SEARCHING ENCRYPTED TEXT DATABASES

Michael A. Shepherd
Computing Science Division
Dept. of Mathematics, Statistics & Computing Science
Dalhousie University
Halifax, Nova Scotia, Canada B3H 3J5

Cryptography is the study of secret communication in which messages being transmitted are disguised by making them incomprehensible to anyone except the legitimate recipients. One important means for accomplishing this is through a cryptographic transformation which is a sequence of mathematical procedures selected from a set of mapping functions under the control of a cryptographic key. A cryptographic transformation is applied to the text which is to be protected, called the plaintext. This encrypts the plaintext into a ciphertext which is intended to be unintelligible to any person except the intended recipients. The recipients possessing the cryptographic key can then decrypt the ciphertext back into plaintext. The overall algorithm containing an ecryption and corresponding decryption is called a cryptosystem.

In a similar manner to protecting sensitive messages, databases of sensitive textual material may be stored in an encrypted form for security. However, there may be considerable overhead associated with this security if the data has to be decrypted for searching. In addition to the overhead of decryption, there is an increased risk that security may be compromised during the time that the data is in plaintext for searching. Therefore, it may be preferable to search the data in its encrypted form and only decrypt the data for output to the user.

The research presented in this paper examines three cryptosystems for the searching of encrypted text databases with respect to their time and space complexities. The cryptosystems are the homophonic cipher, the algebraic cipher, and the Data Encryption Standard (DES).

The homophonic cipher is a substitution encryption in which each character of a particular type of plaintext domain (English literature, French literature, chemistry journal articles, etc.) will have more than one representation according to the distribution of the character frequencies. The purpose of the multiple representations is to flatten out the frequency distribution in the ciphertext in order to hide any clues about the plaintext that might be picked up by examining the frequency distribution of the ciphertext characters.

The algebraic cryptosystem is based on four-character encryption and decryption. The plaintext is encrypted by applying matrix multiplication followed by substitutions and transpositions. The algebraic cryptosystem examined in this research transforms each set of four plaintext characters into a 49-bit binary vector.

The Data Encryption Standard was adopted as a standard in 1977. It is based on the idea that an encryption could be the application of a sequence of linear and nonlinear transformations on the plaintext. It consists of 18 transformations from 64-bit blocks of plaintext into 64-bit blocks of ciphertext under the control of a 64-bit cryptographic key.

Each of these cryptosystems will be tested on a number of sample text database of various numbers of document records and the actual costs, in terms of cpu time for encryption and decryption and the resulting database sizes, will be determined and compared. In addition, the time required to search the encrypted databases will be determined and compared to the costs of searching the plaintext databases.

L'interrogation de bases de données de texte chiffrées

Michael A. Shepherd
Computing Science Division
Dept. of Mathematics, Statistics & Computing Science
Dalhousie University
Halifax, NS

La cryptographie est l'étude des communications confidentielles par lesquelles des messages sont camouflés lors de la transmission de manière à les rendre incompréhensibles pour quiconque sauf leur destinataire légitime. Une des méthodes les plus usitées pour atteindre cet objectif consiste à utiliser une transformation cryptographique, laquelle consiste en un enchaînement de procédures mathématiques choisies parmi un ensemble de fonctions de projection, sous la conduite d'une clef cryptographique.

Une transformation cryptographique est appliquée au texte à protéger, appelé texte en clair. Celle-ci transforme le texte en clair en texte chiffré, inintelligible pour quiconque sauf le destinataire. Un destinataire en possession de la clef cryptographique peut ensuite décrypter le texte chiffré en texte en clair. On appelle cryptosystème l'algorithme global qui comprend le chiffrage et le déchiffrage correspondant.

De la même manière que les messages confidentiels sont protégés, les bases de données de texte confidentielles peuvent être, pour des raisons de sécurité, stockées sous forme chiffrée. Un coût administratif important peut être associé à cette sécurité si les données doivent être déchiffrées lors de l'opération de recherche. Il existe aussi, en plus du coût administratif associé au déchiffrage, un risque que la sécurité soit compromise si les données se trouvent sous forme de texte en clair lors de l'opération de recherche. Par conséquent, il se peut qu'il soit préférable d'effectuer la recherche alors que les données sont encore chiffrées, et de n'effectuer le déchiffrage qu'au moment d'afficher les données pour l'usager.

Le travail de recherche présenté dans cette communication examine la complexité, en termes de temps et d'espace, de trois cryptosystèmes utilisés dans l'interrogation de bases de données de texte chiffrées. Ces cryptosystèmes sont le chiffre homophonique, le chiffre algébrique et le Data Encryption Standard (DES, norme de chiffrement américaine).

Le chiffre homophonique est un chiffre par substitution qui, pour chaque caractère de type spécifique de texte en clair (littérature anglaise, littérature françaises, articles de périodiques spécialisés en chimie, etc.) utilise plus d'une représentation, selon la distribution des fréquences de caractère. Le but de ces représentations multiples est d'aplanir la distribution de fréquence dans le texte chiffré, de manière à dissimuler tout indice quant au texte en clair qui puisse être déduit par analyze de la distribution de fréquence des caractères du texte chiffré.

Le cryptosystème algébrique est basé sur le chiffrage et le déchiffrage de blocs de quatre caractères. Le texte en clair est chiffré par multiplication de matrice suivie de substitutions et de transpositions. Le cryptosystème faisant l'objet de cette étude transforme chaque bloc de quatre caractères du texte en clair en un vecteur binaire de 49 bits.

Le Data Encryption Standard a été adopté en 1977. À l'origine de ce standard se trouve la conception qu'un système de chiffre pourrait résulter de l'application d'un enchaînement de transformations linéaires et non-linéaires à un texte en clair. Il comprend 18 transformations de blocs de 64 bits en blocs de 64 bits de texte chiffré sous le contrôle d'une clef chiffrée de 64 bits.

Chacun de ces cryptosystèmes sera testé sur plusieurs échantillons de bases de données de texte contenant un nombre variable de documents. Les coûts réels seront établis et comparés au niveau du temps processeur requis pour le chiffrage et le déchiffrage et à celui de la taille de la base de données résultante. Le coût d'interrogation des bases de données chiffrées sera également établi et comparé à celui des bases de données de texte en clair.