A Combined Mining Approach and Application in Tax Administration.

Dr. Ela Kumar, Arun Solanki
School of Information and Communication Technology
Gautam Buddha University, Greater Noida

Abstract- This paper reports the development of a model for taxation. This model will work for the tax payers as well as for the administrator. It utilizes the technique of web mining, text mining, data mining and human experience knowledge for creating a knowledge base of taxation. All knowledge from each part is saved in knowledge base through a knowledge management platform. Using this knowledge management platform the administrator and tax payer can retrieve knowledge; send feedback on the basis of actions suggested. This model facilitates to monitor the knowledge management platform. Its application shows the utilization of model for tax administration .Using this model administrator can improve the quality of decisions.

Keywords- Web Mining model, Tax administration, Data mining, Knowledge management.

1. Introduction-

The World Wide Web has recently gained a very high popularity because of this ability to store a large amount of data. Billions of pages are publicly available and is still growing. The web provides an unprecedented freedom to discuss public issues, administration and politics etc. The general public enjoys the right to speak on BBS, message board, self-owned sites and blogs [8]. One can also take part in interactive actions online. Web Mining, that can discover knowledge from huge amount of web pages, has become a high priority research area in computer science [1].

An area of significant public policy importance, where specialists from a range of disciplines interact is taxation. Taxation is a governmental assessment (charge) upon property value, transactions (transfers and sales), licenses granting a right, and/or income. These include Federal and state income taxes, county and city taxes on real property, state and/or local sales tax based on a percentage of each retail transaction, duties on imports from foreign countries, business licenses, Federal tax (and some states' taxes) on the estates of persons who have died, taxes on large gifts, and a state "use" tax in lieu of sales tax imposed on certain goods bought outside of the state.Public finance and economic theories drive much of the policy agenda surrounding the structure of the tax system. The influence of those theories flows naturally into tax administration. However, a revenue authority is either a government department or an agency that reports to and is responsible to government. Public administration and public sector performance management theories therefore provide useful insights into best practice operation [2].

2. Framework of the combined web mining Model

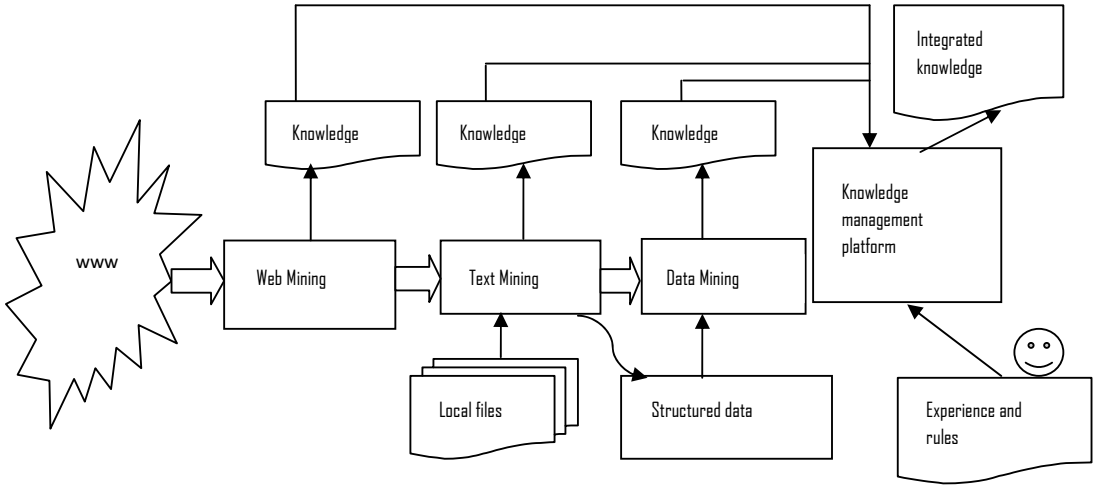

Figure 1 Framework of the combined web mining model.

Web mining is powerful, but it also needs to combine with other technologies and management rules to enhance the effectiveness. Figure1.Shows a framework of combined web mining model.

Web mining extracts knowledge from web data, i.e. web content, web structure, and web usage data [1]. Web content mining is the process of extracting useful information from the contents of web documents. Web mining involves two main steps: Extracting the pages relevant to a query and ranking them according to their quality. Text mining deals with the text documents in general, such as emails, letters, reports and articles, that exists in both intranet and internet environment. Text mining is concerned with the analysis of very large document collections of the extraction of hidden knowledge include topic discovery and tracking, extracting association patterns or abstract, clustering of web documents and classification from text-based data.

Data mining has been used in many fields [3, 9]. More and more leading edge organizations are realizing the data mining provide them the ability to reach their goals in customer relationship management, risk management, fraud and e-business etc [10.]During the process of web mining and text mining, statistical data such as visit-times, last-visit-date, browser number, weights, and number of total words can be collected from mining software or by special software tools, this data set are acted as training sets for data mining. We define two classes for this dataset using a label variable: The pages should be pay attention to (label=1) and not (label=0). After the classification model has been trained, the effect of new pages can be predicted by scoring in earlier time.

Finally, knowledge management is necessary for collecting experts experience and compared with knowledge from data mining [4, 8].

3. Taxation Flow chart (at user end)

Using a browser, the tax payer logs on the tax authority website or portal and then navigates to the 'What's new' section for any information that will affect his current tax responsibility.

Next, the tax payer can navigate to the "Tax Form" section of the portal, where he has two choices: either utilize a wizard to fill out the tax form online or download the tax form to his desktop to complete offline at his

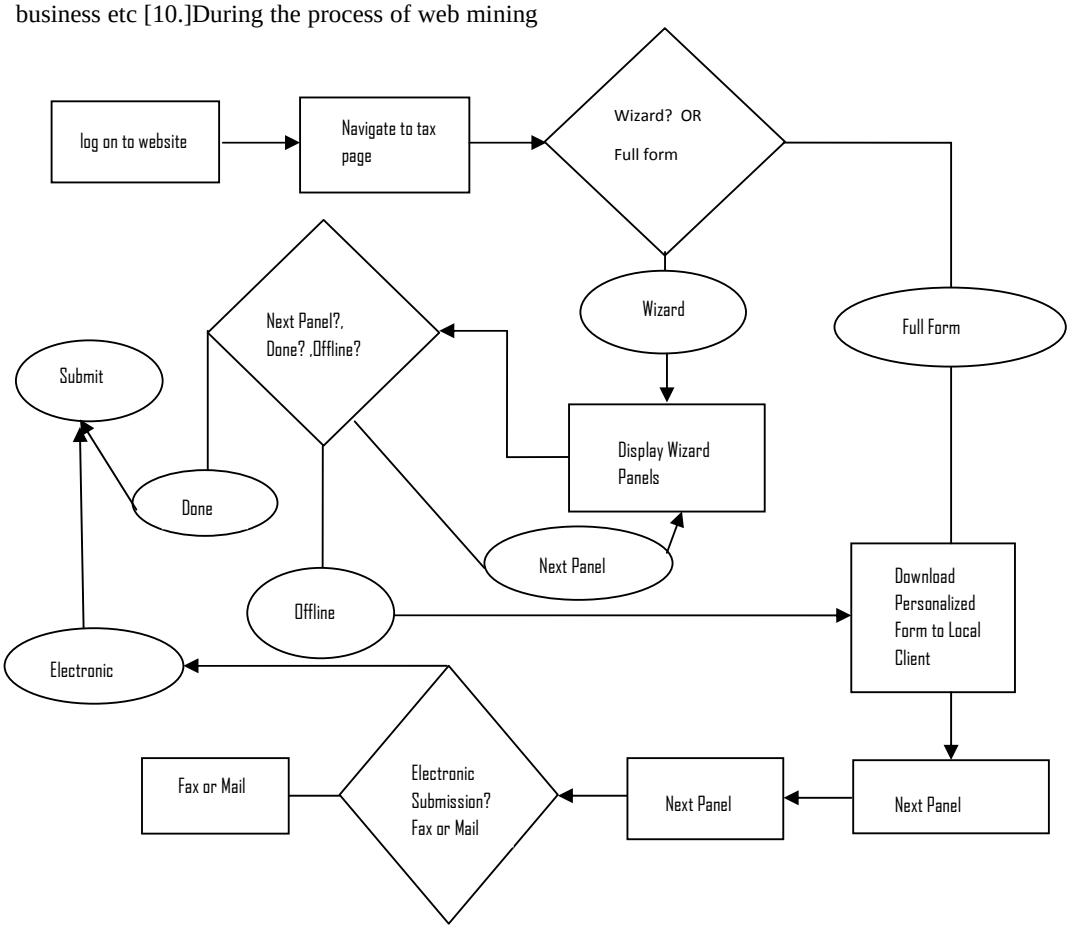

Figure-2 A Flow Diagram of Electronic/Manual Submission of Tax Forms.

convenience. If the tax payer has accessed these services in previous years, many of the fields in the form will be populated with personalized data:

Wizard Option: If the tax payer selects the wizard option, he moves through a series of smaller online forms. When he completes the process, the system automatically aggregates the data into single form that is submitted to the appropriate department.

Offline Option If the taxpayer chooses to complete the form offline, the process is simple and provides several additional options. After completing his tax form, the tax payer can submit it via fax, by mail or electronically. If the tax payer selects the electronic option, he can digitally sign the form, reconnect the agency website or portal and submit it online [7].

4. Taxation Flow Chart (At Administrator end)

In figure 3, using a browser, the tax administrator officer logs on to the system and navigates to an inbox displaying the list of tasks- including tax forms requiring review. The administrator reviews the submitted forms of tax and has the ability to request further information for evaluating the tax form. For example, Administrator may need to view the tax payer's tax submission for the last three years. After analyzing the additional information, administrator can determine the additional information, he/she can determine the appropriate next step- either close the review because there is no issue, notify the payer that a payment is due, or escalate the case to supervisor for further scrutiny. If Administrator forwards the case to her supervisor, he can easily add comments that explain his decision to escalate [7].

5. Data mining process model-

One of the best model for data mining is CRISP-DM (CRoss-Industry Standard Process For data Mining).This is a non-proprietary, documented and freely available data mining model which was conceived in late 1996 by four leaders of the nascent data mining market: Daimler-Benz (now DaimlerChrysler),Integral Solutions Ltd.(ISL), NCR, and OHRA. Developed by industry leaders with input from more than 200 data mining users

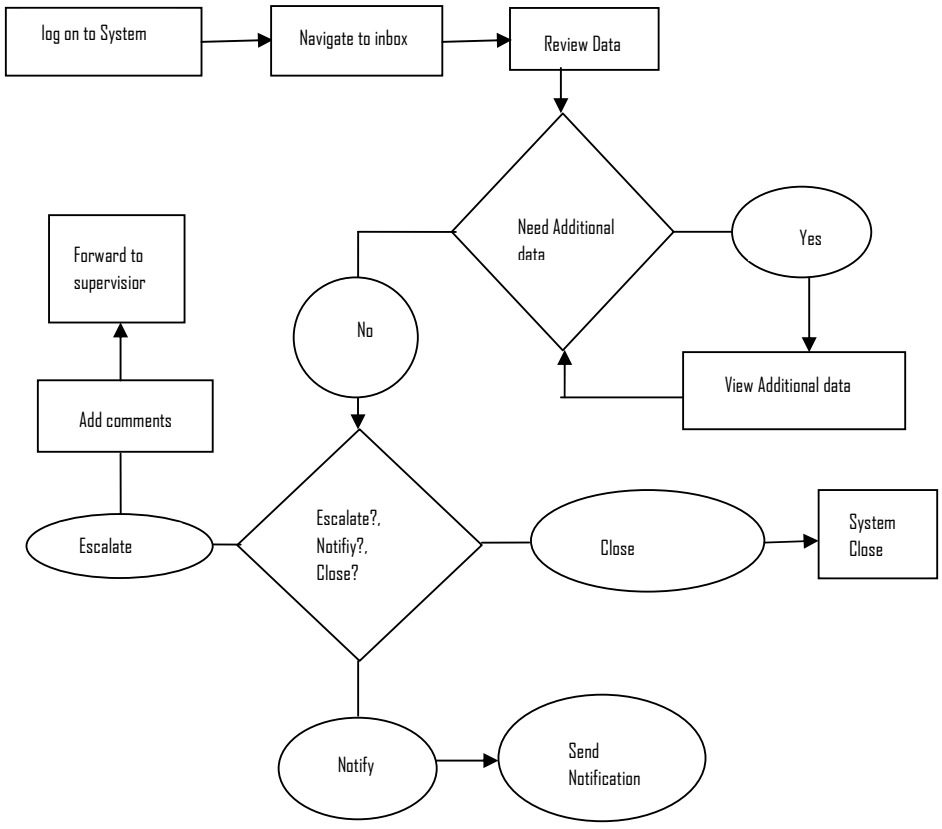

Figure-3 A flow Diagram of Taxation (Administrator end).

and data mining tools and service providers, CRISP-DM is an industry based tool. This model encourages best practices and offers organizations the structure needed to realize better, faster results from data mining [5].

Crisp-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. These phases help organizations understand the data mining process and provide a road map to follow while planning and carrying out a data mining project [5, 6]. The process is shown in figure-4.

In the following we outline each phase briefly-

Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business prospective, and then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to from hypotheses for hidden information.

Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tools(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute section as well as transformation and cleaning of data for modeling tools.

Modeling

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

Evaluation

At this stage in the project you built a model that appears to have high quality from data analysis perspectives. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of data mining results should be reached.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying "live" models within an organization's decision making processes, for example in real-time personalization of web pages or repeated scoring of marketing database. However, depending on the requirements, the deployment phase can be simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up the front what actions need to be carried out in order to actually make use of the created models [6].

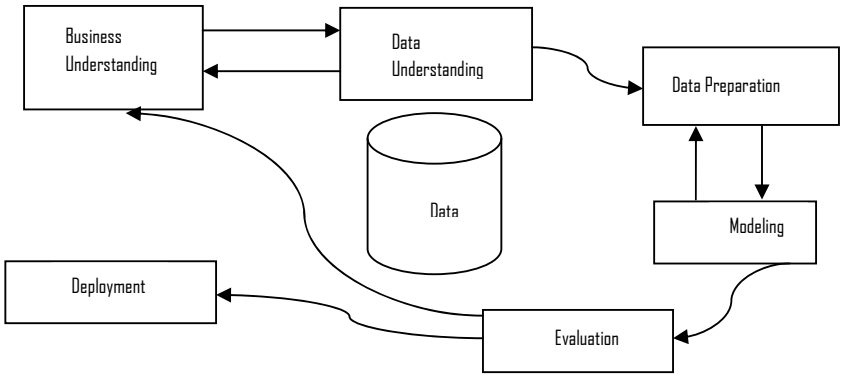

Figure 4. Phases of the CRISP-DM process model.

6. Working of proposed model

In this proposed model, a knowledge management platform (figure 5) is used for data mining process. The knowledge management platform can take care of the new as well as experienced user. This platform is proposed on the basis of CRISP-DM model. Its working flow consists of following steps-

1. When a expert perform data mining, the software is invoked to record what they do, then save in a knowledge
base. The experts can trace back by the records and revise the parameters if they are not satisfied with the mining results.
2. When a new user run into trouble in data mining process, he/she can ask questions to the platform, and then get automatically answers from the knowledge base.
3. After the new user solved their problems, they can write down their experience and save it back into the knowledge base.
4. Developers or experts browser the new user's questions to find out what to do with the working sequence so as to improve the software.
5. With the help of the experts, experience and the new user's feedback, the platform and the algorithms can be improved quickly [8].

This knowledge Management platform is used in the proposed model of web/data mining in tax administration. If an Administrator is new for the system, then this Knowledge Management platform help him for making right decisions according to problems. The proposed model collects information from the WWW. Queries of tax payers also submitted to the database. Except this the knowledge base contains the information of the different type of taxes and their rules. According to this knowledge administrator can calculate the tax of tax payer.

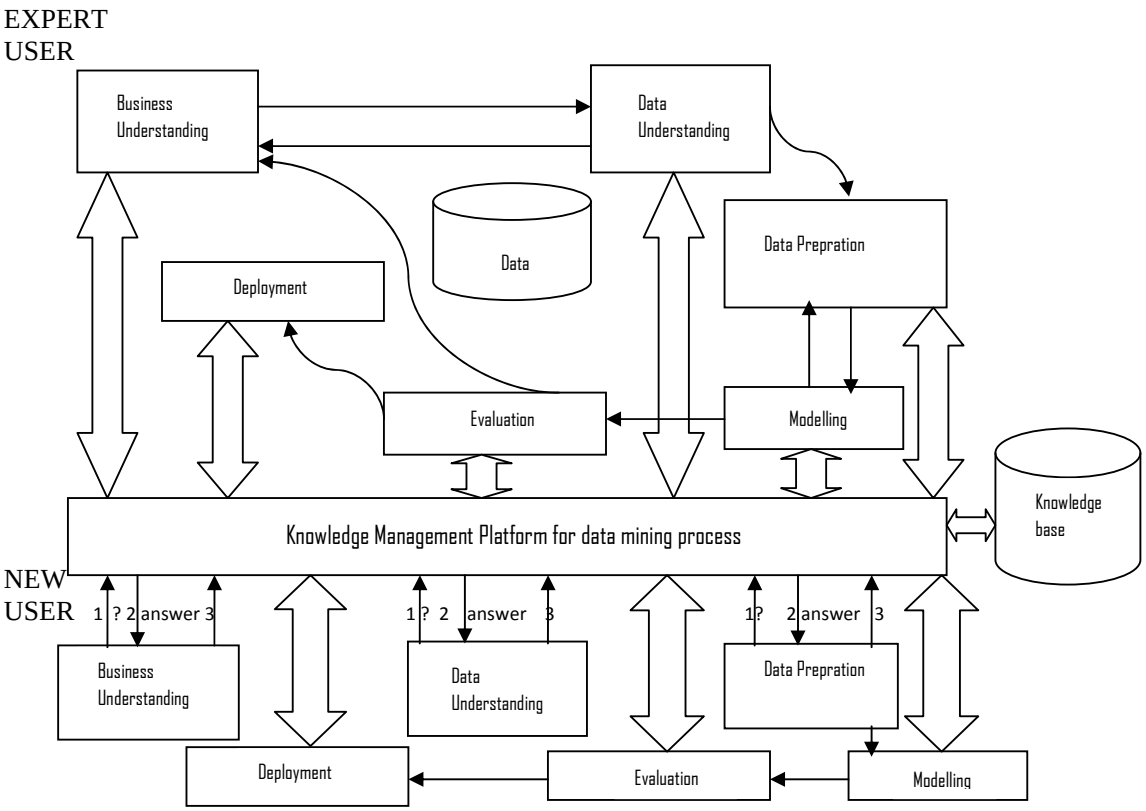

Figure 5. Knowledge Management Platform

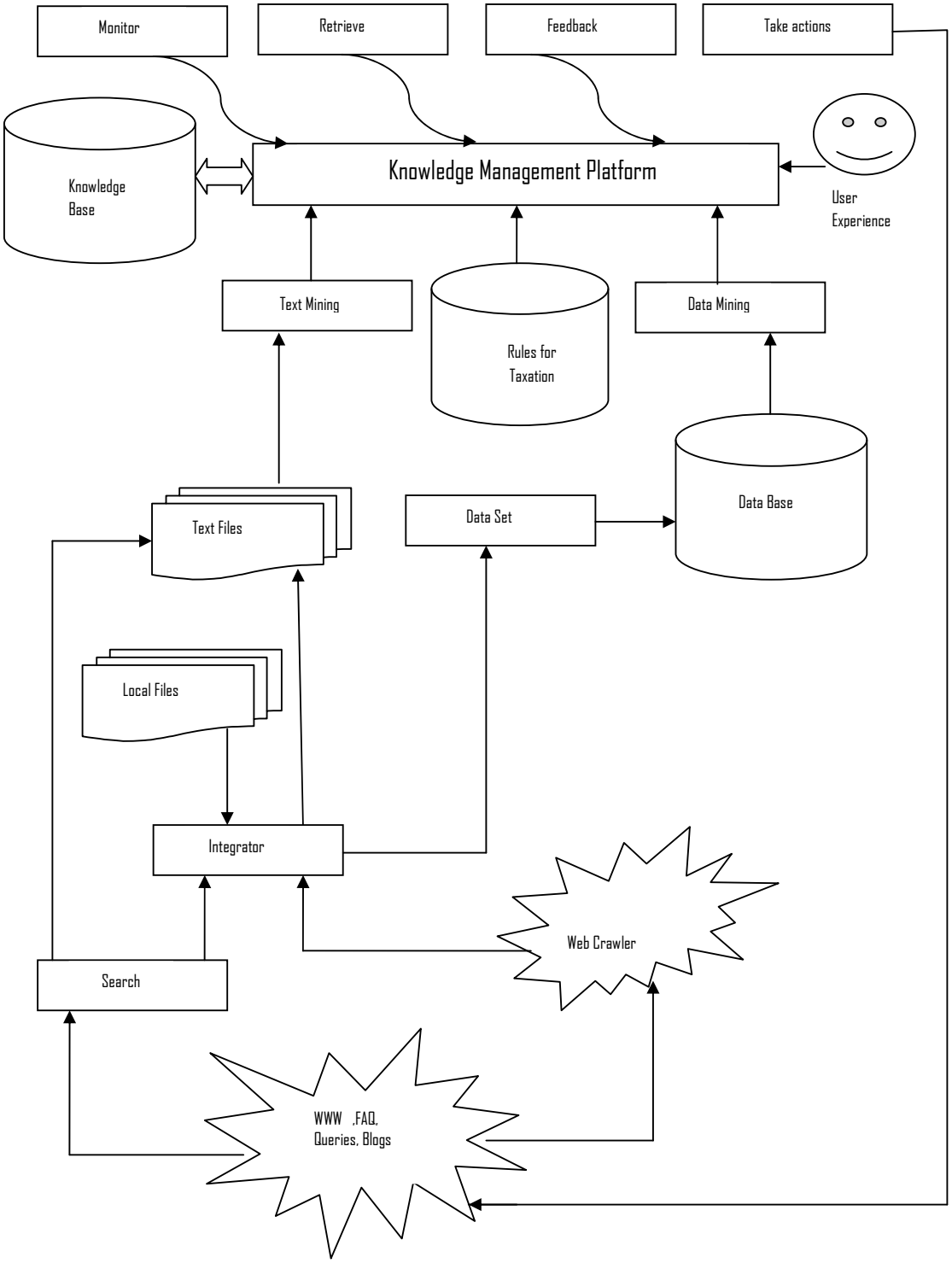

Figure 6. The work process of combined model.

The work process of combined model contains the following steps shown in figure 6.

Step1 Searching the web. Search information on the web manually or by web crawler tools. Useful pages and it's web address are collected to software called integrator. Web crawler counts the information by the statistics and transforms it into structured data set then stores into the database.
Step2 Files Integration. Transfer web pages into been searched into text or XML files and integrate some local files together by integrator. Noise on web pages should be cleaned before applying mining tasks to these files.
Step3 Mining. By web mining and text mining, we can discover interesting topics, clustering of web documents and classification. By data mining, we can get scoring knowledge and know which pages should pay more attention to and decide when to take actions combined with experience.
Step4 Monitoring. Identifying communities of users and information sources [1] on the web. Monitor all useful pages from mining and the new comments.
Step5 Take actions. Since peoples are easily affected by the group consciousness. Proper action such as publish the true facts and guide browser to see the positive side on the key websites, say the right words to invoke good corresponding reactions, advertisement on TV on the right time, communicating with tax payers or press conference etc.
Step6 Feeding Back. After the actions been taken, the effect can be found by remaining. At the same time, the expert's opinion's can be saved as text files or store in the knowledge base. This can make future tax administration wiser and wiser [8].

7. Conclusions

Proposed model for "server side data administration for taxation" is developed using combined approach of web mining, data mining and knowledge management system. This work and paper reporting is part of an ongoing research project namely "Development of Expert System Shell for Server Side Data Management for Large Databases". Development of uniform standard for taxation related activities will create a basic environment for tax submission at user-side and tax administration at back-end side. This model will enable standard methods and procedures so that the entire process of taxation becomes easier for different types of users. The knowledge management platform described here collect information and share with new users. Development of prototype software is also under process which is giving satisfactory result for this model. In next phases of work the functionality of each component of this model like integrator, data mining, web mining, text mining etc will be identified, elaborated and implemented. From the user point of view this approach gives the results of user query consulting knowledge base. User will submit their queries and suggestions to the knowledgebase. From the administrator view point, this approach gives the result for query of administrator. Administrator uses the knowledge base and rules of taxation to sort out the problems regarding taxation. This model gives a new tool to manage tax administration and can also be used for other fields like stock exchange management, economy management, central bureau of investigations, central information commission, election commission, national human right commission, union public service commission, etc.

8. References

[1] Srivastava J.,Desikan P., Kumar V.: Web Mining-Concepts, applications and research directions, In data mining: next generation challenges and future directions. AAAI/MIT press, Boston, MA (2003).
[2] Clinton Alley, Duncan Bentley.: The increasing imperative of cross – disciplinary research in tax Administration, ejournal of tax research (2008) vol.6,no.2,pp.122-144.
[3] Kou G., Y.Peng,etc.: Discovering credit cardholders behavior by multiple criteria linear Programming, annals of operations research vol.135, (2005)261-274.
[4] M. Du. Plessis. Drivers of knowledge management in the corporate environment. International journal of Information management, vol.25, issue 3, June 2005, pp. 193-202.
[5] Colin Shearer, The CRISP-DM model: The new blueprint for data mining, Journal of data warehousing ,vol. 5 No.4 fall 2000 ,pp. 13-22.
[6] http://www.crisp-dm.org/
[7] http://www.adobe.com/government/pdfs/ Tax_Revenue_Auth_SolnBrief_011205.pdf
[8] Xingsen Li, Lingling Zhang, Maoliang Ding, Yong Shi, and Jun Li: A combined web mining model and its application in crisis management, ICCS 2007,PartIII, LNCS 4489,pp 906- 910,2007.
[9] Y.Shi, Data mining. In:M.Zeleny(Ed.), IEBM Handbook of Information Technology in Business,(2002), pp.490-495.
[10] Jack Noonam, Data Mining strategies DM Reviews, July 2000, available online: http://www.dmreview.com/