

SPIRE/EPI-SPIRE Model-Based Multi-modal Information Retrieval from Large Archives

Chung-Sheng Li, Yuan-Chi Chang, John R. Smith and Matt Hill
IBM T. J. Watson Research Center, P. O. Box, 704, Yorktown Heights, NY 10598
E-mail: {cqli,yuanchil,jsmith,mh}@us.ibm.com

Abstract

In this paper, we describe a new paradigm for information retrieval in which the retrieval target is based on a model. Three types of models – linear, finite state, and knowledge models are discussed. These information retrieval scenarios often arise from applications such as environmental epidemiology, oil/gas production and exploration, and precision agriculture/forestry. Traditional model-based data and information processing usually requires the processing of each and every data points. The proposed new framework, in contrast, will process the data progressively using a set of progressive models and utilize indexing techniques specialized for the model to facilitate retrieval, thus achieving a dramatic speedup.

1. Introduction

Most existing information retrieval applications are based on similarity retrieval of templates or examples, such as similarity retrieval of text and image documents. In such retrievals, the query usually consists of a number of keywords or phrase (for text retrieval), or features of an image (for image retrieval) [4-10]. Each of the documents (text or image) in the database is usually represented as one or more vector(s) in a multi-dimensional feature space. The query processing of such similarity retrieval usually involves identifying in the feature space those vectors that have the smallest Euclidean distance to the vector that corresponds to the query target.

This similarity retrieval paradigm, however, is not entirely suitable for many scientific and business decision support applications, which are mostly based on models. For examples:

- Environmental epidemiology: Many environmental epidemic diseases such as Denge fever, Hantavirus Pulmonary Syndrome (HPS), malaria, and Lyme disease are usually carried by insects or rodents. There already exist various disease outbreak models that incorporate remotely sensed images, weather information, GIS (geographic information system) and demographic information to predict the location of the outbreak. The results of these models can then be used to assess the risk of various regions to a major disease outbreak.

- Oil/gas production and exploration: Oil and gas explorations frequently require the integration (or fusing) of information from seismic data and various instruments. The integration and fusing of information is frequently performed based on various knowledge models. As an example, a geologist may be looking for a strata region consisting of shale, on top of sandstone, on top of siltstone. Additional specifications such as the Gamma Ray response has to be higher than a certain number can also be included in the specifications.

The main challenge of applying models to large archives is scalability. Although most of the applications only require the retrieval of a very small subset of the results that maximize or minimize the model, almost all existing methods require applying the model *sequentially* over the entire region of the data. In this paper, we propose a model-based information retrieval framework to address this challenge. The main components of this framework include

- Decompose the model into a progressively represented model,
- Decompose the data in the archive into a progressive data representation which consists of multiple abstraction levels (raw data, features, semantics and metadata) and multiple resolutions, and
- Apply model-specific indexing techniques on the data in the archive.

By using this framework, we have demonstrated that several order-of-magnitude speedups can be achieved for those applications described above.

2. Preliminary

2.1 Linear Time-Invariant Model

A linear time-invariant model refers to a linear regression model and its variations. The model is time invariant as none of the coefficients of the model are time varying. In general, the model has the following form:

$$Y = a_1X_1 + a_2X_2 + \dots + a_nX_n,$$

where each X_i represents the data itself or derived attributes/features from the multi-modal information sources, while the coefficient a_i represents the weights (relative contribution) of the attribute derived from the

data. Well known techniques exist in deriving the “optimal” weights based on collections of data. As an example, in the Hantavirus Pulmonary Syndrome (HPS) example, the risk assessment model for the risk associated with a location (x,y) is:

$$R(x,y) = 0.443X_1 + 0.222X_2 + 0.153X_3 + 0.183 X_4,$$

where X_1 , X_2 , and X_3 correspond to the pixel value of band 4, 5 and 7 of Landsat Thematic Mapper image at location (x,y), while X_4 corresponds to the elevation (in meters) from the corresponding DEM (digital elevation map). In general, the weights of this model can be “trained” by using historical data (previous disease incident reports compiled by those public health officials) in conjunction with the corresponding images and other data. Methods such as linear regression techniques have been widely used to derive the optimal coefficients for the model.

A second example of the linear model is the credit risk modeling, such as the one used by FICO (Fair, Isaac & Co.) for evaluating the credit scores of individuals for loan approval. The complete FICO credit score, which ranges from 300 to 900, has several hundred parameters with a model similar to the one below:

$$FICO = 900 - a_1X_1 - \dots - a_N X_N$$

where the attribute X_i 's include



Figure 1. Regions of high risks to Hantavirus Pulmonary Syndrome.

- Late payments
- The amount of time credit that has been established
- The amount of credit used versus the amount of credit available
- Length of time at present residence

- Employment history
- Negative credit information such as bankruptcies, charge-offs, collections, etc.

The parameters have been chosen such that this model can accurately predict the probability of foreclosures. It has been shown that the probability of foreclosures is less than 2% when the score is higher than 680, while the probability of foreclosures increases to 8% if the score is less than 620.

2.2 Bayesian Network and Knowledge Model

A Bayesian network is a graphical model for probabilistic relationships among a set of variables. It is also known as belief networks, causal networks, probabilistic networks, influence diagrams and knowledge maps. Bayesian networks can readily handle incomplete data sets, allow one to learn about causal relationships, and can be used in conjunction with Bayesian statistical techniques to facilitate combining domain knowledge and

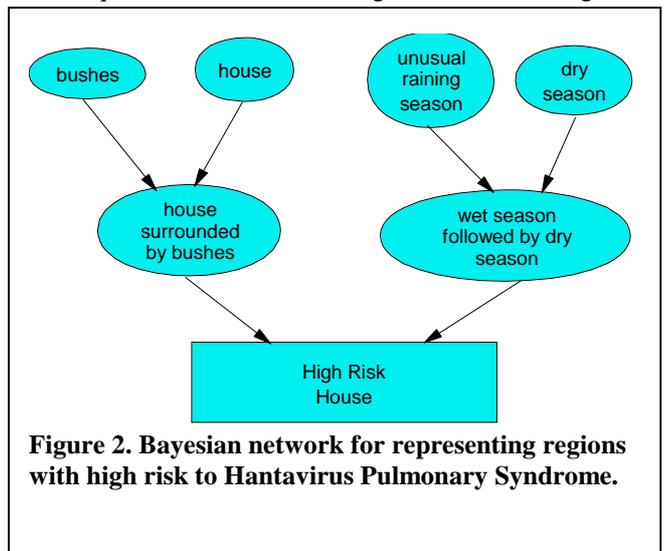


Figure 2. Bayesian network for representing regions with high risk to Hantavirus Pulmonary Syndrome.

data. Consequently, Bayesian network has become a popular representation for encoding expert knowledge in expert systems. Recently, methods have been developed to learn Bayesian networks from data.

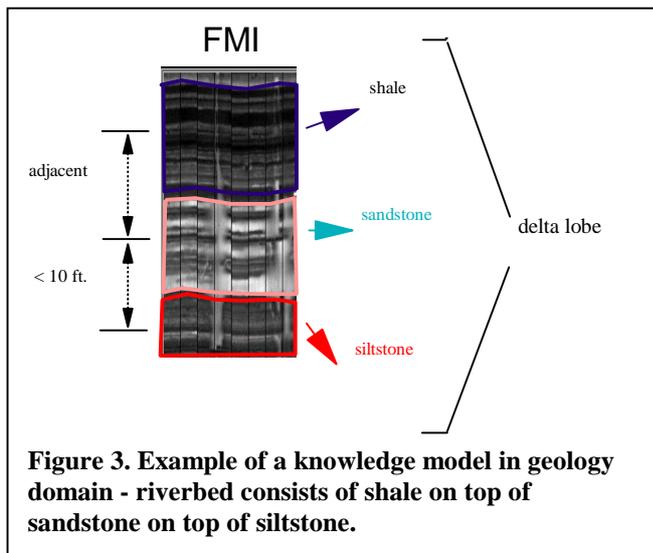
As an example, the high risk houses that are vulnerable to Hantavirus Pulmonary Syndrome can consist of the following rules:

- Area of houses, which are
- surrounded by bushes, and has
- weather pattern of raining season followed by a dry season.

Figure 1 shows an example of such a region based on a high-resolution satellite image. The Bayesian network representation of this model is shown in Fig. 2. Note that this model is multi-modal, as it consists of data from images and weather pattern.

Another example, as shown in Fig. 3, shows the riverbed consisting of:

- shale, on top of
 - sandstones, on top of
 - siltstones, and
 - the Gamma ray of these region is higher than 45.
- Note that this model is also multi-modal, as it consists of data from both images and well log traces (1D series).



3. Model-Based Information Retrieval

Model-based information retrieval emphasizes locating data series or subsets which not only satisfy the query model criterion but also best represent the class of selected data. Similar to other types of information retrieval, the top-K choices based on the ranking evaluated by the model is usually desired. Using the models described in the previous section as examples,

- The linear model is used to locate top-K sets of tuples (X_1, X_2, \dots, X_n) where the model described by the linear regression equation $a_1X_1 + a_2X_2 + \dots + a_nX_n$ is maximized or minimized.
- The finite state model is used to locate the top-K data patterns that satisfy a model that can be described by a finite state machine. When the finite state machine extracted from the data is slightly different from the target finite state machine, it is also possible to define a distance between these two finite state machines based on their similarities.
- The Bayesian network and knowledge models are used to locate the top-K data patterns that satisfy the fuzzy and/or probabilistic rules specified within the model.

These models are usually tightly coupled with various decision support applications, as described in Section 1.

This problem is related to, though different from, the system identification problem. For single-input single-output systems, we denote the input to the system as $\mathbf{X}(t)$, and the output from the system as $\mathbf{Y}(t)$. The system, which is usually characterized by the system transfer

function $\mathbf{H}(s)$, can be computed from the input $\mathbf{X}(s)$ and the output $\mathbf{Y}(s)$ of the system by using the following relationship:

$$\mathbf{H}(s) = \mathbf{Y}(s)/\mathbf{X}(s)$$

where $\mathbf{X}(s)$, $\mathbf{Y}(s)$, and $\mathbf{H}(s)$ are the Laplace transformation of $\mathbf{X}(t)$, $\mathbf{Y}(t)$, and $\mathbf{H}(t)$, respectively. Once the system transfer function is identified, a system representation that has the identical system transfer function can then be constructed.

Similar to system identification problem, most of the existing approaches for model-based information processing usually involve the following process (as shown in Fig. 5):

1. Develop a hypothetical decision model (linear model, finite state model, or knowledge model),
2. Using the available multi-modal data to fit the model and determine the model coefficients,
3. Use the model to identify and retrieve subset of the data that satisfy the model constraints
4. Use the retrieved data to revise the model.
5. Apply the revised model to a much bigger data set.
6. Repeat steps 3 and 4 as many times as necessary.

The first two steps calibrate the model based on a training set, while the remaining steps involve applying the model for information retrieval, potentially on a very large archive. There have been intensive studies elsewhere for developing models based on small training set. Most existing methods will then require sequentially applying the model on the data in the data archive. Substantial re-computation on the entire data set is required even when there is a small revision of the model. There are very little investigations on speeding up model executions for large archives (required at step 5 in the workflow described above). Consequently, it is extremely difficult to make small revisions on the model based on the data archive.

In this paper, we propose the following framework to facilitate a substantial speedup on model-based information retrieval: (1) *progressive model execution* using *progressive data representation* and, (2) *high-dimensional indexing support* for the decision models.

3.1 Progressive Model and Data Representation

Execution of the model progressively allows incremental generation of model predictions. This will enable the execution of the more complete version of the model on those regions that have been predicted to be high risk, thus resulting in more accurate predictions for those regions with high interests sooner.

Progressive models usually involve the generation of increasingly coarse representations of the same

model. In the linear model case, the risk model for the disease at location (x,y) and time t is

$$R(x,y,t) = a_1 X_1(x,y,t) + a_2 X_2(x,y,t) + a_3 X_3(x,y,t) + a_4 R(x,y,t-1)$$

If $|a_1, a_2| \gg |a_3, a_4|$ then, a coarser representation of the model for $Risk(x,y,t)$ will be

$$R^*(x,y,t) \sim a_1 X_1(x,y,t) + a_2 X_2(x,y,t)$$

Consequently, $R(x,y,t)$ and $R^*(x,y,t)$ represent two levels of progressive models.

In general, the generation of progressively coarser representation of a model can be accomplished by analyzing the relative contribution of each parameter to the overall model. This is related to query planning issues in query optimization for object-relational databases. However, query planning usually rearranges the execution order so that operations resulting in maximal filtering will be executed earlier. In contrast, progressive model generation will select those operations that are most relevant to the final results to be executed first.

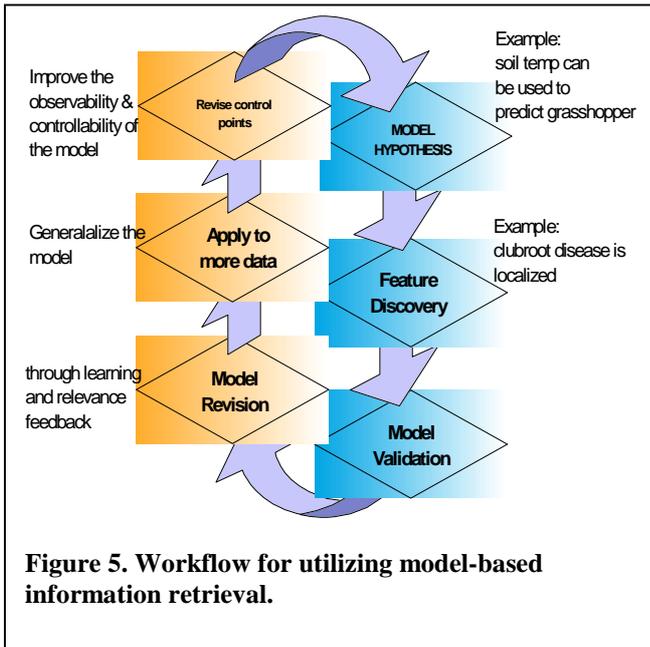


Figure 5. Workflow for utilizing model-based information retrieval.

A complementary dimension that further enhances the incremental model execution is the use of progressive data representations. Two orthogonal dimensions exist for generating progressive data representations. *Multi-resolution representations*, such as wavelets can be used to provide rough approximations of information at low resolutions (low data volumes), with more detailed views at higher resolutions [1-3]. In contrast, *multiple abstraction level representations* rely on the fact that raw information can be processed into alternate formulations such as features (texture, color, shape, etc.) and semantics that require lower data volumes at the expense of fidelity. For example, contours can be computed from a data array,

allowing for very rapid identification of areas with low or high parameter values, but with a loss of accuracy.

Previously, we have shown in [13] that a 30-times speedup can be achieved through applying progressive classification on progressively represented data. This type of classification of satellite images can be viewed as a special case of applying Bayesian network. We have also shown in [12] that a 4-8 times speedup can be accomplished through applying feature extraction progressively on progressively represented data. By applying progressive model execution on progressively represented data, a substantial speedup compared to using either progressive models or progressive data representation can usually be achieved.

3.2 High-Dimensional Model-Based Indexing Support

Most of the high-dimensional indexing techniques such as R^* -tree are optimized for spatial range queries. These techniques have also been utilized for processing similarity-based queries by pruning the search space through range queries [14]. However these techniques are sub-optimal for model-based queries, as these indices do not indicate where to find data points that will maximize the model.

An indexing technique, Onion, based on convex hull was proposed in [11] to address the issue of locating tuples that optimize (either maximize or minimize) a linear model. Experimental results have shown, with three-parameter Gaussian distributed data sets, a speed-up of 13,000 fold is achieved for retrieving the top-one choice while a speed-up of 1,400 fold is achieved for retrieving the top-ten choices, both measured against sequential scan of the unindexed data set.

For finite state and Bayesian-network-based models, indexing techniques may not be suitable. A dynamic programming based search space pruning technique, SPROC (Sequential Processing of Fuzzy Cartesian Queries) was proposed in [15] to reduce the computational complexity from $O(L^M)$ to $O(MKL^2)$, where L is the size of the database, M is the number of rules, and K is the number of retrievals. This complexity is further reduced to $O(ML \log L + \sqrt{LK} + K^2 \log K)$ in [16].

4. Summary

In this paper, we describe a new framework for information retrieval, in which the retrieval is based on a model instead of a simple template as the search target. Three types of models – linear model, finite state model, and Bayesian networked-based model are discussed. Traditional model-based processing techniques usually require the processing of each and every data points by the complete model. The new framework, in contrast, will

execute the model progressively on progressively represented data, with the help from model-specific indices.

By using this framework, we have demonstrated that several order-of-magnitude speedups can be achieved for a number of scientific and business applications such as environmental epidemiology, oil/gas exploration, and precision agriculture/forestry.

5. Acknowledgement

This work was funded in part by IBM/NASA CAN NCC5-305. The authors also appreciate the discussions with Prof. Gregory Glass of Johns Hopkins Univ., as well as Drs. Peter Tilke and Ian Bryant from Schlumberger-Doll Research.

6. References

- [1] J. R. Smith and S.-F. Chang. Joint adaptive space and frequency graph basis selection. In IEEE Proc. Int. Conf. Image Processing (ICIP), Santa Barbara, CA, October 1997.
- [2] J. R. Smith and C.-S. Li. An adaptive view element framework for multi-dimensional data management. In ACM Int Conf on Information and Knowledge Management (CIKM), Kansas City, MO, November 1999.
- [3] J. R. Smith, V. Castelli, and C.-S. Li. Adaptive storage and retrieval of large compressed images. In IS&T/SPIE Symposium on Electronic Imaging: Science and Technology - Storage & Retrieval for Image and Video Databases VII, San Jose, CA, January 1999.
- [4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The {QBIC} system," IEEE Computer, vol. 28, no. 9, pp. 23 -- 32, September 1995.
- [5] J. R. Smith and S.-F. Chang, "Visualeek: A fully automated content-based image query system," in Proc. International Conference on Image Processing, 1996.
- [6] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C. Shu, "Virage: image search engine: an open framework for image management," in Symposium on Electronic Imaging: Science and Technology-Storage & Retrieval for Image and Video Databases {IV}. IS&T/SPIE, 1996, vol. 2670, pp. 76 -- 87.
- [7] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data", IEEE Trans. Pattern Analysis Machine Intell. Special Issue on Digital Libraries", vol.18, no. 8, 1996.
- [8] J. R. Smith and S.-F. Chang, "SaFe: A general framework for integrated spatial and feature image search," IEEE first workshop on Multimedia Signal Processing, pp. 301-306, 1997.
- [9] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Region based image querying," IEEE Proc. CBAIVL'97, pp. 42-49, 1997.
- [10] D. A. Forsyth, and M. M. Fleck, "Body Plans," IEEE Proc. CVPR, pp.678-683, 1997.
- [11] Y.-C. Chang, L. D. Bergman, V. Castelli, C.-S. Li, M.-L. Lo, and J. R. Smith, "The Onion Technique: Indexing for Linear Optimization Queries," accepted for ACM Proc. SIGMOD , 2000.
- [12] C.-S. Li and M.-S. Chen, "Progressive Texture Matching for Earth Observing Satellite Image Database," SPIE Multimedia Storage and Archiving Systems, Boston, Nov. 1996.
- [13] V. Castelli, C.-S. Li, J. J. Turek, and I. Kontoyiannis, "Progressive Classification in the Compressed Domain for Large EOS Satellite Databases," IEEE ICASSP'96 , April, 1996.
- [14] A. Thomasian, V. Castelli, and C.-S. Li, "Clustering and Singular Value Decomposition for Approximate Indexing in High Dimensional Spaces," Conference on Information and Knowledge Management (CIKM'98), Nov., 1998.
- [15] C.-S. Li, J. R. Smith, L. Bergman and V. Castelli, "SPROC: Sequential Processing for Content-Based Retrieval of Composite Objects," SPIE Photonics West'98, San Jose, Jan. 1998.
- [16] C.-S. Li, Y.-C. Chang, J. R. Smith, and L. D. Bergman, "Framework for Efficient Processing of Content-Based Fuzzy Cartesian Queries," SPIE Storage and Retrieval System, Jan. 2000.