ISSN : 0974 - 7435 Volume 10 Issue 9

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(9), 2014 [4257-4262]

Web information retrieval based on data mining technology

Yunpeng Cai
Shenyang Normal University, Liaoning, 110034, (CHINA)
E-mail : caicaiyp@163.com

ABSTRACT

Whether Web information retrieval can be quick and efficient depends on the intelligent degree of information retrieval, and the utilization of Datamining technology can greatly enhance the retrieval function of Wed information. This paper mainly described the application of Datamining in Web information retrieval, and adopted Bayesian network algorithm for conducting the association of related data. Before application, it was needed to conduct simple description on retrieval platform, and then to elaborate and analyze the association rules of Datamining, and finally to conduct simulated analysis of living cases based on Bayesian algorithm. The final obtained experimental result was that Datamining technology combined with the application of Bayesian algorithm played an intelligent and personalized role in Web information retrieval, thus it had great research value.

KEYWORDS

Datamining; Web information retrieval; Bayesian network; Association rules.

INTRODUCTION

Search engine on Wed have partially solved the problem of resource discovery. However, it often returns to users' tens of thousands of retrieved webpage, among which a large portion of webpage have nothing to do with the user's retrieval requirements. Thus users cannot obtain the valuable information they need quickly and accurately. Moreover, search engine aims to discover resource on Web. In terms of the knowledge discovery on Web, though the retrieval accuracy is high, search engine cannot competent. Facing this challenge, Datamining technology has revealed its strong vitality.

The so called Datamining refers to the process of extracting information and knowledge that are implicit and unknown but also potentially useful from the original data that is abundant, incomplete, noisy and random[1]. Or in other words, it refers to the process of discovering valuable knowledge (KDD) from database, and then conducting data analysis, data fusion and decision support. People consider data as the sources of knowledge formation. Similar to mining stone from mineral and mining gold from gravel, Datamining is mining a little bit of information it needs from vast sea of original data. What is involved in Datamining is mostly the structural data. In order to deal with the heterogeneous, nonstructural or semi-structural data on Web, Web Datamining has become an important branch of Datamining research[2]. Web Datamining is originated from Datamining. It aims at disposing nonstructural data, and at the same time using its research product to enhance the accuracy and efficiency of information retrieval. Moreover, it also intends to discover the potentially useful pattern or information from WWW database, thus to make Web information retrieval develop into a new stage.

WEB INFORMATION RETRIEVAL

Web information retrieval structure

The so called Web information retrieval refers to the integration of the following several aspects, such as different fields of expert system, database management system, user model, information retrieval, pattern recognition, natural language understanding, etc. It is to conduct in-depth integration on the relatively advanced knowledge and technology of theses aspects[3], thus Web information retrieval appears. In terms of its structure, there are seven parts of main functional structure: user interaction platform, database system, information collection platform, Web server, system management and operating platform, resource management platform, Data mining module and search tool.

According to the different users, the whole information retrieval of Web also can be divided into users that experience Web service, online community and the platform of data demonstration, information entry staff that can type resources in search database, and system administrators that can conduct interactive processing on data cache, management on Web users and reconnaissance on information.

Association rules of datamining

Data size in Web information is increasingly increased, and only in vast information can the client's demand points be found and can the specialty of search engine in Web be actually reflected. The most mainstream way of internet information search is intelligent search which includes many implement methods, among which the most popular method is artificial intelligence test. It refers to the realization of a series of processes like collecting, searching, filtrating, etc from multifarious information. In the integration of information, it can embody its advantages, which play a guiding role in Web users and have users to pay more attention to their searching information. The association rules of Data mining can be used in vast data, and can search the correlative relation among them, thus to realize a kind of retrieval mode of user resource and the integration of user information resource[4].

Suppose $I = \{i_1, i_2, ...i_m\}$ is retrieval set which also can be called as item set, D is the database of affairs. Any affair can be represented by {TID, T}, of which $T = \{i_1, i_2, ...i_k\}, i_j \subseteq I(j = 1,2,..., m)$ is used

for signifying the item set that involved in affair. The content searched by users can be defined as the following three types:

Definition 1: the relevancy of association rules, that is the ratio between affair set that simultaneously contains X and Y and all of affair sets, which is usually represented by support (X⇒Y). That is:

$$\sup port(X \Rightarrow Y) = \frac{\left|\{T : X \bigcup Y \subseteq T, T \in D\}\right|}{\left|D\right|} \tag{1}$$

Definition 2: reliability of associative rules: that is the ratio between affair item that simultaneously contains X and Y and affair that contain X, which is generally represented by confidence (X⇒Y), that is:

$$confidence = (X \Rightarrow Y) = \frac{\left|\{T : X \bigcup Y \subseteq T, T \in D\}\right|}{\left|T : X \subseteq T, T \in D\right|} \tag{2}$$

Definition 3: if the relevancy support (X) of item set is larger than the threshold value minsup, then this item set is called frequent item set.

NETWORK INFORMATION RETRIEVAL BASED ON BAYES

Bayesian network algorithms

Use mathematical knowledge to analyze Bayesian network model: define certain random variable set as $x = \{x_1, x_2, \cdots x_n\}$, of which x_i stands for the vector quantity of dimension m. According to Bayesian network, we can get: B=<G, θ>. That is, one Bayesian network model contains two parts which are respectively used to represent different fields, thus to conduct quantitative and qualitative description. Then network parameter θ of Bayes and its relevant network structure G are represented[5].

The first part is mainly based on Bayesian network model structure (G), which represents the directed acyclic graph (DAG). It is made up of the collection of certain sequence node and directed edge set, of which node stands for element and directed edge stands for the relation among elements. Use arrows of these edges to represent the relation among different elements which is also called the decision relation. Bayesian network model can stands for causality, but is also not limited to it[5].

According to the chain rule of probability, we can get:

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \left| X_1, X_2, \cdots, X_{i-1}\right.) \tag{3}$$

As for variable X_i, we can find the minimum subset $Parent(X_i) \subseteq \{X_1, X_2, \cdots, X_{I-1}\}$ that independent on it, and to have it satisfy:

$$P(X_i \left| X_1, X_2, \cdots, X_{i-1}\right.) = P(X_i \left| Parent(X_i)\right.) \tag{4}$$

Therefore, if the numerical value $\langle x_1, \cdots, x_n \rangle$ is given to variable $\langle X_1, \cdots, X_n \rangle$, then the following formulation can be used to represent the joint probability distribution of Bayesian network:

$$P(x_1,\cdots,x_n)=\prod_{i=1}^{n}P(x_i|Val(Parent(X_i)))$$
$$=\prod_{i=1}^{n}\theta_{x_i|Val(Parent(X_i))} \quad \textbf{(5)}$$

We can establish the Bayesian network based on internet information retrieval through the following three steps:

(1) Determine the parameter of model and the possible value, conduct the reconnaissance of information according to practical problems, and have the variable that need to think over as sunset and need to notice the irrelevancy among them.

(2) As for how to define the network system structure based on Bayes: mainly by means of data learning and ways of how to determine the interdependency among each of variable provided by experts, thus to establish the expression that is relatively independent of condition and based on assertory and directed loop-free structural chart. As for arbitrary variable X_i, it all can find out a minimum subset $Parent(X_i)\subseteq\{X_1,X_2,\cdots,X_{I-1}\}$, and there exist dependency between variable in $Parent(X_i)$ and variable X_i. Therefore, the construction of Bayesian network structure model mainly includes the following two steps: to rank all variables $X_1,X_2,\cdots,X_n$ according to certain order; then to meet the parent node set of formula (2) $Parent(X_i)$ $(i=1,2,\cdots,n)$.

(3) It is needed to affirm the network relevant parameter of Bayesian model: to appoint or through learning obtain the local probability $P(x_i|Parent(X_i))$.

Obviously, the above mentioned step (2) and (3) are the core of whole Bayesian network model structure. Without doubt, the above mentioned each step may be alternatively conducted, not merely the simple and successive treating process.

Bayesian network model can solve the incomplete and uncertain specialty of problem. It has certain effect on the unstable character of some comparatively complicated equipment and system fault brought by relevancy character. At present, it has been popularized, promoted and implemented in many industrial fields. In the known Bayesian network structure, through edge to signify and measure the dependency among various elements, and at the same time, use probability to express the strength of this dependent degree[5].

The gather of certain variable randomized by definition, such as $x=\{X_1,X_2,\cdots X_n\}$, and as for X_i, it refers to the base vector of certain dimension m. Then according to Bayesian Belief network rule, we can get the probability distribution situation above variable x that meet certain given unite condition. The definition of Bayesian Belief network is formulated as follows:

$$B=<G,\theta> \quad \textbf{(6)}$$

In the first part of the above formula, G stands for directed acyclic graph. The graph fixed point expressed as the random variable $X_1,X_2,\cdots X_n$ in finite set x, and the graph arc stands for certain determined function dependency. If now there is an arc from variable Y to X, then Y is the parents of X and X is the rear-guard of Y. Suppose parents is given, then the corresponding single variable in this graph of the corresponding node is not subsequent but independent. In graph G, all parents variables of X_i is represented by set $Pa(X_i)$.

In the second part, θ stands for the correlated quantization parameter in network. Corresponding to X_i, the incidence of the value x_i of $pa(X_i)$ happens or not is based on the probability of condition.

Therefore, aiming at the specific certain application and practice and based on Bayesian Belief network, the probability distribution situation of the comprehensive conditions based on the set x of variable is set:

$$P_B(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P_B(X_i | Pa(X_i)) \quad \textbf{(7)}$$

Generally, the process of establishing network model based on Bayes is as follows: suppose there existing one group of sample $D = \{x_1, x_2, \cdots, x_n\}$, x_i is the living example of X_i, and search the most suitable network to match this sample. The most common method is to introduce evaluation function $S(B \mid D)$ which is applied to evaluate the satisfaction degree of network structure to the given sample, and to search all of optimized indicator that exist in network.

In the known network modeling algorithm, when the network structure is determined, then it is needed to search the related parent variable of X_i at one time from n-1 nodes waiting for analysis. The algorithm in traditional significance did not consider the interrelation among single elements, and wasted lots of energy in finding unreasonable variable indicator. For example, in the implication as follows: $X \to Y \to Z$, we can find that there exist dependency between X and Y, Y and Z, X and Z. Then we consider that the variable of X and Y are only the parent nodes of Z. Therefore, we know if Y is considered as the parent node of Z, then X will go against the generation of Z and it has no assistance at all. Then it is drawn that the comparatively extensively used certain new algorithm at present is the construction method of certain Bayesian prototype faith model network system based on "compressive candidate". Correlated depend metric function I (X, Y) is set to measure the degree of coupling between two independent variables. The larger the I (X, Y) is, which indicate the stronger the degree of coupling between variable X and Y is, and then the larger the possibility of the set membership between X and Y is; however, if the value of I (X, Y) is quite small, then it is indicated that the probability of becoming parent and child between X and Y is very small. Based on this and through this indicator, at the same time of selecting parent node, we should focus on the parent variable $Y_{i1}, Y_{i2}, \cdots Y_{ik}, k << n$ of X_i under the maximum probability.

Main process of bayesian network intelligent retrieval

Bayesian network is adopted to realize the main flow chart of intelligent retrieval. First is to construct retrieval model, and the method of inputting search of retrieval model mainly is the keyword search. The output of retrieval model is the information resources fed backed by internet retrieval platform system. The function module of retrieval model mainly includes three parts: calculate conditional probability, simulate vector space model and generate frequent vector set[6].

CONCLUSION

This paper adopted Bayesian network to finish the data association of Web information retrieval which has the advantage of personalized information retrieval. Web information retrieval platform needs to conduct targeted screening on mass information according to user's habit, thus to feed back the information that users most wish to obtain more fast and accurate and to realize the function of intelligent retrieval. The application of Bayesian network algorithm to Web information retrieval has good adaptability and expansive application future.

REFERENCES

[1] M.Fan, X.F.Meng, et al; Translate, Datamining-Concept and Technology, Beijing: China Machine Press **(2007)**.

[2] X.Q,Cheng, J.F.Guo, X.L.Le; A Retrospective of Web Information Retrieval and Mining, Journal of Chinese Information Processing, **6**, 111-117 **(2011)**.

[3] Z.H.Xu, J.Qin, T.Zhen; An Approach of Web Information Retrieval Based on Agent-Aglet, Microelectronics & Computer, **1**, 168-176 **(2012)**.

[4] J.L.Lv, S.W.Che; Mining Association Rules Based on Correlation Measure, Journal of Zhejiang University (Science Edition), **3**, 284-288 **(2012)**.

[5] X.C.Li, G.M.Liu; The Analysis Method of Critical Chain Project Management Bayesian Network Model Based on WBS, Machine Design and Manufacturing Engineering, **7**, 1-4 **(2011)**.

[6] J.Meng, P.Wang, J.Zhang, X.K.Wang; Minimal Association Rules Mining Based on Itemset Dependency, Computer Science, **1**, 183-186, 217 **(2013)**.