December 1993

An Information Retrieveal Model for Crime Investigation

Jua-Hwang Wang
Central Policy University Taiwan

I-Long Lin
Central Policy University Taiwan

An Information Retrieval Model for Crime Investigation

Jau-Hwang Wang I-Long Lin

Department of Information Management
Central Police University
Taiwan, R.O.C.

ABSTRACT

Very often the suspect of a committed crime is one of those who have committed ones before (especially in narcotics and theft crime). Therefore, the investigation scope of a committed crime can be significantly narrowed down if an information system can be used to determine whether the suspect is one of those who have committed ones before, or the crime is committed by one who has no record. We propose to an information retrieval model which represents each criminal by a set of *profile characteristics* (key words in document retrieval) by parsing each old crime investigation report. The criminal records are then organized, stored, and indexed by the profile characteristics. The characteristics found in the crime scene, called *crime scene characteristics*, can then be matched with the criminal profile characteristics and relevant criminals are retrieved to determine the investigation direction. Mainly, the proposed IR system consists of a parser and a retrieval subsystem. The parser scans the crime investigation reports, extracts the characteristics, and builds a profile for each criminal. The retrieval subsystem takes the characteristics found at a crime scene and the profile database as inputs and retrieves relevant criminals for determining the direction of investigation. Each relevant record is also weighted by a certainty factor.

Key Words: Information Retrieval, Crime Investigation, Criminal Profile, Profile Characteristics, Crime Scene Characteristics, Certainty Factor.

1. Introduction

Very often the suspect of a committed crime is one of those who have committed ones before (especially in narcotics and theft[BRAN84]). Therefore, the investigation scope of a committed crime can be significantly narrowed down if an information system can be used to determine whether the suspect is one of those who have committed ones before, or the crime is committed by one who has no record. The National Police Administration has established the police information system for several years and has accumulated quite large amount of data about criminals. It is mature at this moment to build such an application on top of the criminal data.

The objective of crime investigation is to discover the evidence left in crime scene and then identify the suspect. It is a narrowing process. At the time a detective arrives at a crime scene, he may know nothing about the crime except that a crime has happened. However, each discovery at the crime scene may help the detective to narrow down the investigation scope significantly, and eventually isolate the scope to a few possibilities. The ability of a detective to be able to narrow down the investigation scope often depends on his/her experience, i.e., depends on how much information he/she has gathered in his/her career as a detective. We believe that this information retrieval process can be simulated by a computer system.

Information retrieval (IR) is a discipline involved with the organization, analysis, storage, searching, and dissemination of information. IR systems are designed to make available a given stored collection of information items with the objective of providing reference that would contain the information needed by the users. It has been widely used in document retrieval. Each document is analyzed, represented (usually by a set of key words), stored, and can be retrieved by matching them with the user's query (e.g., also represented as a set of key words). This requires the application of some automatic or manual analyzing technique to the full text or some surrogate (e.g., abstract) of the documents in order to identify the key words to be used in their representation. Each key word (term) may also be associated with a weight factor to reflect its importance as an indicator of the content of a document. The user request may be in the form of a natural language statement or a boolean expression and may be represented as a set of (key word, weight). The documents desired by a user can then be retrieved by matching the query with each set of key words representing each document.

Several mathematics models for document retrieval systems have been developed [CROF82, SALT83, SALT89, TAHA76]. They are used to formally represent the basic characteristics, functional components and the retrieval processes of document retrieval systems. Two basic categories, namely the vector processing model and the boolean retrieval model have been employed in information retrieval. Boolean model has advantage to provide a better structure to formulate the user query. However, it has no provision for associating weight either to the documents or to the queries. The output obtained in response to a query is not ranked in any order of presumed importance to the user's request[WONG86].

Therefore, we propose to build an information retrieval system based on the vector space model information retrieval system for crime investigation. Each criminal is represented by a set of *profile characteristics*[RESS88] (key words in document retrieval) by parsing each crime investigation report. Each characteristic is also weighted with an importance factor. The characteristics found in the crime scene can then be matched with the profile of each criminal to determine the investigation direction (i.e., narrow down the investigation scope).

This paper is organized as following, section 2 describes the vector space model, section 3 gives the system architecture and describes the work to be done, and section 4 gives the conclusion.

2. The Vector Space Model

The vector space model[SALT89] represents both queries and documents by term (i.e., key word) sets and computes global similarities between queries and documents. It assumes that an available term set is used to identified both stored records and information requests[SALT89]. Both queries and documents can be represented as term vectors of the form

$$D_i=(a_{i1},a_{i2},\ldots,a_{it}) \quad (1)$$

and

$$Q_j=(q_{j1},q_{j2},\ldots,q_{jt}) \quad (2)$$

where the coefficients a_{ik} and q_{jk} represent the values of term k in document D_i or query Q_j, respectively. Typically a_{ik} (or q_{jk}) is set equal to 1 when term k appears in document D_i (or in query Q_j), and to 0 when the term is absent from the vector. Alternatively, the vector coefficients could take on numeric values, the size of the coefficient depending on the importance of the term in the respective document or query.

Consider now a situation in which t distinct terms are available to characterize record content. Each of the t terms can then be identified with a term vector T_i and a vector space is defined whenever the T vectors are linearly independent. In such a space,

any vector can be represented as a linear combination of the t term vectors. Hence the rth document D_r can be written as

$$D_r=\sum_{i=1}^{t} a_{ri} T_i \quad (3)$$

where the a_{ri}s are interpreted as the components of D_r along the vector T_i.

The similarity between any two vectors x and y can be measured by the product $x \cdot y = |x||y|\cos\alpha$, where $|x|$ is the length of x and α is the angle between the two vectors. Hence given a document D_r and a query Q_s represented in the form by expression (3), the document-query similarity can be computed as

$$D_r \cdot Q_s = \sum_{i,j=1}^{t} a_{ri} q_{sj} T_i \cdot T_j \quad (4)$$

Computing the similarity values of expression (4) thus depends on a specification of the document and query components, as well as knowledge of the term correlations $T_i \cdot T_j$ for all term pairs. The vector components can be generated by an indexing operation. Furthermore, by assuming that the terms are in fact uncorrelated, i.e., the term vectors are orthogonal, the t term vectors form a proper basis for the vector space. The similarity computation of expression (4) is then reduced to the simple sum-of-products form of expression (5):

$$sim(D_r, Q_s)=\sum_{i,j=1}^{t} a_{ri} q_{sj}. \quad (5)$$

3. System Architecture

Mainly, the proposed IR system consists of a parser and a retrieval subsystem. The parser scans the crime investigation reports, characterizes each criminal, and builds a criminal record database. The system architecture is shown in Figure 1.

Definition 3.1: *crime scene characteristics*[RESS88] are those elements of physical evidence found at the crime scene that may reveal behavioral traits of the suspect. For example, the crime scene of a homicide can include the point of abduction, locations where the victim was held, the murder scene, and the final body location. Examples of the crime scene characteristics may include the use of restraints, manner of death, depersonalization of the victim, possible staging of the crime, and the mount of physical evidence at the crime scene.

Definition 3.2: *profile characteristics*[RESS88] are those variables that identify the offender as an individual and together form a composite picture of the suspect. Profile characteristics are usually determined as a result of analysis of the scene characteristics of old crime investigation reports and can include sex, age, occupation, intelligence, acquaintance with the victim, residence, and mode of transportation.

Basically, the information retrieval system takes the criminal records (e.g., crime investigation reports) and requests (e.g., queries by users) as inputs, and retrieves relevant criminal records in response to the requests. In principle, the retrieval of stored criminal records in answering to requests must be based on determining similarities between queries and the stored items, and then retrieving those items which are sufficiently similar to the corresponding queries.

The stored records and requests may be unstructured text items, and the retrieval decision may depend on the content of the corresponding texts. In these circumstances, a direct comparison between records and queries is inconvenient. Therefore, intermediate steps are needed to transform user's requests into formal query statements and criminal records into formal representations (e.g., index term vectors) before the comparison is actually carried out. A detailed system architecture is shown in Figure 2.

The content analysis and query formulation is performed ahead of the query-record comparisons. The content analysis is the process of characterization (e.g., extracting profile characteristics for each criminal). Examples are automatic indexing -- i.e.,

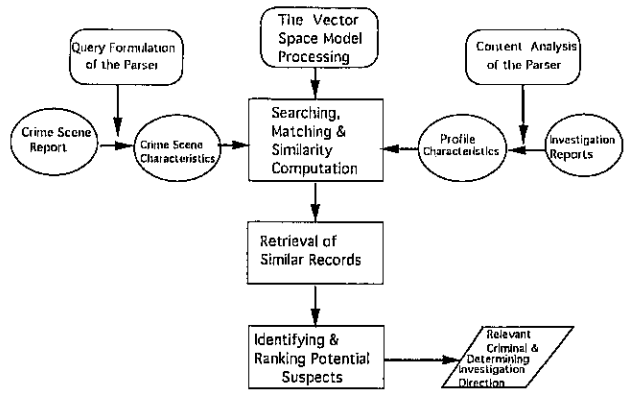

Figure 1. The System Architecture

choosing appropriate descriptors, automatic abstracting, and natural-language understanding. Most automatic or half-automatic indexing systems have "stop lists" of common words that may not be chosen as descriptors. Textual descriptors can be phrases-combined descriptors-or separately assigned words[Bart 85]. Queries could be formulated in natural language, as a set of query

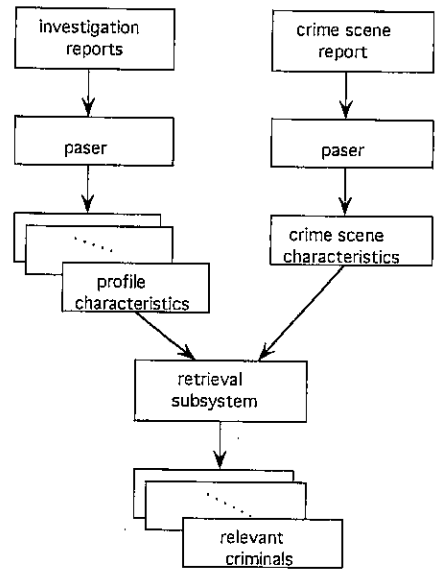

Figure 2. The Overall System Architecture.

descriptors, or as descriptors combined by operators (e.g., Boolean or contextual operators) that follow a certain syntax. A query can be regarded as a virtual item. Query evaluation can be regarded as the process of retrieving relevant data items (e.g., potential suspects). We propose to construct a parser to process the stored records (i.e., crime investigation reports) and represent them by sets of characteristics, that is profile characteristic vectors. Each characteristic may be assigned a weight to reflects its relative importance. Queries (i.e., crime scene investigation reports) may similarly be expressed by using sets of weighted characteristics, called crime scene characteristic vector.

The Parser

As described above, the identification and subsequent retrieval of potential suspects in response to incoming requests depends on the degree of similarity between the suspect's profile and the characteristics found at crime scene. Currently, neither the criminal profiles nor the crime scene characteristics are available. They can be generated by analyzing the old crime investigation reports and the crime scene investigation reports.

The process of constructing surrogates for each documents by assigning identifiers to text items is known as *indexing*. Indexing operations can be performed either by trained persons or by machine. It is proposed that a *parser* should be used to perform the task, i.e., *automatic indexing*, since the volume of text being indexed is large and the indexing operation is repetitive. The parser takes the old crime investigation reports and analyses each record and generates a profile for each criminal. Similarly, the crime scene investigation report is parsed and crime scene characteristics are generated.

In order to generate the criminal profiles characteristic vectors from the crime reports and the crime scene characteristic vector from the crime scene investigation report, a deep analysis of each criminal record, which dealing with meaning and the intent of the items may then be required. However, little success has been achieved in semantic text processing[SALT89]. For practical purpose, content-based information processing is possible only in special circumstance[HAYE83]:

1. When the environment is severely limited and can be represented by a few entities and their relationships,
2. The documents fulfill special functions that automatically place the texts in particular contexts from which most of the usual ambiguities in interpretation are absent.

Therefore, the following approaches are possible alternatives to overcome this difficulty:

1. A basic characteristic database is created manually and used by the parser to analyze the crime reports and generates profile characteristics for each criminal.
2. The investigation reports are required to be written in a specific form such that it is possible to interpret investigation reports automatically.
3. Apply artificial intelligence technology[ICOV86].

All three approaches will be studied and evaluated.

Furthermore, a program will be constructed to scan all the investigation reports, gather the statistics, and determine the weight of each characteristic. The weight of a characteristic is a function of its frequencies in an individual reports and in the remainder of the collections[SALT89].

The outputs of the parser are:

$$C_i=(c_{i1},c_{i2},\ldots,c_{it}) \quad (6)$$

and

$$Q_j=(c_{j1},c_{j2},\ldots,c_{jt}) \quad (7)$$

where each C_i, Q_j, and c_n represent a criminal, a query, and a characteristic respectively.

The Retrieval Subsystem

As described, the query evaluation is to retrieve the relevant stored items. The search operations must effectively identify stored records (e.g., potential suspects). In practice, the operational retrieval environments are characterized by two main requirements:

(1) Access to the files must be carried out more or less instantaneously, usually while users wait at computer terminals.

(2) The number of criminal records to be searched may be very large--of the order of tens of thousands.

The fast-access requirement eliminates sequential searches of the profile database. A good solution is to provide a separate dense index for each characteristic. That is, each characteristic is associated with a list criminal identifiers having the characteristic. The set of indexes for all possible characteristics and the identifiers is collectively known as an inverted index or inverted file [SALT83].

The inverted-index for each characteristic can be obtained by transposing the criminal-characteristic array (e.g., $C_i = (c_{i1}, c_{i2},\ldots c_{it})$). Each characteristic c_i is then associated with a list $(C_1, C_2,\ldots, C_n)$, where C_ks are those criminals who have characteristic c_i.

With an inverted index, the evaluation of a particular query can then be done in three steps, (1) perform set operation on the lists of the characteristics found at crime scene, (2) compute the **dot** product of each C_i (obtained in (1)) and the query Q_j, and (3), output relevant criminals.

4. Conclusions

We have proposed an information retrieval model for crime investigation. The system applies the techniques used in document retrieval based on the observation that the operation of document retrieval is similar to the retrieval of criminal records. Each criminal is represented by a set of profile characteristics by parsing each old crime investigation report. The characteristics found in the crime scene can then be matched with those records in the criminal record database to determine the investigation direction. The proposed IR system consists of a parser and a retrieval subsystem. The parser scans the crime investigation reports, characterizes each criminal, and builds a criminal profile database. The retrieval subsystem takes the criminal profile database and the characteristics found at a crime scene and retrieves relevant criminal records. The various issues addressed in this proposal will be investigated in the years to come.

5. References

[BART 85] Martin Bartschi,"An Overview of Information Retrieval Subjects," IEEE computer, may 1985.

[BRAN 84] Paul and P. Brantingham, *Patterns in Crime*, Macmillan Publishing Company, New York,1984.

[CROF 82] W. Croft, "Experiments with Representation in a Document Retrieval Systems," COINS TR-82-21, University of Massachusetts, 1982.

[EDDI 92] Peter Eddison,"Full-Text Information Retrieval: an Overview," special report IMC journal, 1992.

[HAYE 83] P. J. Hayes, and J. G. Carbonell, "A Tutorial on Techniques and Applications for Natural Language Processing," TR-83-158, Carnegie-Mellon University, 1983.

[ICOV 84] D. J. Icove, "Automated Crime Profiling," FBI Law Enforcement Bulletin, Dec., 1986.

[LANC 79] F.W. Lancaster, *Information Retrieval Systems: Characteristic, Testing and Evaluation*, Second Edition, John Wiley and Sons, New York, 1979.

[RESS 88] R. K. Ressler, A. W. Burgess, and J. E. Douglas, *Sexual Homicide: Patterns and Motives*, Lexington Books, 1988.

[SALT 83] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Co., New York, 1983.

[SALT 89] G. Salton, *Automatic Text Processing: the Transformation, Analysis, Retrieval of Information by Computer*, Addison Wesley,1989.

[TAHA 76] V. Tahani, "A Fuzzy Model of Document Retrieval Systems," *Information Processing & Management*, Vol. 12, 1976.

[WANG 93] J.-H. Wang, and I.-L. Lin, "An Information Retrieval Model for Crime Investigation", to appear in the proceedings of the PACIS Conference, Kao-Shiung, Taiwan, May 30, 1993.

[WONG 86] S. K. M. Wang, W. Ziarko, V. V. Raghavan, and P. C. N. Wong, "On Extending the Vector Space Model for Boolean Query Processing, Proceedings of the 1986-ACM Conference on Research and Development in Information Retrieval, 1986.