Conceptual Data Modeling and Information Retrieval System Design

개념적 데이터 모델링과 정보검색 시스템 디자인

오 삼 균(Sam-Gyun Oh)*

Contents

ABSTRACTS

The purpose of this paper is to show how conceptual data modeling can enhance current information retrieval (IR) systems. The conceptual database design provides for: 1) data mining capability to discover new knowledge based on the relationships between entities; and 2) integrating current separate databases into one IR system (e.g., integrating ISI Citation, a thesaurus, and bibliographic databases into one retrieval system). Further, as new user requirements are unfolded, modifications of IR systems based on conceptual data modeling will be much easier to make than they were in the current IR systems because conceptual modeling facilitates flexible modifications. The enhanced Entity-Relationship (ER) model was employed in this study to develop conceptual schemas of IR data.

초 록

이 논문의 목적은 개념적인 데이터 모델링이 기존의 정보 검색(IR) 시스템을 어떤 식으로 보다 향상시킬 수 있는지를 보여주는 것이다. 개념적인 데이터베이스 디자인은 1)개체들간의 관계에 기반하여 새로운 지식을 발견해 내는 데이터 마이닝 능력과 2)기존의 개별적으로 분리된 데이터베이스를 하나의 정보검색 시스템 안으로의 결합을 위해 사용된다 (예: ISI 인용, 시소러스, 서지 데이터베이스를 하나의 정보검색 시스템 안에 결집시킴). 더 나아가서, 개념적인 모델링은 수정을 용이하게 하므로, 새로운 이용자의 요구가 가미될 때마다, 개념적인 데이터 모델링에 기반한 정보검색 시스템을 수정하는 것은 기존의 정보검색 시스템 상에서보다 훨씬 수월해질 수 있다. 보다 향상된 개체-관계(Entity-Relationship) 모델이 이 논문에서 다룬 정보검색 데이터의 개념적 스키마를 개발하는데 사용되었다.

* Sam-Gyun Oh is an assistant professor in the Department of Library and Information Science at SungKyunKwan University.
접수일자 1999년 11월 1일

1. Introduction

Current IR systems are mainly based on the flat file structure. This creates difficulty in maintaining data integrity in the database and makes it hard to utilize the relationships that exist among many entities associated with text, image or multimedia retrieval environments. Current IR systems as a whole have been particularly inadequate in answering specific queries and, except in the case of full-text database systems, do not provide users with the capability to search factual and relationship information expressed in the documents. Furthermore, what these full-text database systems allow, i.e., the option of searching terms in the documents themselves, can be limiting because the relationship between the terms have not been defined.

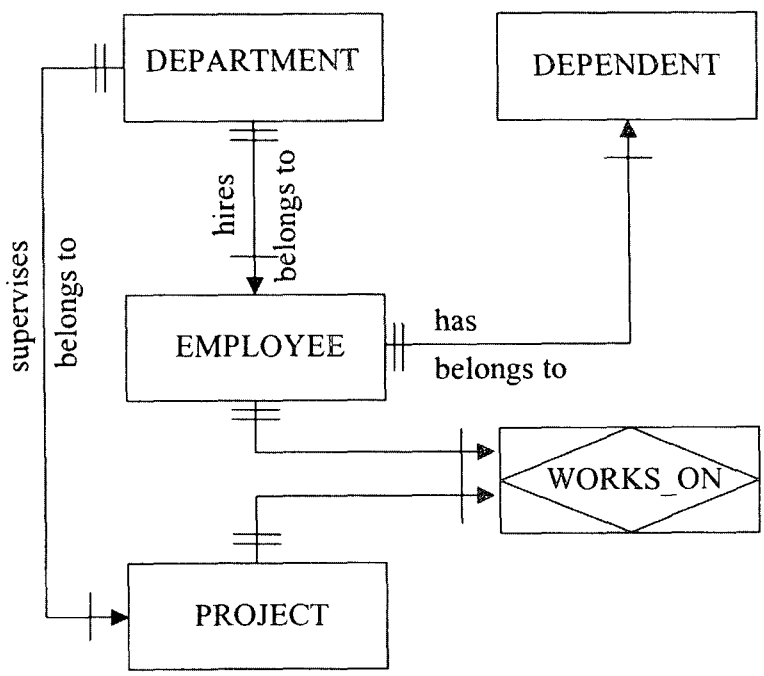
One way to overcome these difficulties is to employ conceptual data modeling when designing a database of any kind. The Entity-Relationship (ER) model is a conceptual data modeling technique that is widely used in database design. This paper explores the benefits of applying ER modeling to IR data to represent factual and relationship information expressed in the documents and to capture other entities and relationships associated with documents.

2. The Entity-Relationship (ER) Model

The ER model (Chen, 1976) is an early semantic data model that unifies features of the traditional models (hierarchical, network and relational) to facilitate the incorporation of semantic information. The enhancement of the model has been continued to date by several researchers (Brachman & McGuinness, 1988; Bachman, 1996; Moody 1996, Martin, Jacobson & Kendall 1997, Muller 1999). As indicated by the name, the two primary modeling constructs are the entity and the relationship. The ER model forms the basis of ER diagrams, which represent conceptual models of databases based on the needs of end users. These diagrams depict the ER model' s three main components: entities, attributes, and relationships. The ER model possesses the requirements of a good semantic data model because it: 1) is expressive enough to point out commonly

occurring distinctions between types of data, relationships, and constraints; 2) is simple and its concepts can be easily understood; 3) has a small number of basic concepts that are distinct and non-overlapping in meaning; and 4) has a diagrammatic notation for displaying a conceptual schema that is easy to interpret (Elmasri & Navathe, 1994).

An entity is a "thing" in the real world with an independent existence. An entity may be an object with a physical existence such as a particular person, car, house, or employee. It may also be an object with a conceptual existence such as a company, a job, or a university course. Each entity has particular properties, called attributes, which describe it. For example, an employee may be described by the employee's name, age, address, and salary. A particular entity will have a value for each of its attributes. A relationship is an association between entities. In the original model described by Chen, relationships do not have attributes. However, a special case arises when connectivity between two entities is many-to-many, or when there is a need to keep track of transactions



〈Figure 2. 1〉 An Example of Entity-Relationship Diagram with Connec-tivity and Cardinality

between the entities involved. A recursive entity is one that consists of a relationship existing between occurrences of the same entity type.

The name of an entity is enclosed by a double quote and the name of a relationship is underlined. A "department" hires one or many "employees", but each "employee" belongs to a "department". A "department" also supervises many "projects", but each "project" belongs to a "department". Each "employee" can have many "dependents", but each "dependent" belongs to an "employee". Each "employee" can work on many "projects" and each "project" can have many "employees" participating in the project.

Now that the basic concepts of the ER model have been discussed, the benefits of applying conceptual data modeling techniques to IR data are explained below.

3. Conceptual Modeling of IR Data

Many authors in the past attempted relational approaches to IR (Macleod, 1981; Dobosz et al, 1981; Rybinski & Szymanski, 1981; Crawford, 1981; Schek, 1982; Stonebraker, 1983; Rybinsik et al, 1985; Blair, 1988, Green, 1996). They found that a relational document retrieval system does not just contain documents, but also a great deal of valuable information of an inferential or tacit nature. Blair (1988), for example, contended that one of the most important facilities of a good document retrieval system is its associative searching capability. This permits the user to discover semantic relationships between the subject index terms that have been assigned to documents on the database. The primary use of associative searching is to semantically broaden a user's subject search. He further argued that relational database management systems (RDBMS) provides easy and flexible access to data and also facilitates managing database.

However, it is important to note that the relational model is an implementation model, which means that it is tied to a particular technology. The relational model also lacks expressive power in terms of describing relationships between entities. What is proposed in this paper is a conceptual data modeling approach to IR, which is independent of implementation models. The conceptual

modeling approach to IR has several advantages:1) It helps us focus our attention on entities, attributes and relationships needed to meet the users' need before we get absorbed in implementation details; 2) it allows us to implement a conceptual model using different implementation models so we can compare them; and 3) it helps us employ a user-centered approach in designing IR systems because the conceptual approach enhances communication between database designers, users, and system designers.

The idea that conceptual models can be beneficial to IR design is not new. Agosti et al (1989) expressed the need for a conceptual paradigm for the design of advanced IR applications, and Ingwersen (1992) also asserted that IR systems should provide users with different ways to access documents because users have various cognitive views. To illustrate how ER modeling can provide this cognitive variety, this paper presents two examples of conceptual schemas to model documents. The first example is a set of conceptual schemas to model empirical facts, research results reported in empirical journals. A system constructed based on these schemas will allow users to search information using empirical variables, statistical relationships and the strength and direction of those relationships. The second example included here is a proposal for conceptual schema specific to journal articles that shows how data modeling can be applied to model the relationships associated with documents, subjects, and authors.

4. Conceptual Modeling of Empirical Facts: the EFR System

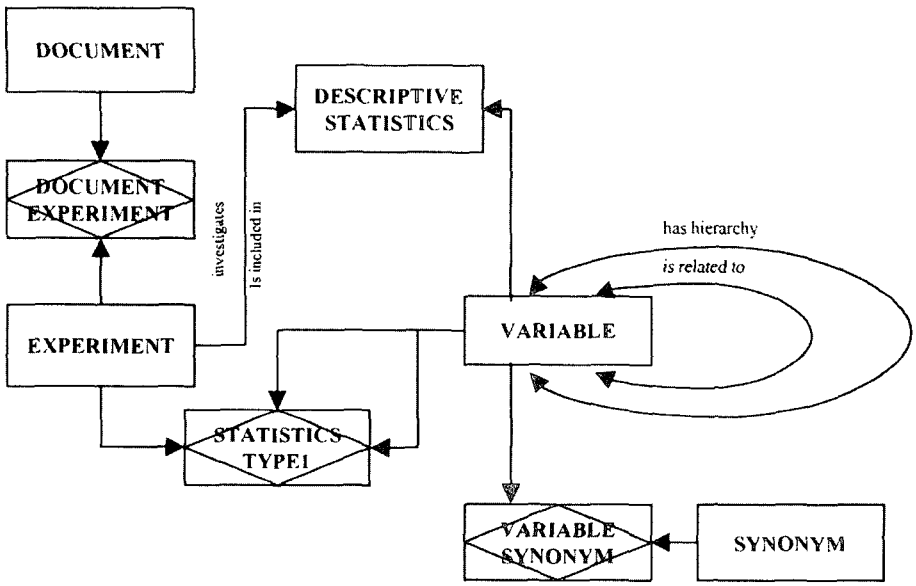
Conventional IR systems that employ isolated term assignments seem inadequate for queries that are specific and empirical in nature. If, on the other hand, retrieval systems provide a link to represent the relationships between the variables of interest as reported in the documents. That is, precision might be enhanced for specific and empirical queries when the relationships between the index terms were specified in retrieval systems.

An empirical fact is defined here as a statement of relationship between two variables. For example, a

variable (contraceptive use) has a moderate correlation with another variable (religious conviction). This is an empirical fact. One empirical fact can be represented by three attributes (variable 1, variable 2 and correlation value). This paper proposes a conceptual model of empirical facts as a way to model the contents of document. Modeling of empirical facts has not been done before. To develop conceptual schemas of empirical facts, they are classified into three groups based on the nature of statistical tests and complexities. Figures 4. 1 through 4. 3 are the conceptual schemas that represent the entities and relationships involved in various statistical tests.

- The first schema (figure 4. 1) represents a situation in which a variable has a statistical relationship with another variable. In other words, given an occurrence of this type of statistical relationship, there are only two variables associated with that particular instance (descriptive statistics or correlation). The first schema captures this relationship and models statistical analyses such as studies that investigate correlation between variables.
- The second schema (figure 4. 2) represents a situation in which a variable has a statistical relationship with a set of variables. Given an occurrence of this type of statistical relationship, there can be a statistical relationship between a dependent variable and a set of independent variables. Statistical tests such as ANOVA and regression fall into this category.
- The third schema (figure 4. 3) represents a situation in which a set of variables has a statistical relationship with another set of variables. It means that given an occurrence of this type of statistical relationship, there are two sets of variables: a set of dependent variables and a set of independent variables. The statistical test such as multivariate regression falls into this category.
- The last schema (figure 4. 4) combines and integrates all these schemas into one conceptual schema to represent a wide range of statistical analyses.

Each of these schemas is discussed below.



〈Figure 4. 1〉 A Conceptual Schema of Correlation Type Statistics

4. 1 A Conceptual Schema of Descriptive and Correlation Type Statistics

The ER diagram in Figure 4. 1 is a conceptual schema to map empirical facts that report either descriptive statistics or correlation type studies.

To explain the above schema, we start with the entity "document". The relationship type is specified either below or above the line and the entity type is enclosed within a double quote. A "document" may report more than one "experiment". The connectivity between them is many-to-many because a "document" may report more than one "experiment" and an "experiment" can be reported in more than one "document". This type of connectivity requires a composite entity ("document-experiment") to keep track of where experiments are reported. An "experiment" reports descriptive statistics about many "variables" and a "variable" can be reported in many "experiments". The connectivity between them is many-to-many, which requires a composite entity ("descriptive-statistics") to keep track of where the descriptive statistics of "variables" are reported. A "variable" has a correlation with other "variables". This is a recursive relationship. The connectivity in this relationship is

many-to-many, which requires a composite entity ("correlation").

The "variable" entity also has three recursive relationships with itself: 1) a "variable" is related to other "variables"; a "variable" can be broader than other "variables"; and a "variable" can be narrower than other "variables". The connectivity between them is many-to-many, which requires a composite entity for each relationship. Actual drawing of the composite entity is omitted if the relationship is recursive. A "variable" can have many "synonyms" and each "synonym" can also be associated with more than one "variable". Other entities associated with the document (e.g., "author", "descriptor") are omitted here because the focus of this conceptual schema is on modeling empirical facts reported in the document.

Logical Schema: The above conceptual schema can be translated into various logical schemas (relational, hierarchical, network, or object-oriented). An example of relational logical schema is presented in this paper. The following is the outline of an algorithm that can map an ER schema into the corresponding relational database schema (Elmasri & Navathe, 1994).

- For each regular entity type E in the ER schema, create a relation R that includes all the simple attributes of E. For a composite attribute, include only the simple component attributes. Choose one of the key attributes of E as primary key for R. If the chosen key of E is composite, then the set of simple attributes that form it will together form the primary key of R.
- For each weak entity type W in the ER schema with owner entity E, create a relation R and include all simple attributes (or simple components of composite attributes) of W as attributes of R. In addition, include as foreign key attributes of R the primary key attribute(s) of the relation that corresponds to the owner entity type E; this takes care of the identifying relationship type of W. The primary key of R is the combination of the primary key of the owner and the partial key of the weak entity type W.
- For each binary 1:1 relationship type R in the ER schema, identify the relations S and T that correspond to the entity types participating in R. Choose

one of the relations for example, S, and include as the foreign key in S the primary key of T. It is better to choose an entity type with total participation in R in the role of S. We include all the simple attributes (or simple components of a composite attribute) of the 1:1 relationship type R as attributes of S.

- For each regular (non-weak) binary 1:N relationship type R, identify the relation S that represents the participating entity type at the N-side of the relationship type. Include as foreign key in S the primary key of the relation T that represents the other entity type participating in R; this is because each entity instance on the N-side is related to at most one entity instance on the 1-side of the relationship type.
- For each binary M:N relationship type R, create a new relation S to represent R. Include as the foreign key attributes in S the primary keys of the relations that attributes of the M:N relationship type (or simple components of composite attributes) as attributes of S.
- For each multi-valued attribute A, create a new relation R that includes an attribute corresponding to A plus the primary key attribute K of the relation that represents the entity type or relationship type that has A as an attribute. The primary key of R is then the combination of A and K. If the multi-valued attribute is composite, include its simple components.
- For each n-ary relationship type R, $n>2$, create a new relation S to represent R. Include as foreign key attributes in S the primary keys of the relations that represent the participating entity types. Also include any simple attributes of the n-ary relationship type (or simple components of composite attributes) as attributes of S. The primary key of S is usually a combination of all the foreign keys that reference the relations representing the participating entity types. However, if the participation constraint (min, max) of one of the entities type E participating in R has max = 1, then the primary key of S can be the single foreign key attribute that references the relation E'

〈Table 4. 1〉 A Logical Schema of Descriptive Statistics and Correlation Type StatisticsLegend: Primary Key is Underlined

Relation	Attributes
Document	Dno, DocTitle, Year, PubType, Source, Language, Identifier, Abstract, HardData?
Experiment	Eno, Experiment Purpose, Experiment Procedures, Research Design, Sample Size
Document Experiment	Dno, Eno
Variable	Variable, Definition
Synonym	Synonym
Variable Synonym	Variable, Synonym
Related	Variable, RelatedVariable
Broad	Variable, BroadVariable
Narrow	Variable, NarrowVariable
Descriptive Statistics	Dno, Eno, Variable, Mean, Median, Standard Deviation
Statistics Type1	Dno, Eno, Variable1, Variable2, Statistical Test, Statistical Value, Significance Level

corresponding to E; this is because in this case each entity e in E will participate in at most one relationship instance of R and can hence uniquely identify that relationship instance.

The logical schema that corresponds to the above conceptual schema in Figure 4. 1 presented in Table 4. 1.

The attributes (RelatedVariable, BroadVariable, NarrowVariable, Variable1, and Variable2) specified in Table 4. 1 are foreign keys which get their value from the "VARIABLE" authority table.

The Empirical Fact Retrieval (EFR) system, a prototype IR system based on the above conceptual and logical schema, was actually constructed, with an experiment demonstrating its efficiency and effectiveness. It can provide the following type of search capabilities:

- One can find studies that investigated a particular variable and further qualify the search by the value of mean, median, or standard deviation, if needed.
- One can also find all the variables investigated in association with a particular variable and further qualify the search by either the direction of relationship or the significance level, if needed.
- One can also find all documents that investigated relationships between two variables and further qualify the search either by the direction of the

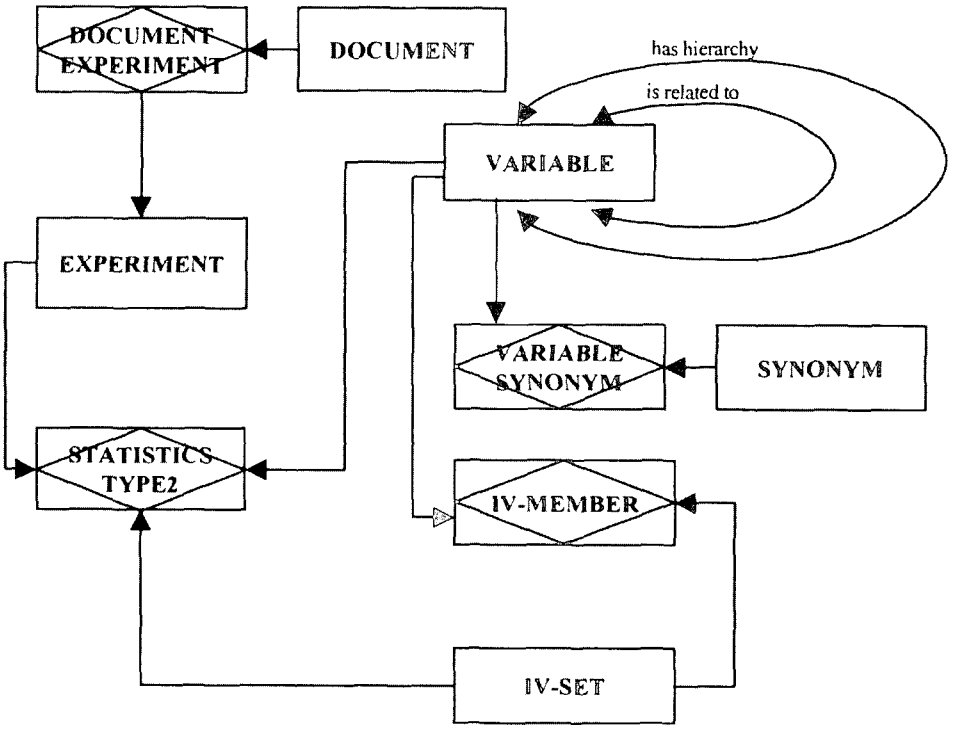
relationship or the significance level, if needed.

- One can also expand a search by including variable synonyms or related variables.
- One can also expand a search using broader variables or restrict a search using narrower variables.

4. 2 A Conceptual Schema of ANOVA or Regression Type Statistics

The following ER diagram maps empirical facts that are the results of analyses of either variance or regression analysis. In ANOVA and regression analysis, a dependent variable is associated with a set of independent variables. In other words, statistical tests are carried out to see what impacts the independent variables have on a dependent variable.

The schema in Figure 4. 2 maps ANOVA and regression type of statistics. There are three new entities introduced in the figure: "STATISTICS-TYPE2", "IV-SET" and "IV-MEMBER". A set of independent variables is grouped as an "IV-SET". The composite entity



〈Figure 4. 2〉 A Conceptual Schema of ANOVA or Regression Type Statistics

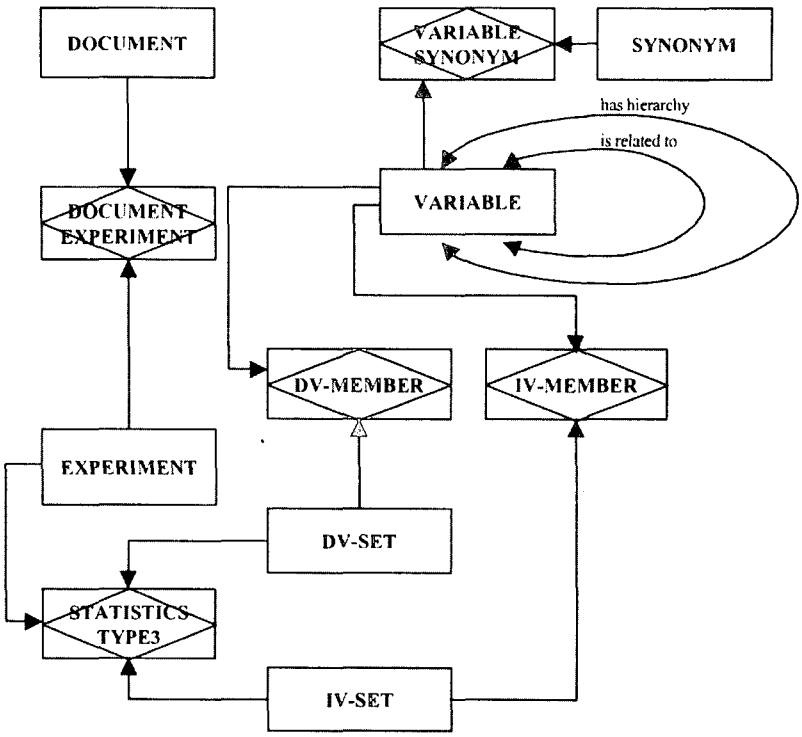
"IV-MEMBER" records individual variables that belong to an "IV-SET". Another composite entity, "STATISTICS-TYPE2," stores statistics between one variable (a dependent variable) and a set of variables (independent variables). A single variable recorded in the "STATISTICS-TYPE2" entity is automatically considered as a dependent variable. A dependent variable is associated with one to many independent variables. There is a many-to-many connectivity between the "VARIABLE" entity and the "IV-SET" entity. The composite entity "IV-MEMBER" records which variable is associated with which "IV-SET" and the other composite entity, "STATISTICS-TYPE2," keeps track of each statistical test performed between a particular dependent variable and a set of independent variables. In the ANOVA statistics, a dependent variable is associated with zero to many main and interaction effects, treated as independent variables. The main and interaction effects are grouped as the "IV-SET" entity. All other entities and relationships in this figure are the same as in the previous figure.

The logical schema that corresponds to the above conceptual schema (Figure 4. 2) is presented below (Table 4. 2).

If an IR system is constructed based on the above conceptual and

〈Table 4. 2〉 A Logical Schema of ANOVA/Regression Type StatisticsLegend: Primary Key is Underlined

Relation	Attributes
Document	Dno, DocTitle, Year, PubType, Source, Language, Identifier, Abstract, HardData?
Experiment	Eno, Experiment Purpose, Experiment Procedures, Research Design, Sample Size
Document Experiment	Dno, Eno
Variable	Variable, Definition
Synonym	Synonym
VariableSynonym	Variable, Synonym
Related	Variable, RelatedVariable
Broad	Variable, BroadVariable
Narrow	Variable, NarrowVariable
IV-Set	IvsetNo
IV-Member	IvsetNo, Ivariable
Statistics Type2	Dno, Eno, Variable, IVsetNo, Statistical Test, Statistical Value, Significance Level



〈Figure 4. 3〉 A Conceptual Schema of Multivariate Type Statistics

logical schema, the following types of queries can be answered:

- Given a particular dependent variable, one can find all the independent variables found to have a statistically significant impact on that variable. The search can be specified further by setting a significance level.
- Given a particular dependent variable, one can also find all the main or interaction effects found to have a statistically significant impact on that variable. The search can be specified further by setting a significance level.

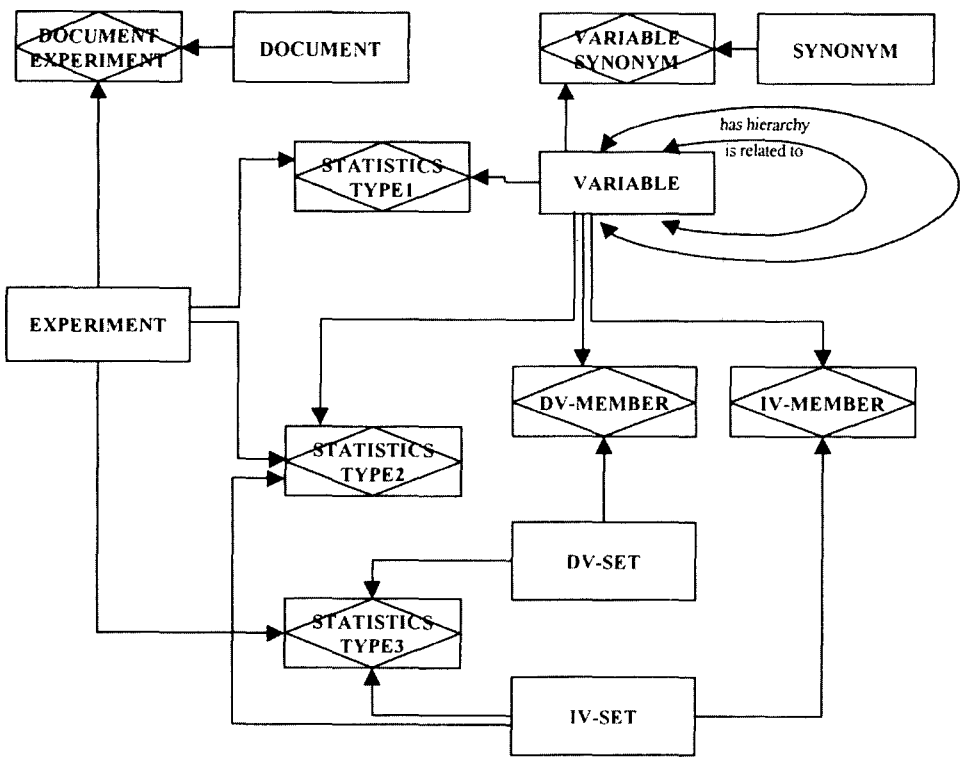
4. 3 A Conceptual Schema of Multivariate Type Statistics

The most complex type of statistical analysis is to investigate how a set of multiple dependent variables is affected by multiple independent variables. Multivariate analysis fits into this category, and is described in Figure 4. 3.

Conceptual schema explained:

〈Table 4. 3〉 A Logical Schema of Multivariate Type StatisticsLegend: Primary Key is Underlined

Relation	Attributes
Document	Dno, DocTitle, Year, PubType, Source, Language, Identifier, Abstract, HardData?
Experiment	Eno, Experiment Purpose, Experiment Procedures, Research Design, Sample Size
DocumentExperiment	Dno, Eno
Variable	Variable, Definition
Synonym	Synonym
VariableSynonym	Variable, Synonym
Related	Variable, RelatedVariable
Broad	Variable, BroadVariable
Narrow	Variable, NarrowVariable
IV-Member	IVsetNo, Ivariable
DV-Set	DvsetNo
DV-Member	DVsetNo, Dvariable
Statistics Type3	Dno, Eno, Variable, IvsetNo, DVsetNo, Statistical Test, Statistical Value, Significance Level



〈Figure 4. 4〉 An Integrated Conceptual Schema of Empirical Facts

The schema in Figure 3. 3 maps all the statistical tests that are similar to multivariate regression analyses. With complex statistics, there are relationships between a set of dependent variables and a set of independent variables. To map this situation, each set of dependent and independent variables is grouped as a "DV-SET" and "IV-SET" entity respectively. There is a many-to-many connectivity between these SET entities and the "VARIABLE" entity, and the " IV-MEMBER" and " DV-MEMBER" entities keep track of all the variables that belong to each SET entity. The composite entity " STATISTICS-TYPE3" records all the statistical values related to the occurrences of these SET entities.

The logical schema that corresponds to the above conceptual schema (Figure 4. 3) is presented in Table 4. 3.

If an IR system is constructed based on the above conceptual and logical schema, the following type of queries can be answered:

- Given a set of dependent variables, one can find sets of independent variables found to have statistically significant impact on those variables. The search can further be specified by setting a significance level.

4. 4 An Integrated Conceptual Schema of Empirical Facts

The conceptual schema in Figure 4. 4 integrates the previous conceptual schemata into one.

The logical schema that corresponds to this integrated conceptual schema is presented in Table 4. 4.

In summary, if an IR system is constructed based on this conceptual and logical schema, all the types of queries mentioned above can be answered.

- One can find studies that investigated a particular variable and further qualify the search by the value of mean, median, or standard deviation, if needed.
- One can also find all the variables investigated in association with a particular variable and further qualify the search either by the direction of relationship or the significance level, if needed.
- One can also find all documents that investigated relationships between two variables and further qualify the search either by the direction of the relationship or the significance

〈Table 4. 4〉 An Integrated Logical Schema of Empirical FactsLegend: Primary Key is Underlined

Relation	Attributes
Document	Dno, DocTitle, Year, PubType, Source, Language, Identifier, Abstract, HardData?
Experiment	Eno, Experiment Purpose, Experiment Procedures, Research Design, Sample Size
DocumentExperiment	Dno, Eno
Variable	Variable, Definition
Synonym	Synonym
VariableSynonym	Variable, Synonym
Related	Variable, RelatedVariable
Broad	Variable, BroadVariable
Narrow	Variable, NarrowVariable
IV-Set	IvsetNo
IV-Member	IVsetNo, Ivariable
DV-Set	DvsetNo
DV-MemberD	VsetNo, Dvariable
Descriptive Statistics	Dno, Eno, Variable, Mean, Median, Standard Deviation
Statistics Type 1	Dno, Eno, Variable1, Variable2, Statistical Test, Statistical Value, Significance Level
Statistics Type 2	Dno, Eno, Variable, IVsetNo, Statistical Test, Statistical Value, Significance Level
Statistics Type3	Dno, Eno, Variable, IVsetNo, DVsetNo, Statistical Test, Statistical Value, Significance Level

level, if needed.

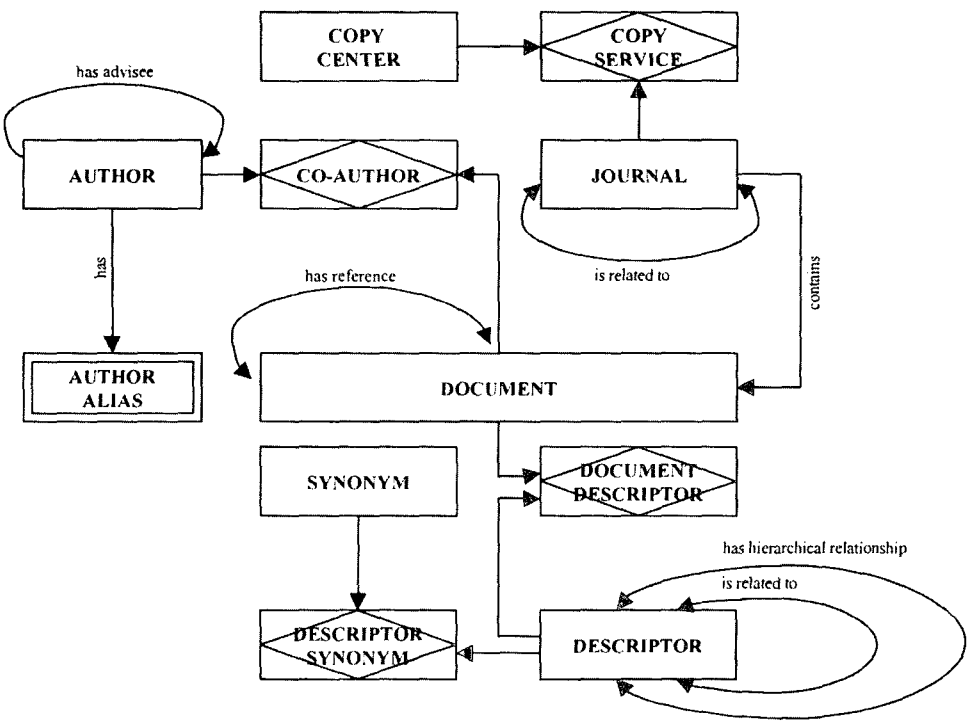
- One can also expand a search by including variable synonyms or related variables.
- One can also expand a search using broader variables or restrict a search using narrower variables.
- Given a particular dependent variable, one can also find all the independent variables found to have a statistically significant impact on that variable. The search can be specified further by setting a significance level.
- Given a particular dependent variable, one can also find all the main or interaction effects found to have a statistically significant impact on that variable. The search can be specified further by setting a significance level.
- Given a set of dependent variables, one can also find sets of independent variables found to have a statistically significant impact on those variables. The search can be specified further by setting a significance level.

As mentioned in the previous section, the main purpose of this paper is to present and discuss the benefits of conceptual modeling of IR data. The first example is to develop a conceptual schema of empirical facts, research results reported in empirical journals. The second example is to develop a conceptual schema of bibliographic relationships specific to journal articles, which shows how data modeling can be applied to model the relationships associated with documents, subjects, and authors.

So far, the discussion has been focused on modeling statistical results and benefits of modeling those data for the retrieval effectiveness. However, the following section describes how data modeling can also be applied to develop a conceptual schema of bibliographic relationships in order to improve retrieval effectiveness.

5. Conceptual Modeling of Bibliographic Relationships

To show how the ER model can be



〈Figure 5. 1〉 A Conceptual Model of Bibliographic Relationships Manifested in Journal Articles

used for other applications, a conceptual schema of biographic relationships manifested in journal articles (Figure 5. 1). This ER diagram depicts the entities and relationships associated with journal articles and their authors.

The ER diagram in Figure 5. 1 maps the major attributes, entities and relationships related to journal articles and their authors. To explain the schema, we again discuss connec-tivity and cardinality between entities, and we specify them from the perspective of the entire life cycle of the database being con-structed. Let's examine the entity "AUTHOR" first. The entity "AUTHOR" has an advisee relationship with itself. For example, an "AUTHOR" may have many advisees, but an advisee can have only one advisor. This is a hierarchical recursive relationship. In hard sciences, valuable information can be found by tracing advisor - advisee relationships because an advisee tends to work in the area that is close to his or her advisor. This conceptual model will allow one to search that kind of information. This model also keeps track of names that "AUTHORs" have written under so that a search can be expanded to include all the author aliases with the help of the entity " AUTHOR-ALIAS".

As for the entity "DOCUMENT", a "document" can be written by more than one "author" and an "author" may publish more than one "document" in the life of the database. A "DOCUMENT" also cites many "DOCUMENTs" and a part-icular "DOCUMENT" can be cited by many "DOCUMENTs". The Institute of Scientific Information (ISI) currently maintains this information, but it is not incorporated with other types of value-adding information (e.g., controlled vocabulary, etc.) in bibliographic databases, which makes the ISI database less effective. Using this conceptual schema, a direct connection between the ISI databases and bibliographic databases can be established. The entity " DOCUMENT" also has a relationship with the entity "DESCRIPTOR". A "DOCUMENT" is represented by one or more " DESCRIPTORs" and a " DESCRIPTOR" can be assigned to many "DOCUMENTs". The entity "DOCUMENT" is related to the entity "DESCRIPTOR" through the composite entity " DOCUMENT-DESCRIPTOR".

The entity "DESCRIPTOR" has many recursive relationships: 1) a

〈Table 5. 1〉 Logical Schema of IR Data Model of Journal Articles and its AuthorsLegend: Primary Key is Underlined

Relation	Attributes
Document	Dno, Title, Year, PubType, Issn, Source, Language, Identifier, Abstract, HardData?
Reference Trace	SourceDno, ReferencedDno
Author	AuNo, Name, Email, Web, Phone, Fax, NameQualifier, AdvisorNo
Co-Author	Dno, AuNo, Position
Author Alias	AuNo, AliasName
Journal	Issn, JnTitle, JnStartYear
Related Journal	Issn, RelatedIssn
Copy Center	CenterId, Name, Email, Phone, Fax, Web
Copy Service	CenterId, Issn
Descriptor	Descriptor
Related	Descriptor, RelatedTerm
Broad	Descriptor, BroadTerm
Narrow	Descriptor, NarrowTerm
Synonym	Synonym
DescriptorSynonym	Descriptor, Synonym
Document Descriptor	Dno, Descriptor, Weight

"DESCRIPTOR" has many "SYNONYMs" and a "SYNONYM" can also be associated with many "DESCRIPTORs"; 2) a "DESCRIPTOR" is associated with many related terms; and 3) a "DESCRIPTOR" is associated with many broad and narrow terms.

The above is the logical schema (Table 5. 1) that corresponds to the conceptual schema presented in figure 5. 1.

5. 1 Added Search Capabilities by Modeling of Bibliographic Relationships

If an IR system is implemented based on the above conceptual and logical schema, the following new capabilities can be added to the current IR systems:

A user can:

- broaden a search by including the articles written by co-authors.
- broaden a search by including an author's aliases.
- expand a search by including the works by the advisor and advisees.
- broaden a search by including synonyms associated with a descriptor.
- broaden a search by including

related terms associated with descriptors.

- be assisted in broadening a search by utilizing the recursive-relationships among descriptors.
- be assisted in narrowing a search by utilizing the recursive-relationships among descriptors.
- search the ISI database using descriptors and expand the search result using other entities related to the descriptor entity.
- easily find where to order the copy of articles if relevant citations are found.

6. Discussion and Conclusion

The advantages of ER modeling for information retrieval can be summarized as follows:

- **Expanding Search Functionality**: Because an IR system based on the ER schema can integrate the functions of thesauri and citation searching, users can easily expand their search in ways that were not available before. Users can start with a subject search, find an interesting article, then move on to a citation search to see who cited that work or whom it cited without changing databases. It is also easy to create either a subject or keyword list that users can browse instead of typing. This kind of access is particularly useful when users have poorly defined search requests.
- **Collocating Related Items**: The ER model can also enhance cataloging. One objective of cataloging is to relate and display together (collocate) the editions which a library has of a given work and the works which it has of a given author (Lubetzky, 1960). Collocating different editions, translations of the same work, and different works based on the same work becomes much easier if a system is constructed based on the ER schema because the main strength of the ER model is the capability of capturing the relationships that exist between entities in a succinct manner. The relationship between a work and all the different editions, translations, and variations (artistic renderings of the work, etc.) can be internally captured so that collocating or grouping them to satisfy the second

purpose of cataloging (collocating related items) can be better accommodated. The ER model is a good candidate to fulfill the objectives of cataloging.

In addition, the EFR model can also be used to model the contents of documents, as the example in this article has demonstrated. For example, we all know that Milton wrote 'Paradise Lost'. This work has many editions, many translations into other languages, and many audio and video renditions. All these can be collocated and grouped by their characteristics if an IR system is based on an ER schema. The current MARC structure makes it difficult to accommodate different groupings for useful displays because of its inadequacy in defining relationships between entities. If we carefully define the relationships between entities and implement them, grouping and collocating related items can be effectively and efficiently accomplished.

- **Flexibility and Modularity**: The ER model increases the flexibility of system updating and maintenance because of its modularity. If additional entities or relationships are identified after a retrieval system is built, those can be easily integrated into the existing conceptual schema and system. For example, the EFR system described earlier did not implement any controlled vocabulary for empirical variables. Considering that researchers may use a variety of different expressions to describe the same variable concept, it could be important to provide users with a controlled list of empirical variables in each empirical discipline. Provision of a controlled vocabulary of empirical variables is likely to improve retrieval. Including this kind of new feature is relatively easy if a system is developed using an ER schema because you can add a new table to include a controlled vocabulary of variables and link them with the variables that were already entered in the database.
- **Enhancing Communication**: One of the most difficult problems of database design is the fact that designers, programmers, and users tend to see data in different terms. Different views of data are likely to

lead to database designs that do not reflect users' information needs. To solve these problems, communication among database designers, programmers, and users should be as free of ambiguities as possible. Data modeling helps the process of communication by reducing difficult real world complexities to more easily understood abstractions that define entities and the relations among them. The ER model, the most common and well-known data model, has been considered the best tool in designing industry databases. In addition, ER modeling, if it is applied to information retrieval functions such as cataloging, will enhance communication among participants because it has rich set of diagramming notations to represent relationships that exist in the bibliographic records.

In conclusion, the ER model is a powerful tool to represent all the entities and relationships related to IR data: 1) modeling external entities and relationships (documents and authors); and 2) modeling internal entities and relationships (empirical facts). The ER model has sound and rigorous design principles. The use of conceptual data modeling in IR will help us to build user-centered IR systems and to design effective IR systems because we can take better advantage of the efforts made by IR researchers. There are numerous advantages if we employ the ER model in mapping IR data. More rigorous research is highly recommended.

References

Agosti, M. (1989). "Towards data modeling in information retrieval." *Journal of Information Science*, 15: 307-319.

Bhatia, S.K., Deogun, J.S. & Raghavan, V.V. (1995). "Conceptual query formulation and retrieval." Journal of Intelligent Information Systems: *Integrating Artificial Intelligence and Database Technologies*. 5(3): 183-209.

Blair, D.C. (1988). "An extended relational document retrieval model." *Information Processing & Management*, 24(3), 349-371.

Brachman, R.J. & McGuinness, D.L. (1988). "Knowledge representation, connectionism, and conceptual retrieval." Proceedings of the 11th *International ACM SIGIR Conference on Research and Development in Information Retrieval*, Grenoble, France, 161-174.

Bachman, C.W. (1996). "Impact of object-oriented thinking on ER modeling." *Entity-relationship approach, ER '96: 15th International Conference on the Entity-Relationship Approach*, Manchester, United Kingdom, 1-4.

Crawford, R.G. (1981). "The relational Model in Information retrieval." *Journal of American Society for Information Science*, 32(1): 51-64.

Dobosz, J., & Szymanski, B. (1981). "An implementation of relational interface to an information retrieval system." *Information Systems*, 6(3), 219-228.

Elmasri, R. & Navathe, S.B. (1994). *Fundamentals of database systems*. (2nd ed.). Redwood City, CA: The Benjamin/ Cimmings Publishing Company, Inc.

Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

Maclead, I.A. (1981). "A data base management system for document retrieval applications." *Information Systems*, 4(2): 131-137.

Martin, F, Jacobson, I & Kendall, S. (1997). *UML distilled: Applying the standard object modeling language*. Reading, MA: Addison-Wesley.

Moody, D. (1996). "Graphical entity relationship models: Towards a more user understanding representation of data." *Entity-relationship approach, ER '96: 15th International Conference on the Entity-Relationship Approach*, Manchester, United Kingdom, 227-244.

Muller, R.J. (1999). *Database design for smarties: Using UML for data modeling.* San Francisco: Morgan Kaufmann.

Oh, S.G. (1996). An empirical fact retrieval system: An entity-relationship and relational approach. Ph.D. Dissertation -- Syracuse University.

Ozkarahan, E. & Can, F. (1991). "Multi-media document representation and retrieval." *ACM Computer Science Conference: Preparing for the 21st Century. Proceedings.* San Antonio, TX USA. PP 420-9. March 1991.

Rob, P. & Coronel, C. (1995). *Database systems: Design, implementation, and management.* Boyd & Fraser.

Rybinski, H. & Szymanski, B. (1981). "Multilevel information system - towards more flexible information retrieval systems." *Information Processing & Management*, 17(5), 277-290.

Rzeczkowski, W. & Subieta, K. (1985). "Stored queries - a data organization for query optimization." *Data and Knowledge Engineering* 3, 29-48.

Schek, H.J & Pistor, P. (1982) "Data structures for an integrated data base management and information retrieval system." *In Proceedings of the Eight International Conference on Very Large Data Bases*, pp. 197-207, Mexico City.

Stonebraker, M. (1983). "Document processing in a relational data base system." *ACM Transaction on Office Information Systems*, 1, 143-158.