Effective Content Based Data Retrieval Algorithm for Data Mining

Aws Saad Shawkat, H K Sawant

Abstract—Data mining techniques are used in a many research areas, including mathematics, cybernetics, genetics and marketing. Web mining, a type of data mining used in customer relationship management (CRM), takes advantage of the huge amount of information gathered by a Web site to look for patterns in user behavior. Data mining is sorting through data to identify patterns and establish relationships. Data mining parameters include: Regression - In statistics, regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. Sequence or path analysis - looking for patterns where one event leads to another later event. Classification - looking for new patterns. Clustering - finding and visually documenting groups. Decision Trees – Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.

I. INTRODUCTION

Data mining is an iterative process that typically involves the following phases:

a) Problem definition: A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required.

b) Data exploration: Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

c) Data preparation: Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

d) Modeling: Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

e) Evaluation: Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions: Does the model achieve the business objective? Have all business issues been considered? At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

f) Deployment: Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

II. CLUSTER ANALYSIS

Cluster analysis is an exploratory data analysis tool for solving classification problems. Its object is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type.

A) Types of Cluster Analysis

Hierarchical Clustering: Hierarchical algorithms find successive clusters using previously established clusters. These algorithms can be either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divide algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Partitional clustering: Partitional algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering.
B) Types of Agglomerative Clustering

Most agglomerative hierarchical clustering algorithms are variants of the single-link or complete-link algorithms. These two basic algorithms differ only in the way they characterize the similarity between a pair of clusters.

In the single-link method, the distance between two clusters is the minimum of the distances between all pairs of samples drawn from the two clusters (one element from the first cluster, the other from the second).

In the complete-link algorithm, the distance between two clusters is the maximum of all distances between all pairs drawn from the two clusters. A graphical illustration of these two distance measures

C) K MEAN Algorithm

Typically, this criterion is met when there is no reassignment of any sample from one cluster to another that will cause a decrease of the total squared error.

K-means algorithm is popular because it is easy to implement, and its time and space complexity is relatively small.

ALGORITHM

- Typically, this criterion is met when there is no reassignment of any sample from one cluster to another that will cause a decrease of the total squared error.
- K-means algorithm is popular because it is easy to implement, and its time and space complexity is relatively small.
- A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function if the initial partition is not properly chosen.

III. DATA REDUCTION

Means reducing the number of cases or variables in a data matrix. The basic operations in a data-reduction process are delete column, delete a row, and reduce the number of values in a column. These operations attempt to preserve the character of the original data by deleting data that are nonessential. There are other operations that reduce dimensions, but the new data are unrecognizable when compared to the original data set, and these operations are mentioned here just briefly because they are highly application-dependent.

A) Entropy

A method for unsupervised feature selection or ranking based on entropy measure is a relatively simple technique; but with a large number of features its complexity increases significantly.

The similarity measure between two samples can be defined as

\[ S_{ij} = e^{-\alpha D_{ij}} \]

Where \( D_{ij} \) is the distance between the two samples \( x_i \) and \( x_j \), and \( \alpha \) is a parameter mathematically expressed as

\[ D = \frac{\sqrt[2]{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2 / \max(k) - \min(k))}}{n} \]

\[ \alpha = -\frac{(\ln 0.5)}{D} \]

Hence, \( \alpha \) is determined by the data. But, in a successfully implemented practical application, it was used a constant value of \( \alpha = 0.5 \). Normalized Euclidean distance measure is used to calculate the distance \( D_{ij} \) between two samples \( x_i \) and \( x_j \):

\[ D_{ij} = \left[ \sum_{k=1}^{n} \left( \frac{|x_{ik} - x_{jk}|}{\max(k) - \min(k)} \right)^2 \right]^{1/2} \]

- where \( n \) is the number of dimensions and \( \max(k) \) and \( \min(k) \) are maximum and minimum values used for normalization of the k-th dimension.
- All features are not numeric. The similarity for nominal variables is measured directly using Hamming distance.

\[ S_{ij} = \frac{\sum_{k=1}^{n} |x_{ik} - x_{jk}|}{n} \]

where

The total number of variables is equal to \( n \). For mixed data, we can discretize numeric values (Binning) and transform numeric features into nominal features before we apply this similarity measure.

If the two measures are close, then the reduced set of features will satisfactorily approximate the original set. For a data set of \( N \) samples, the entropy measure is

\[ E = -\sum_{i=1}^{N} \sum_{j=1}^{N} (S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log(1 - S_{ij})) \]

where \( S_{ij} \) is the similarity between samples \( x_i \) and \( x_j \).

This measure is computed in each of the iterations as a basis for deciding the ranking of features. We rank features by gradually removing the least important feature in maintaining the order in the configurations of data. The steps of the algorithm are base on sequential backward ranking, and they have been successfully tested on several real-world applications.

B. Linear Regression:

In statistics, linear regression refers to any approach to modeling the relationship between one or more variables denoted \( y \) and one or more variables denoted \( X \), such that the model depends linearly on the unknown parameters to be estimated from the data.

Linear regression has many practical uses. Most applications of linear regression fall into one of the following two broad categories:
• If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.

• Given a variable y and a number of variables X₁, ..., Xₖ that may be related to y, then linear regression analysis can be applied to quantify the strength of the relationship between y and the Xᵢ to assess which Xᵢ may have no relationship with y at all, and to identify which subsets of the Xᵢ contain redundant information about y, thus once one of them is known, the others are no longer informative.

IV. IMPLEMENTATION

The document processor prepares processes, and inputs the data, documents, pages, or sites that users search against. The document processor performs some or all of the following steps:

• Normalizes the document stream to a predefined format.
• Breaks the document stream into desired retrievable units.
• Isolates and metatags subdocument pieces.
• Identifies potential indexable elements in documents.
• Deletes stop words.
• Stems terms.
• Extracts index entries.
• Computes weights.
• Creates and updates the main inverted file against which the Data Extractor searches in order to match queries to documents.

Steps 1-3: Preprocessing. While essential and potentially important in affecting the outcome of a search, these first three steps simply standardize the multiple formats encountered when deriving documents from various providers or handling various Web sites. The steps serve to merge all the data into a single consistent data structure that all the downstream processes can handle. The need for a well-formed, consistent format is of relative importance in direct proportion to the sophistication of later steps of document processing. Step two is important because the pointers stored in the inverted file will enable a system to retrieve various sized units — either site, page, document, section, paragraph, or sentence.

Step 4: Identify elements to index. Identifying potential indexable elements in documents dramatically affects the nature and quality of the document representation that the engine will search against. In designating the system, we must define the word "term." Is it the alpha-numeric characters between blank spaces or punctuation?

If so, what about non-compositional phrases (phrases in which the separate words do not convey the meaning of the phrase, like "skunk works" or "hot dog"), multi-word proper names, or inter-word symbols such as hyphens or apostrophes that can denote the difference between "small business men" versus small-business men." Each Data Extractor depends on a set of rules that its document processor must execute to determine what action is to be taken by the "tokenizer," i.e. the software used to define a term suitable for indexing.

Step 5: Deleting stop words. This step helps save system resources by eliminating from further processing, as well as potential matching, those terms that have little value in finding useful documents in response to a customer's query. This step used to matter much more than it does now when memory has become so much cheaper and systems so much faster, but since stop words may comprise up to 40 percent of text words in a document, it still has some significance. A stop word list typically consists of those word classes known to convey little substantive meaning, such as articles (a, the), conjunctions (and, but), interjections (oh, but), prepositions (in, over), pronouns (he, it), and forms of the "to be" verb (is, are). To delete stop words, an algorithm compares index term candidates in the documents against a stop word list and eliminates certain terms from inclusion in the index for searching.

Step 6: Term Stemming. Stemming removes word suffixes, perhaps recursively in layer after layer of processing. The process has two goals. In terms of efficiency, stemming reduces the number of unique words in the index, which in turn reduces the storage space required for the index and speeds up the search process. In terms of effectiveness, stemming improves recall by reducing all forms of the word to a base or stemmed form. For example, if a user asks for analyze, they may also want documents which contain analysis, analyzing, analyzer, analyzes, and analyzed. Therefore, the document processor stems document terms to anal- so that documents which include various forms of anal- will have equal likelihood of being retrieved; this would not occur if the engine only indexed variant forms separately and required the user to enter all. Of course, stemming does have a downside. It may negatively affect precision in that all forms of a stem will match, when, in fact, a successful query for the user would have come from matching only the word form actually used in the query.

Systems may implement either a strong stemming algorithm or a weak stemming algorithm. A strong stemming algorithm will strip off both inflectional suffixes (-s, -es, -ed) and derivational suffixes (-able, -aciousness, -ability), while a weak stemming algorithm will strip off only the inflectional suffixes (-s, -es, -ed).

Step 7: Extract index entries. Having completed steps 1 through 6, the document processor extracts the remaining entries from the original document.
The output of step 7 is then inserted and stored in an inverted file that lists the index entries and an indication of their position and frequency of occurrence. The specific nature of the index entries, however, will vary based on the decision in Step 4 concerning what constitutes an “indexable term.” More sophisticated document processors will have phrase recognizers, as well as Named Entity recognizers and Categorizers, to insure index entries such as Milosevic are tagged as a Person and entries such as Yugoslavia and Serbia as Countries.

Step 8: Term weight assignment. Weights are assigned to terms in the index file. The simplest of Data Extractors just assign a binary weight: 1 for presence and 0 for absence. The more sophisticated the Data Extractor, the more complex the weighting scheme. Measuring the frequency of occurrence of a term in the document creates more sophisticated weighting, with length-normalization of frequencies still more sophisticated. Extensive experience in information retrieval research over many years has clearly demonstrated that the optimal weighting comes from use of “tf/idf.” This algorithm measures the frequency of occurrence of each term within a document. Then it compares that frequency against the frequency of occurrence in the entire database.

\[ w_i = tf_i * \log \left( \frac{D}{df_i} \right) \]

Term Weight =

where

- \( tf_i \) = term frequency (term counts) or number of times a term i occurs in a document. This accounts for local information.
- \( df_i \) = document frequency or number of documents containing term i
- \( D \) = number of documents in a database.

V. CONCLUSION

The model in which every decision is based on the comparison of two numbers within constant time is called simply a decision tree model. It was introduced to establish computational complexity of sorting and searching, advantages of applying is Easy to understand, Map nicely to a set of business rules, Applied to real problems, Make no prior assumptions about the data, Able to process both numerical and categorical data.

Data mining techniques are used in a many research areas, including mathematics, cybernetics, genetics and marketing. Web mining, a type of data mining used in customer relationship management (CRM), takes advantage of the huge amount of information gathered by a Web site to look for patterns in user behavior.

REFERENCES


Aws Saad Shawkat from Information Technology Department at Bharati Vidyapeeth Deemed University College of Engineering, Dhanuskodi, Pune India. His areas of interest are Software Engineering and Data mining.

H K Sawant is working as an Professor in Information Technology Department at Bharati Vidyapeeth Deemed University College of Engineering, Dhanuskodi, Pune India. He was awarded his Master of Technology Degree from IIT Bombay. His areas of interest are Computer Network, Software Engineering and Multimedia System. He has Twenty years experience in teaching and research. He has published more than twenty research papers in journals and conferences. He has also guided ten postgraduate students.