existing rudimentary approaches to accessing encrypted data include separating the decipherment and mining operations into sequential stages. A negative implication of this approach is that obtaining a sufficiently large text sample for analysis may require exposing too much plaintext for too long a period. Conversely, decrypting too little data may fail to reveal the proper context of the information and lead to flawed analysis.

### **Approaches to Secure Mining**

It may seem strange to contemplate allowing encrypted text to be mined at all. However, text that is valuable for mining isn't necessarily public information.

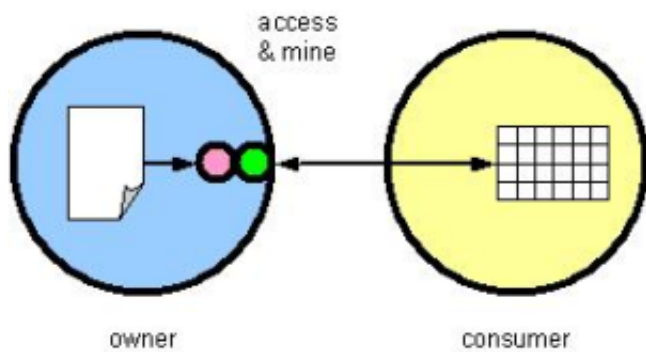
Furthermore, mining text may not necessarily compromise data security, considering that the result of data mining may simply be an aggregation -- or the rules that govern an inference engine or a neural network -- rather than the details of the text itself. Consequently, some form of control is necessary because data is not always mined by its owner.

What emerges from this is the need for an engagement between the interests of the data miner and those of the data owner. In matters of law, fault may be avoided by adhering to the terms of a contract. In IT, it is avoided by adhering to a protocol. Therefore, what remains is to develop both text mining strategies and protocols that efficiently engage streams of encrypted text for data mining without violating security policy.



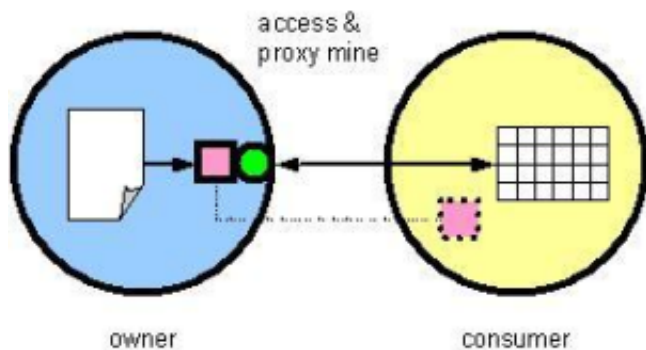
(click on image to enlarge)

We can apply a service-provision metaphor to text data mining by defining the service as either (a) the simple access to the data (fig. 1), or (b) the mining operation itself, which is conducted by the owner on behalf of the mining interest or "consumer" (fig. 2). In the first case, the consumer retains the mining function, perhaps because the consumer's techniques are valued intellectual property. During the mining, the consumer has access to the text in its original form. In the latter case, mining is provided as a service. This simplifies the interface to the data and allows the owner to restrict any view on the data. This approach requires the consumer to trust the mining methods of the owner. The quality of the mining and/or analysis is only as good as the technology to which the owner has subscribed.



(click on image to enlarge)

A third approach (fig. 3) allows the consumer to first provide the data owner with the "method" of mining in the form of a mining object, to which the owner will subject the data. This "middle" approach both protects access to the text and enables the use of the consumer's competitive technology.

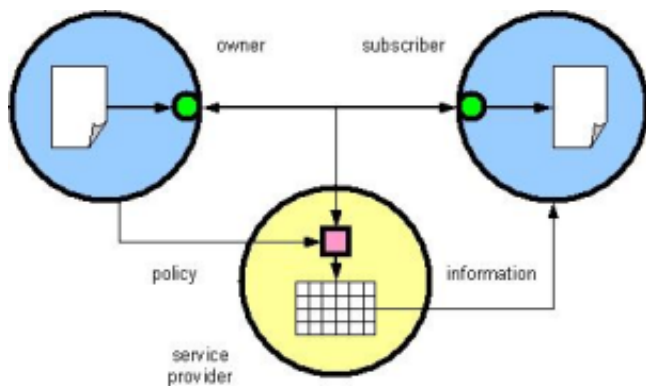


(click on image to enlarge)

To determine the nature of an interface between the consumer and the data owner, we first enumerate the rights of the data owner with respect to data access. This is crucial because we must ensure that the consumer's "methods" do not conflict with the owner's data security policies. For example, the data owner may have the "rights" to:

1. Restrict mining to aggregations (sums, averages) as opposed to allowing specifics (names and numbers)
2. Restrict mining to generalizations ("most," "some," "many") as opposed to direct measures ("maximum," "minimum," "average")
3. Restrict any access to certain data elements, such as identification numbers (SSN, credit-card numbers, etc.) and/or data related to certain groups of individuals such as minors
4. Restrict mining to (or from) certain date ranges

Pursuing such "qualitative" attributes implies not only an ability to symbolize and encode representations of such dimensions, but it also implies a uniformly acceptable process to identify new criteria and extend the protocol dynamically, following the model of the ITU-T X.690 extensible standard for object encoding rules.



(click on image to enlarge)

So far, we've approached text data mining assuming that data being mined always resides in a static "place" at the time it is being mined. An alternative scenario (fig. 4) envisions secure data in transit being subject to mining en route. Secure data transport technologies such as HTTP Applicability Statement 2 support the inclusion of various metadata specifying how the data was "packaged" (i.e. compressed, encrypted, digitally signed, etc). Enabling data in transit to be securely mined can be accomplished by extending this metadata to include the owner's mining "policy" and sufficient technology to enforce the owner's restrictions.

The approaches presented here highlight some emerging challenges facing data and text mining in a technological environment growing increasingly sensitive to security and privacy concerns. **EDT**

---

**David Walling** is CTO of [nuBridges](#), an e-business security provider.

---

▶ **Next Article in Data Management: [Intel Asks Devs to Help Get LANs, SANs to Play Nice](#)**

### **Related Resources**

- [PinPoint - Data Mining, Analysis and Database Marketing](#)
- [Maximizing ROI from Data Mining](#)
- [IBM DB2 Data Warehouse Edition](#)
- [MicroStrategy 8 -- Business Intelligence Software Solutions](#)

---

Copyright © 1998-2008 ECT News Network, Inc. All Rights Reserved. See [Terms of Service](#) and [Privacy Policy](#). [How To Advertise](#).