自動資料鏈結分析技術之開發與應用

楊境榮、陳鴻文

E-mail: 9422529@mail.dyu.edu.tw

摘 要

資料探勘(data mining)技術的主要目的是從大量資料中，自動分析出隱藏的知識規則。然而在一般的關聯分析(association analysis)及序列樣式分析(sequential pattern analysis)的研究中，往往在過濾大項目集合(large itemsets)時，由於門檻值設定的緣故，使得一些出現頻率較低、實質上卻具有高價值的資料項目不易被發掘出來，或引入過多可信度太低的關聯法則。為了嘗試解決此一問題，有些研究提出多重支持度的做法，但仍需事先針對資料進行適當的分類和分析瞭解，再個別地設定適當的最小支持度；然而著重於頻率門檻值的研究，似乎仍無法廣泛處理真實世界的問題。此外，鑑於以往的鏈結分析(link analysis)，除了特定問題，如Google搜尋引擎使用PageRank技術來評估網頁重要性權重的超鏈結分析(hyperlink analysis)，一般偏向將資料利用視覺化工具呈現後，仰賴專家主觀目測其規律性；但對於高複雜的關係圖形，如食物鏈、電話通連、交通網、人際關係等的分析，則往往成效不彰。 在一些社會網絡分析研究中指出，出現頻率高的強鏈結(strong tie)雖能聯繫具有相同特質的個體，以形成群聚現象；然而透過弱鏈結(weak tie)則可聯繫不同的群體，即為不同群體之間的橋樑(bridge)。所以在傳播訊息時，往往是透過弱鏈結的運作，來將訊息傳播得更遠。因此基於弱鏈結的觀念，本研究嘗試利用圖形理論中尋找二元連通元件(biconnected component)及強連通元件(strongly connected component)演算法，將所找出的二元連通元件或強連通元件(節點數大於2者)鏈結從圖形中刪除後，再依據剩餘鏈結在結構上的重要性，經過類似分類處理，可找出存在於圖形結構中的潛在弱鏈結(包含弱鏈結)及關鍵弱鏈結路徑(critical path formed by potential weak link)。如此將可彌補關聯法則分析演算法中，無法探勘出稀少關鍵資料規則的缺失，進而找出具有高可信度、高價值性的鏈結。 為了驗證所提出自動化鏈結分析演算法的可行性和效能，本研究採用了美國聯邦能源管制委員會(Federal Energy Regulatory Commission)在調查安隆(Enron)公司時，所公佈的真實電子郵件資料集(Enron Email Dataset)。實驗結果確可有效地分析出人際關係叢集及弱鏈結橋樑等特徵，故可進一步提供社會學者據以分析社會網絡、政府偵防組織犯罪，或網路管理者設計郵件系統功能及辨識垃圾郵件等。

關鍵詞 : 資料探勘 ; 鏈結分析 ; 關聯分析 ; 弱鏈結 ; 自動化鏈結分析

目錄

參考文獻

[1]吳寶秀，「台灣製造業員工個人社會網絡分析」，東海大學社會學研究所碩士論文，民國79年。

[2]林傑彬與劉明德，「資料採掘與OLAP理論與實務」，文魁資訊股份有限公司，民國91年。

[3]胡守仁譯，「連結:混沌、複雜之後，最具開創性的小世界理論」，天下文化，民國91年。

[4]鄭讚源，「社會網絡、社會整合與學習家庭:機會與挑戰」，載於教育部主辦《學習型家庭理論與實務研討會資料》，台灣師範大學，民國88年。

[5]蕭新煌與龔宜君，「東南亞台商與華人之商業網絡關係」，中央研究院東南亞區域研究所，民國87年。

[6]N.M. Adams, G. Blunt, D.J. Hand and M.G. Kelly,“ Data mining for fun and profit, ”Statistical Science, Vol. 15, No. 2, pp. 111-131, 2000.

[7]P. Adriaans and D. Zantinge,“ DATA MINING, ”ADDISON-WESLEY, 1999.

[8]R. Agrawal, T. Imilienski and A. Swami,“ Mining Association Rules between Sets of Items in Large Databases, ”In Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 207-216, 1993.

[9]M. J. A. Berry and G. Linoff, Data Mining Techniques: For Marketing Sale and Customer Support, John Wiley & Sons, 1997.
[10]S. Brin and L. Page, “ The Anatomy of Large-Scale Hypertextual Web Search Engine, ” In Proceedings of the 7th InternationalWorldWideWeb Conference, pp. 107-117, 1998.
[11]P. Cabena, P. Hadjinian, R. Stadler, J. Verhees and A. Zanasi, Discovering Data Mining From Concept to Implementation, Prentice-Hall Inc, 1997.
[12]M. S. Chen, J. S. Park and P. S. Yu, “ Efficient Data Mining for Path Traversal Patterns, ” IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. 2, pp. 209-221, 1998.
[13]F. L. Chung and C. L. Lui, “ A post-analysis framework for mining generalized association rules with multiple minimum supports, ” Workshop Notes of KDD'2000 Workshop on Post-Processing in Machine Learning and Data Mining, pp.9-14, 2000.
[14]U. Fayyad, S. G. Piatetsky and P. Smyth, “ From data mining to knowledge discovery in database, ” AI magazine, Vol. 17, pp. 37-54, 1996.
[15]L. Garton, C. Haythornthwaite and B. Wellman, “ Studying Online Social Networks, ” Journal of Computer-Medicated Communication, Vol. 3, No. 1, 1997.
[16]M. S. Granovetter, “ The strength of weak ties, ” American Journal of Sociology, Vol. 78, pp. 1360-1380, 1973.
[17]J. Han and M. Kamber, Data Mining : Concepts and Techniques, John Wiley & Son, 2001.
[18]M. R. Henzinger, “ Hyperlink Analysis for the Web, ” IEEE INTERNET COMPUTING, Vol. 5, pp. 1089-7801, 2001.
[19]M. P. Johson and R. M. Milardo, “ Networkinterference in pair relationship : A social psychological recasting of Slater's (1963) theory of social regression, ” Joural of Marriage and the Family, Vol. 46, pp. 893-899, 1984.
[20]C. Kleissner, “ Data mining for the enterprise, ” In Proceedings of the 35th Hawaii International Conference, Vol. 7, pp. 295-304, 1998.
[21]D. Knoke and J. H. Kuklinski, Network Analysis, Beverly Hills: Sage Publications, 1982.
[22]R. Lempel and A. Soffer, “ PicASHOW: Pictorial Authority Search by Hyperlinks on the Web, ” In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 438-448, 2001.
[23]B. Liu, W. Hsu, and Y. Ma, “ Mining Association Rules with Multiple Minimum Supports, ” In Proceedings of the 1999 International Conference on Knowledge Discovery and Data Mining, pp. 337-341, 1999.
[24]P. Marsden, “ Network Data and Measurement, ” Annual Review of Sociology, Vol. 16, pp. 435-463, 1990.
[25]S. Milgram, “ The small world problem, ” Psychology Today, Vol. 2, pp. 60-67, 1967.
[26]S. Wasserman and K. Faust, Social network analysis: Methods and application, Cambridge University Press, 1997.
[27]D. J. Watts and S. H. Strogatz, “ Collective Dynamics of Small-World, ” Networks, Vol. 393, pp. 440-442, 1998.
[28]M. A. Weiss, Data Structures and Algorithm Analysis in C. The Benjamin / Cummings Publishing Company, 1993.
[29]P. C. Wang, “ Visual Data Mining, ” IEEE Computer Graphics and Applications, Vol. 19, No. 5, 1999.
[30]C. Westphal and T. Blaxton, Data Mining Solutions, John Wiley &Sons, 1998.
[31]H. Yun, D. Ha, B. Hwang and K. H. Ryu, “ Mining Association Rules on Significant Rare Data Using Relative Support, ” The Journal of Systems and Software, Vol. 67, pp. 181 – 191, 2003.
[32]Enron Email Dataset， http://www-2.cs.cmu.edu/~enron/ [33]Enron Dataset， http://www.isi.edu/~adibi/Enron/Enron.htm