

frilly / October 27, 2015 05:50PM

創業公司怎麼搞巨量資料？

這是一個不說自己搞巨量資料就不好意思和人打招呼的時代。阿里巴巴用巨量資料貸款，百度用巨量資料推廣告。那麼，沒有巨頭的天然優勢，創業公司具體怎麼搞巨量資料？

開搞之前，先明確一下目前巨量資料主要是用來預判，挖掘用戶的需求（也有用過去的巨量資料判斷現在的需求，如巨量資料種菜判斷菜目前的長勢情況）。那麼，預判的依據是什麼？就是所謂樣本，例如我通過前幾年的平均氣溫判斷今年的溫度是高低。

這裡還可以分為三種情況：一種是我的巨量資料軟體自帶了很多樣本可以作為分析依據，一種是我把之前的交易資料作為樣本。第二種較先進，要求系統有自我學習建立樣本模型的能力，也就是邊干邊學。第三種最牛，不光能做出預判，還能告訴你下一步怎麼做，例如一家賣拖拉機的公司，為用戶提供何時在哪裡種植何種作物，犁地的最佳路線。首先說一下資料來源問題，可以是平時經營過程中積攢下來的也可以向別的公司購買（像京東那樣買騰訊的資料，目前國內貴州巨量資料交易中心和DATACOMB很出名，後面細說），技術關鍵：積攢資料同時要根據自己的需求為用戶和產品貼標籤，例如我希望了解和用戶年齡段有關的規律，那麼年齡就是必須採集的一個標籤。積攢資料的方式大致分兩類，一類是圍繞用戶，一類是圍繞產品，將兩者每次交易的資料都採集下來，可用人力，例如飯店服務員每次交易記錄下菜品，性別，滿意度之類，這一步又叫企業資料化。

然後是資料的初步處理：租伺服器，建立自己的資料分析架構，例如日報表月報表這些都OUT了，應該明確適合公司的格式，如果需要這些定製需求的報表，目前流行的巨量資料的解決方案大部分都是以Hadoop為基礎架構。什麼是Hadoop？簡單來說Hadoop是一個分布式計算的解決方案，分布式通俗來說就是把一件事分布到幾台計算機上運行。由多台計算機同時運行和存儲資料，比一台計算機運行速度快，而且如果資料量大了，或者報表複雜導致運算速度慢，只要再加計算機就解決了。

當每台計算機運算完畢後，會把中間結果集中到一台計算機上，再把這些中間結果匯總起來得出最終結果。把手頭的資料進行預處理，包括將不同資料庫的資料導入到一個資料庫中，資料的粗選，分析，分類，會用到EMC的GreenPlum、Oracle的Exadata，以及基於MySQL的列式存儲Infobright等，而一些批處理，或者基於半結構化資料的需求可以使用Hadoop。先別頭大，如果你不是淘寶京東這樣的巨頭沒有那麼巨量資料量可以不這麼麻煩，剔除那些和其他資料差別很大的就行

接下來，就是最核心的，資料挖掘。這一步可以挖掘出你意想不到的信息，比如，買啤酒的顧客通常會買尿布，持續買無香精面霜的顧客通常接下來會買母嬰用品。這些資料怎麼挖？需要用到以下工具：

1. 決策樹，各種可能性的展示圖，主要用來精確分類。例如，樹的根部是雜貨鋪里所有商品，往上走就是第一個分叉，分叉處有一個判斷標準-吃的還是用的，於是所有商品被分開為兩個樹枝，吃的再往上走又出現分叉，於是被分成甜的和鹹的...

2. 神經網路，例如，計算機預判迎面走來的這個軟妹子會買奶茶，因為根據神經網路顯示，上次買咖啡，上上次買酸奶的，這次99.99%的概率會買奶茶。什麼？這貨買的啤酒？那計算機就會根據這次的結果修改神經網路。沒錯，神經網路是具有學習能力的模擬人腦總結教訓的分析模式，他會自己得到一個穩定準確的結果。

3. 關聯規則，例如上文說的買啤酒的一定買尿布。

先看一個簡單的例子，假如有下面資料集，每一組資料ti表示不同的顧客一次在商場購買的商品的集合：

t1: 牛肉、雞肉、牛奶

t2: 牛肉、乳酪

t3: 乳酪、靴子

t4: 牛肉、雞肉、乳酪

t5: 牛肉、雞肉、衣服、乳酪、牛奶

t6: 雞肉、衣服、牛奶

t7: 雞肉、牛奶、衣服

假如有一條規則：牛肉→雞肉，那麼同時購買牛肉和雞肉的顧客比例是3/7，而購買牛肉的顧客過一段時間也購買了雞肉的顧客比例是3/4。這兩個比例參數是很重要的衡量指標，它們在關聯規則中稱作支持度（support）和置信度（confidence）。前者反映了牛肉雞肉組合的吸引力，後者可用來預判消費行為。好了，現指定你想要的支持度和置信度是多少，然後把牛奶雞肉，衣服乳酪等各種組合用程序算一遍，找出支持度和置信度滿足你要求的組合。

其中，假如一條關聯規則的支持度和置信度很高時，不代表這個規則之間就一定存在某種關聯。舉個最簡單的例子，假如X和Y是最近的兩個比較熱門的商品，大家去商場都要買，比如某款手機和某款衣服，都是最新款的，深受大家的喜愛，那麼這條關聯規則的支持度和置信度都很高，但是它們之間沒有必然的聯繫。所以，搞巨量資料既要對專業精通又要對你公司的業務精通，這也是現在公司經常碰到的難題，此事古難全啊。

除了這些，巨量資料還經常用到一些數學工具如，線性回歸，支持向量機等來幫助進行比對和分類。不要小看這些輔助工具，有時候他們是主力。美國警察用電力公司的資料與培養大麻的LED燈用電曲線進行比對，數學工具就是線性回歸，只要發現哪家用電曲線和樣本相符（相符的程度用方差表示，標準老警自己定），直接踹門抓人一抓一個準。這些數學工具很多都是用來分類的，分類在巨量資料中是個大事，例如國外有的門戶網站首頁分類用巨量資料直接篩選出詳細頁使用頻率較高的辭彙作為各個大類的標題。

工具列完了，下面資料挖掘正式開幕：

巨量資料對於什麼時候用人力什麼時候用計算機也能看出來功力，中小公司更應該注意巧妙利用人力，可避免在技術上過多拼財力。比如以上的例子中大家可以看到，相似不相似，是不是關聯，這些標準都是人定的。計算機常負責一些同一個演算法算大批量資料的活。美國巨量資料公司的老大Palantir就因精於此道而著稱。中小公司在這方面也有做的不錯的，最近開始流行的圖片新聞APP 網新聞的策略就很值得借鑒：

網新聞主要是用關係圖，時間軸各種圖片展示更豐富信息，一條新聞，用戶希望在哪一點上獲取更深入的信息是網新聞最想分析的，明確了這個目的又積累了一定用戶行為的資料後，網新聞做了如下工作：把一個新聞，舉個栗子，如釋永信事件，讓小編寫完這條新聞就分析用戶關注這條新聞的點在哪，並且把各種可能性列出來。列出的結果是：

1、用時間軸展示少林寺背後的利益紛爭事件，用關係圖展示各利益相關方的關係，這兩條可以滿足求知慾，用戶也可以說出去秀知識。

2、說明佛門中的禪宗一派特點就是不避世俗。古代就在寺廟做小買賣。進而拓展到律宗，密宗是怎麼回事。

3、傳聞釋永信嫖娼，有情婦，桃色因素是一個。

4、中國法律目前對寺廟住持是怎麼定性的，有哪些法律來約束，觸犯了這些法律會受到什麼樣的懲罰，對事件目前形勢未來預判也是一個吸引人的因素...

這樣列出來後，進行關聯性分析，點擊量高的新聞，一定是具備因素2和因素3？還是因素1和因素3？...網新聞用這種方法得出很多有價值的信息，如用戶對用圖片解釋「經理人採購指數」感興趣還是對展示某經濟指標對自己生活的影響更感興趣，然後用這些資料指導小編擺圖寫稿。

另外，不要覺得分析資料預判用戶需求很屌，其實路邊賣手抓餅的也經常能猜出你這次要不要放辣椒，就是調用的他腦中的巨量資料。可以讓用戶刷打折卡或者是通過QQ定外賣時顯示他平時的偏好。這個方法需要對資料預處理去掉那些和其他資料偏離較大的資料可獲得基本的「用戶畫像」。另外還可以結合其他因素資料豐富「用戶畫像」，例如天氣炎熱或寒冷的情況下用戶的喜好，每到周末是不是多買一些等。

還有一個很重要的使用人力的節點就是先想好你了解什麼。例如，買啤酒的顧客會買尿布這個例子中，資料挖掘工程師其實提前就想到有必要查看一下啤酒和其他商品的關聯，才有了這個發現，而不是計算機自己去找這兩者關係。這需要洞察力！福布斯網站發文稱「首先制定好策略，然後奔著結果找出答案」，並由此得出「問題比答案重要」的結論。

使用巨量資料的流程就這些了。總結一下哪些地方使用人力：收集什麼樣的資料需要人來做決定，想好要弄清楚哪些規律需要人來做決定。電腦主要用來存儲，以及用數學工具來進行具體計算。哈哈，原來巨量資料不是巨頭們買一個塞滿房子的超級計算機把海量資料統統輸進去然後「度昂」一下就出來結果了！我們小公司一樣可以搞啊！

由於「問題比答案重要」，再說一下巨量資料領域提出的幾個有代表性的「問題」。既然除了利用之前積累的資料了解用戶偏好向用戶推銷這次的產品，還可以用巨量資料讓用戶方便的消費來提高用戶體驗：快餐業的視頻分析。該公司通過視頻分析等候隊列的長度，然後自動變化電子菜單顯示的內容。如果隊列較長，則顯示可以快速供給的食物；如果隊列較短，則顯示那些利潤較高但準備時間相對長的食物。這個案例門檻極低，只要人為設定隊伍多長算長，做食物的速度多塊算快即可，也可以人工標定一下哪些是快食哪些是慢食。

不僅可以預判需求予以滿足，營銷時還可以定點營銷。此類打法常常是從別的機構購得資料或者連錢都不花從政府開放的資料中拿，一家領先的專業時裝零售商，通過當地的百貨商店、網路及其郵購目錄業務為客戶提供服務。公司希望向客戶提供差異化服務，如何定位公司的差異化，他們通過從Twitter 和Facebook 上收集社交信息，更深入的理解化妝品的營銷模式，隨後他們認識到必須保留兩類有價值的客戶：高消費者和高影響者。然後通過免費化妝服務，對這兩類消費者進行精準的宣傳。這個案例中，需要用到的工具有用於歸類的線性回歸，聚類分析等。

資料的來源問題：由於沒有巨頭那樣的資料積累，中小企業經常想到購買資料，或使用政府公開的免費資料。目前國內掌握巨量資料的巨頭不夠開放，形成了讓仁人志士頭大的「資料堰塞湖」。雖然也有開放資料的，例如中國首個巨量資料開放平台DataComb，對於掌握一些資料的公司來說也多了一些資料變現的渠道，然而對疏通「堰塞湖」並沒有太大作用。也有好消息，8月19日，國務院常務會議審議通過《關於促進巨量資料發展的行動綱要》，全面公開應該就在眼前了。在美國最重要的資料開放平台就是奧巴馬政府在2009年推出的Data.gov，奧巴馬同學對搞巨量資料一向蠻拼的。

另外，創業公司切記，巨頭們整天嚷嚷的「巨量資料不必追求精確，犧牲精確性可以換取效率和更多規律的發現」，可問題是那是在資料量異常大的情況下，對中小企業資料量不太大時務必要精確。

總的來說，巨量[[url=http://www.finereport.com/tw/](http://www.finereport.com/tw/)]資料分析[[url](#)]可謂是兵無常勢，水無常形，真正厲害的資料挖掘大師是碰到什麼樣的形勢知道用什麼樣的工具來解決的人。中小企業玩巨量資料切記：1.不炒概念。2.不燒錢。切實用巨量資料提升企業競爭力才是王道。

4500+企業選擇FineReport報表與 [[url=http://www.finereport.com/tw/](http://www.finereport.com/tw/)]BI[[url](#)]

[[url=http://www.finereport.com/tw/](http://www.finereport.com/tw/)]商業智慧[[url](#)]工具【免費下載】

opensource開發，類excel設計，全方位異質資料庫整合，資料填報、Flash列印、權限控制、行动應用、客制化、交互分析、報表協同作業管理系統。
分享自：hong
