How Sweet and Ripe are the Fruits? Data Mining Techniques for Classifying and Predicting ‘Quick-Wins’ Direct Capital Investment in Indonesia as One Approach to Business intelligence Orientation and Knowledge Management Scenarios of Indonesian Enterprises

Ali Fauzi
Master of Information Technology Department - Swiss German University, Indonesia

Abstract

The existence of big data of Indonesian FDI (foreign direct investment)/ CDI (capital direct investment) has not been exploited somehow to give further ideas and decision making basis. Example of data exploitation by data mining techniques are for clustering/labeling using K-Mean and classification/prediction using Naïve Bayesian of such DCI categories. One of DCI form is the ‘Quick-Wins’, a.k.a. ‘Low-Hanging-Fruits’ Direct Capital Investment (DCI), or named shortly as QWDI. Despite its mentioned unfavorable factors, i.e. exploitation of natural resources, low added-value creation, low skill-low wages employment, environmental impacts, etc., QWDI , to have great contribution for quick and high job creation, export market penetration and advancement of technology potential. By using some basic data mining techniques as complements to usual statistical/query analysis, or analysis by similar studies or researches, this study has been intended to enable government planners, starting-up companies or financial institutions for further CDI development. The idea of business intelligence orientation and knowledge generation scenarios is also one of precious basis. At its turn, Information and Communication Technology (ICT)’s enablement will have strategic role for Indonesian enterprises growth and as a fundamental for ‘knowledge based economy’ in Indonesia.

Keywords: Direct Capital Investment; Data Mining Techniques; Business Intelligence; Knowledge Management Scenarios
1. Introduction

First priority in Indonesian Investment Roadmap (Presidential Decree No. 16, 2012) is to catch ‘Quick Wins’ Direct Investment (QWDI), a.k.a. low hanging fruits’ investment (See Figure 1). On the following/paralleling steps, Indonesia must also prioritize on: Infrastructure, foods and energy sectors, large capital/large impact investments, knowledge based economy investment. As an initiative or solution aimed to yield rapid positive results with minimal effort (Eustance, 2008), QWDI’s contribution to enable ignitions for those further priority steps is still high. In this study, QWDI samples comprise almost all business sectors in all regions in Indonesia ranging from agriculture/plantation, mining, industry, trade and services. We have found quite significant differentiation in terms of Employment Creation Index (ECI-employment creation times per capita income compared to total capital investment amount) and Smart Capital Index (SCI-intangible capital amount, i.e. Intellectual Property/IP matters: training and development, software/application development, initial research projects, etc. compared to total capital investment amount) of DCI.

2. Backgrounds

Several facts regarding QWDI implantation in Indonesia as of today’s issues have been defined as follow: 1. Lack of quantification/rating of these ‘quick wins/low hanging fruits’ investments based on some fundamental externality factors, e.g. job creation, export potential, transfer of technology, forward/backward linkage potentials, 2. Need for quick profiling/classification based on possible criteria: source of capital (by country), business
sector, startup/expanding company state, location, technology orientation, etc., 3. Big data of overall capital direct investment make such ‘pattern’ somehow unclear to be developed, 4. Lack of information on how business intelligence and knowledge generation scenarios have been done by Indonesian enterprises. Such pool of data as foundation for examination is available currently from the Data Center of BKPM (Investment Coordinating Board of Indonesia) as one of business support agencies in Indonesia. Indeed, the intention to see (somehow to urge) whether DCI has really initialized technology and knowledge spillovers (Petters, 2009) and innovation, and whether it has managed to exploit the power of Information and Communication Technology (ICT) in Indonesian business fields will somehow be the central issue of our current study and hopefully for further studies.

FDI or DCI is not a panacea for economic growth and employment creation (Ernst, 2005). So far, rule-of-thumb consideration for DCI might be only based on the amount of investment. In practice, large investment used to be roughly associated to large externalities. But large capital (internality) doesn’t always mean high magnitude of externalities, i.e. job creation, high export market penetration, large transfer of technology potential, etc.

3. Data Mining Techniques for QWDI Clustering and Prediction

Use of Data Mining techniques is said quite powerful to give some precious information and knowledge building based on very large data (which is sometimes unstructured, patterns-non-visible) available (Tan, 2006).

First Technique: Clustering using K-Mean Method

![Figure 2 –QWDI Clustering process using K-Mean Method in Rapidminer](image)
Mostly used querying tools and statistical analysis is said only can read the lines, but somehow are failed to see what actually exist between the lines. The availability of some predictive analytics tools or namely Open Source Business Intelligence, which began with the conception of On Line Analytical Processing (OLAP) System, has been very helpful to enable us to deal with such situation. In this study we have been able to use some data mining techniques using Rapidminer which was first developed in 2001 by some researcher at the Artificial Intelligence Unit of the Dortmund University of Technology. The Rapidminer 5.0 or older version is still available as open source software under certified Open Source License (OSL), while the newest version Rapidminer 6.0 (with some feature enhancement) is now released with commercial lines.

Following the need to have QWDI rating based on two externality factors: Employment Creation Index (ECI) and Smart Capital Index (SCI), firstly we use clustering technique by K-Mean Method. The data available for this process contains around 3783 rows of data, which was sorted by focusing only the 2 attributes, i.e. ECI and SCI (other attributes, i.e. country name, location and line of business are left behind for a while). We have been preprocessing this data by excluding some outliers, which was less than 0.1 % (40 rows) of existing data. The process and graphic result from Rapidminer could be seen in Figure 2.

Figure 3–QWDI Clusters: Sweet, Sour, Bitter, Ripe
Second Technique: Classification and Prediction using Naïve Bayesian

From the clustering result we have been given 4 clusters that will be differing QWDI by ECI-SCI combination (Figure 3). This result indeed is actually beyond our expectation. At the beginning we have expected there will be such high ECI-high SCI cluster (supposedly will be at the right-top of the chart/figure 3), but such cluster is in fact does not exist. Following the fact, by the given result then we have consequently given labels for each cluster formed. The ‘Sweet QWDI’ (high employment creation-in red color) should be most favorable than ‘Sour QWDI’ (middle employment creation-in light green color). Most of QWDI seems to be shown up as ‘Bitter QWDI’ (low employment creation-in blue color), but that was being paid somehow, in case many of them also tend to be ‘Ripe QWDI’ (high smart capital-in soft green color). This result from data clustering and further by giving labels to Direct Capital Investment (DCI) is indeed one of noble findings of this study. Such clustering and labeling of DCI seems has not been exercised or proposed by previous studies or researches.

After we made such clustering which means also assigning labels (Sweet, Sour, Bitter, Ripe) for every data row, we have then exploited other attributes of the data, i.e. country of origin, location and line of business of QWDI investment projects. These attributes are essential for the next task to be accomplished, which is to make classification, and furthermore for predictive forecasting. Classification was done using Naïve Bayesian method (which also widely used for text processing). First, we were sending Training Data Set (the same 3783 rows data, with more attributes involved, i.e. country of origin, location, line of business and Labor-Smart Capital State). Sample of our Training Data Set (first rows) is in Figure 4. to Naïve Bayesian Classification or processor, than we examine a given Sample Data Set to follow that classification rule. Process in Rapidminer could be seen in Figure 5.

<table>
<thead>
<tr>
<th>Country Name</th>
<th>Line of Business</th>
<th>Location</th>
<th>Labor-Smart Capital State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thailand</td>
<td>Industri kaptai dan perahu</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Bitter</td>
</tr>
<tr>
<td>Thailand</td>
<td>Industri cop, perdagangan beras, kegiatan konsultasi</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Bitter</td>
</tr>
<tr>
<td>R.R. China</td>
<td>Jasa pengurusan transportasi</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Bitter</td>
</tr>
<tr>
<td>Gabungan Negara</td>
<td>Portal Web</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Bitter</td>
</tr>
<tr>
<td>Gabungan Negara</td>
<td>Perdagangan besar</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Bitter</td>
</tr>
<tr>
<td>Gabungan Negara</td>
<td>Industri semen dan industri mortar atau beton sapi</td>
<td>Banten</td>
<td>Bitter</td>
</tr>
<tr>
<td>India</td>
<td>Industri pakaian jadi (konveksi) dari tekstil</td>
<td>Jawa Barat</td>
<td>Bitter</td>
</tr>
<tr>
<td>R.R. China</td>
<td>Perdagangan besar</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Ripe</td>
</tr>
<tr>
<td>Singapore</td>
<td>Industri karet remah (crump rubber)</td>
<td>Sumatera Barat</td>
<td>Bitter</td>
</tr>
<tr>
<td>Gabungan Negara</td>
<td>Industri semiconductors dan komponen elektronik lain</td>
<td>Jawa Barat</td>
<td>Bitter</td>
</tr>
<tr>
<td>Gabungan Negara</td>
<td>Industri semiconductors dan komponen elektronik lain</td>
<td>Banten</td>
<td>Bitter</td>
</tr>
<tr>
<td>Gabungan Negara</td>
<td>Kegiatan konsultasi manajemen bisnis</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Ripe</td>
</tr>
<tr>
<td>Luwesburg</td>
<td>Industri produksi farmasi</td>
<td>Jawa Barat</td>
<td>Bitter</td>
</tr>
<tr>
<td>Luwesburg</td>
<td>Perdagangan beras</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Bitter</td>
</tr>
<tr>
<td>Kepa Selayar</td>
<td>Penggataan Krak dan industri mortar atau beton asap</td>
<td>Sumatera Utara</td>
<td>Bitter</td>
</tr>
<tr>
<td>Singapura</td>
<td>Jasa perawatan dan sejenis usaha tanpa hak</td>
<td>Daerah Khusus Ibu Kota Jakarta</td>
<td>Bitter</td>
</tr>
<tr>
<td>Singapura</td>
<td>Industri borbali beras dari bengkak buatan beras</td>
<td>Kepulauan Riau</td>
<td>Bitter</td>
</tr>
<tr>
<td>Uni Emirat Arab</td>
<td>Pembingungan tanah kedok</td>
<td>Aceh</td>
<td>Bitter</td>
</tr>
<tr>
<td>Malaysia</td>
<td>Pertambangan batubara</td>
<td>Aceh</td>
<td>Bitter</td>
</tr>
</tbody>
</table>

Figure 4 – Training Data Set Sample (Total: 3783 rows)
The classification rule will further assume what Labor-Smart Capital State (as key attribute: Sweet, Ripe, Sour or Bitter) an investment (which comes from certain country, in certain line of business, in certain region) has. This classification/prediction is indeed the core of data mining techniques in this study which will bring somehow more valuable findings.

Making a prediction from classification rule is then could be done. Here, a sample data set could be given to the classifier. Our sample data set is shown in Figure 6. As an illustration we give 2 attributes (source country name and line of business) all the same for every region to be exercised. Supposedly, a German businessman aims to build a DCI project in Indonesia in hotel and cottage industry, and he wants to have an idea of job-technology mode to be applied. For this need, he simply can examine for every possible location in Indonesia. By this classifier, a set of confidence level for every label are given. One of the results as an example, that in Bengkulu, he found out that a relatively more business intelligence orientation (‘Ripe’ confidence level is 0.451, compared to only 0.156 in Banten) has been applied by previous investors. By decision/policy maker side, i.e. central or local government, this sector should be more preferred because confidence levels for Sweet, Sour and Ripe are relatively high (respectively 0.007, 0.028, and 0.451).

From the illustration, the prospective hotel investor should further verify and examine the fact that sort of business intelligence
orientation would be done as prompted by quick prediction. In his intended location, it seems that some hoteliers have been investing significant amounts of their investment to sort of hotel management software, central reservation system or other enhanced business intelligence features. It could also be assumed from these facts that such business in the region has good prospective market in the future and somewhat tough competition forecast, that hoteliers did some innovations/modernizations.

Further variations of prediction could be made by selecting which attributes to be set, and which attribute being exercised. We can see that from this data mining technique alone, which is to build such classifier as one of our research result, a very wide variety of findings and further sources of analytics are possible. For this sample of interest, we previously did not have any idea or assumption of what is really happen in reality. But, after doing some examinations/exercises using the tools, then the results are quite close to the reality according to what we’ve informed.

4. Technology vs. Employment

Correlation between technology development (innovation) and employment creation is somewhat viewed as being substitutive or being applied as synergy. At substitutive/leverage meaning, when we decide to introduce new technology, one thing that should be taken is downsizing, i.e. reducing the employment. At the meaning of synergy, the increased productivity resulting from technological enhancement may not threaten employment (Hernandez, 2001). Innovation instead could be a motor to increased competitiveness and furthermore overall economic growth. Higher income for more skillful employee and increasing consumption levels at its turn will compensate the jobs lost and would create new employment.
The synergy seems to be the favorable situation that we are expecting of. However, one thing that should be assured for such situation is that technology enhancement should be done in a high readiness environment which mainly and basically supported by one of important factors, i.e. Information and Communication Technology (ICT) as general purposes tools providing enablement for every sector of interest and the other support is human resources responsiveness for technology enhancement.

From Figure 8 (derived from our data query), we can see the profile of Indonesian QWDI enterprises. With relatively uniform spread of business categories, we only can see what business lines are altogether contributing as current Indonesian economic backbone.

Figure 9 is a brief resume from one of our data mining work in this study (clustering), explaining how these enterprises are classified into 4 groups based on labor-technology mode or business intelligence orientation (Sweet, Sour, Bitter and Ripe). It is notable that most of Indonesian enterprises (QWDI) have not only relatively low employment creation, but also relatively low initial orientation for business intelligence (Bitter state, 78% of the samples). Significant portion has shown a quite high initial effort for business intelligence, with employment creation is relatively low (Ripe state, 17%).
Still from the query and statistical tools, we have examined some business fields that already have significant business intelligence orientation (See Figure 4, measured in index/percentage), most of them are from service sector (educational services, travel agent, civil and architecture services, head office and management consultant, programming and computer consultancy, food and beverage services and civil construction). Besides the trading sector (grocery and retail), some certain mining sectors (coal mining, mining services) as well have shown significant orientation for business intelligence.

5. IT Enablement Importance

From IT Strategic Planning point of view, those above initial facts could open eyes to realize that there are still wide spaces for IT enablement (as a basic support for knowledge based economy) in Indonesia and that we can identify each sector characteristics to apply such IT enablement. But before we move further to business intelligence orientation and knowledge management scenario, it is important to say that clustering (using data mining) and further making some query, statistical analysis are somehow still very limited to tell the complete story and to cope with growing needs of further analysis.

6. Knowledge Generation Scenarios of Indonesian Enterprises

We have learnt so far, the process of mining the historical data, which are establishing such grouping (clustering) and classifying the majority of Indonesian businesses, and further generating such predictive modeling. For some categories which have been managed to be revealed (Sweet, Sour, Bitter and Ripe; as measures for business intelligence orientation), it is then interesting to see which knowledge management scenarios are recommended for each category. As we can see from Figure 11 (matrix), each knowledge-generation activities (Keri E. Pearlson, 2010), i.e. Research and Development (R&D), Adaptation, Buy or Rent, Shared Problem Solving, Communities of Practice are being rated for given categories.

As we already also managed to identify some tendencies from data query and clustering, we have been able to make such redlines. “Bitter” investments mostly comprise of basic and intermediate manufacturing industries (e.g. chemical/petrochemical, metal and basic metal, rubber and plastic, automotive, non-metal minerals); they are mostly capital intensive, use relatively low number of employment and mostly applying full adoption of overseas technologies. “Sour” investments mostly are of emerging manufacturing industries (e.g. food and beverages, textile, building materials, modern handicraft and creative arts); they could be capital intensive but also labor intensive, use relatively medium to high number of employment and applying a mix between overseas and local competence technologies. “Sweet” investments mostly come from ‘third-world’ manufacturing industries (e.g. garment, leather and footwear); they are labor intensive and applying very basic technologies. “Ripe” investments tend to be of research oriented industries and market oriented services sectors (e.g. mining, pharmacy, hotel and restaurant, travel agent, educational, construction, retail and general trading, ICT providers); they are technology intensive, use relatively low number of employment and...
must maintain technology or market intelligence enhancement in order to survive and grow.

As we already established above mentioned redlines, further suggestion is all about knowledge management scenarios (knowledge generation processes). A “Bitter” company, for example must prioritize first on R&D (indicated as No. 1 in the column in Figure 11), doing (and funding) researches of many aspects within company’s business process, involving some task force groups, and then secondly hiring new personals or contracting consultants (Buy or Rent). The company must also maintain adaptation along the product lifecycle (third priority), before being able to establish communities in practices platform (fourth) and furthermore sophisticated shared problem solving. This roadmap must regard also the nature of expenditure (example; ‘bitter’ means low labor expenses) of businesses under concern, which then suggest R&D (usually relatively high expense) as top priority.

7. Conclusion and Remarks

The use of data mining techniques is proved to be powerful in processing our large data of direct capital investment in Indonesia and more possibilities of having the correlation between more involved factors/attributes. This research has proposed idea of labeling, classification and prediction of one big portion of capital direct investment in Indonesia, which is Low Hanging Fruits/Quick Wins Direct Investment. Compared with some of similar researches (which use data mining approach) in CDI/FDI field, namely the impact of financial development on FDI (Korgaonar, 2012), and data mining in banking and its application (Pulakkazhy, 2013), this research has approached the correlation between variation of CDI/FDI (source country of capital, line of business, and location, or other variables possible) with quite new and emerging parameter, which is technology orientation. Though the findings (labeling) and classification rule have to be proven by further examination, this research could be used as a quick prediction tool for technology orientation. The importance of IT enablement along with its wide proposed subjects (e.g. Enterprise Resource Planning (ERP), Data Security, Enterprise Architecture, etc.) at its turn quite important theme for Indonesian knowledge based economy, which will be the core of our further researches.

8. References


