Design and Implementation of a Generator of Large, Dense, or Sparse Databases to Test Association Rules Miner

By
Hussien Al-Khafaji
Prof. Dr. Alaa Al-Hamami
Prof. Dr. Abbas F. Abdul-Kader

Abstract

Association rules discovery has emerged as a very important problem in knowledge discovery in database and data mining. A number of algorithms is presented to mine association rules. There are many factors that affect the efficiency of rules mining algorithms, such as largeness, denseness, and sparseness of databases used to be mined, in addition to number of items, number and average sizes of transactions, number and average sizes of frequent itemsets, and number and average sizes of potentially maximal itemsets. It is impossible to change present real-world database's characteristics to fairly test and determine the best and worst cases of rule-mining algorithms, to be efficiently used for present and future databases. So the researchers attend to construct artificial database to qualitative and quantitative presence of the above mentioned factors to test the efficiency of rule-mining algorithms and programs. The construction of such databases consumes very large amount of time and efforts. This research presents a software system, generator, to construct artificial databases commensurate with the user's object of desire. The generator depends on mathematical model that simulates the behavior of real-world databases.

Keywords: KDD, Knowledge Discovery, Data Mining, Association Rule.

الخلاصة

لمناجم البيانات أغراض عديدة، والكشف عن قواعد الارتباط واحد من أهم أغراضها. هنالك عوامل عديدة تأثر في كفاءة الخوارزميات المعدة للكشف عن قواعد الارتباط في قواعد البيانات منها كبر، كثافة، وتناثر القاعدة. كذلك عدد ومعدل حجم الصفقات، عدد العناصر، عدد ومعدل حجم مجموعات العناصر المتكررة، عدد ومعدل حجم المجاميع اللاجزئية. إن قواعد بيانات التطبيقات الحقيقية، (إن وجدت)، لا تسمح بأجزاء اختبارات منصفة لخوارزميات الكشف عن قواعد الارتباط لافتقارها إلى توافر وترابط العوامل المذكورة أعلاه لتثبين الحالات الأسوأ والأحسن لهذه الخوارزميات. إن مثل هذه الخوارزميات يجب أن تختبر لمختلف حالات البيانات لقياس ملائمتها وقدرتها على مواكبة تغير عوامل التأثير نظراً لحجم قواعد البيانات التزايد باستمرار. نظراً لذلك عمد الباحثون إلى بناء قواعد بيانات اختباريه ليتسنى لهم إجراء هذه الاختبارات المتنوعة بتنوع العوامل المأثرة في الكفاءة. إن بناء هذه القواعد يحتاج جهود كبيرة لفرق عمل مختصة، علماً إن تغيير قيمة أي عامل مؤثر بالزيادة أو النقصان يغير التوزيعات الإحصائية للعناصر والصفقات. البحث يقدم نظاماً برمجياً، مولداً لقواعد البيانات، لإنشاء قواعد بيانات حسب طلب المستخدم لا يكلفه إلا زمن التنفيذ. وبذلك يمكن توليد قواعد مختلفة تسمح بإجراء اختبارات متنوعة ومنصفة.

Hiding", IBM Systems Journal, Vol. 35, No. 3&4, 1996.

5. Boney Laurence, Tewfik Ahmed H., hamdy Khaled N., "Digital Watermarking for Audio Signals", An Internet survey, 1996.

6. Cox Ingemar, Kilian Joe, Leighton Tom, Shamoon and Talal,"A Secure Robustfor Multimedia", An Internet survey, 1998.

7. Czerwinski Steve, Fromm Richara and Hodes Toad, "Digital Music Distribution and Audio Watermarking", Computer Science Division / University of California, 1997.

8. Fridrich Jiri, Baldoza Arnold C. and Simard Richard J. "Robust Digital Watermarking Based on Key-Dependent Basis Function", Information Hiding: Second International Workshop, Proceedings, and vol. 1525 of Lecture Notes in Computer Science, Springer, 1998, pp.32-47.

9. Gruhl Daniel, Lu Anthony, and Bender Walter, "Echo Hiding", Information Hiding First International Workshop, Proceedings, vol. 1174 of Lecture Notes in Computer science, Springer, 1996, PP. 1-5.

10. Jajodia Sushil, "Advances in Information Security", Kluwer Academic Publishers, 2000.

11. Jarol Scott, "Visual Basic Multimedia", Edited by Jeff Duntemann, Galgotia, publication PVT, ITD., 1994.

12. Katzenbeisser S. and Petitcolas F., "Information Hiding Techniques for Steganography and Digital watermarking", Artech House, 2000, USA.

13. Kientzle Tim, "A Programmer's Guide to Sound", An imprint of Addison Wesley Longman-Inc, 2000.

14. Kurrbayashi Minoru & Tanka Halsukazu ,"A watermarking Scheme Based on Characteristics of addition among DCT coefficients", Information Security, Third International Workshop , ISW 2000 Wollongong, Australia, December, Proceeding Springier, 2000.

15. Lacy Jack, quackenbush Schuyler R., Reibman Amy, Snyder James H., "Intellectual Property Protection Systems and Digital Watermarking", Information Hiding: Second International Workshop, Proceedings, Vol. 1525 of Lecture Notes in Computer Science, Springer, 1998, pp.158-168.

16. Password Philips Research, "Digital Signal Processing for Audio and Speech", An Internet survey, 2000.

17. Qian Lintiaan, N. Klara, "Non Invertible Watermarking Methods for Mpeg Encoded "An Internet survey, 1998.

18. Zian Zhao, "Applying Digital Watermarking Techniques to Online Multimedia Commerce", An Internet survey, 1998.

19. Zhao Zian, "A Digital Watermarking System for Multimedia Copyright Protection", An Internet survey, 1998.

f. Mean Watermarking

a. Changing the size of the file will affect embedding/extraction of the watermark.

b. Apply on stereo and mono files.

c. Changing the watermarking will not affect the time of embedding/extraction.

d. The waves of audio files do not differ after embedding.

e. The sound of a file does not differ after embedding.

f. The watermark is of type number.

g. Changes one byte of the sound file.

13. Conclusions:

1. WES is applied to Wave of Windows, Mid and Songs.
2. ASCII code watermarking is the method most affected by noise according to the methods used by WES system, because the number of Bytes depends on the characters of watermarking.
3. WES does not embed the watermark in the header.
4. WES uses multiple watermarks, which means multiple users. Each user has a special watermark, and this is one requirement of watermark.
5. The Statistical watermark takes more time if the file size is different, especially the Maximum Watermark.
6. BIT and sample change one bit of byte sound until end of watermarking.
7. WES is applied to stereo and mono for all methods, but in ASCII watermarking it is preferred to apply to the stereo two channels rather than mono.

14. Recommendations:

1. Perform compression to sound after embedding the watermarking.
2. Perform WES to another format of Wave.
3. Improve our system by using it on video.
4. Cipher the watermarking before embedding in sound file.
5. Use other methods of watermarking like echo hiding, spread spectrum and phase coding as discussed above.

References:

1. Anderson Ross, "Stretching the Limits of Steganography", Information Hiding: First International Workshop, Proceedings, vol. 1174 of Lecture Notes in Computer science, Springer, 1996, PP. 39-48.
2. Arnold Michael, "Audio Watermarking: Features, Application and Algorithms", An Internet survey, 1998.
3. Bassia P. and Pitas I., "Robust Audio watermarking in the Time Domain", An Internet survey, 1998.
4. Bender W., Gruhl D., Morimoto N. and A. Lu. , "Techniques for Data

e. The sound of a stereo file does not change after embedding.

f. The watermark is of type string.

g. The bytes sound that affects are equal to the characters of the watermark.

b. Sample Watermarking

a. Changing the size of the file will affect embedding/extraction of the watermark.

b. Apply on stereo and mono files.

c. Changing the watermarking will not affect the time of embedding/extraction.

d. The waves of audio files are not affected after embedding.

e. The sound of a file does not differ after embedding.

f. The watermark is of type string.

g. Changes only one bit of the sound until the end of watermark.

c. Frequency Watermarking

a. Changing the size of the file will affect embedding/extraction of the watermark.

b. Apply on stereo and mono files.

c. Changing the watermarking will minimize the time of embedding/extraction.

d. The waves of audio files do not differ after embedding.

e. The sound of a file does not differ after embedding.

f. The watermark depends on the affected frequency of sound file.

g. Changes five bytes of the sound file.

d. Maximum Watermarking

a. Changing the size of the file will affect embedding/extraction of the watermark.

b. Apply on stereo and mono files.

c. Changing the watermarking will not affect the time of embedding/extraction.

d. The waves of audio files do not differ after embedding.

e. The sound of a file does not differ after embedding.

f. The watermark is of type number.

g. Changes one byte of the sound file.

e. Minimum Watermarking

a. Changing the size of the file will affect embedding/extraction of the watermark.

b. Apply on stereo and mono files.

c. Changing the watermarking will not affect the time of embedding/extraction.

d. The waves of audio files do not differ after embedding.

e. The sound of a file does not differ after embedding.

f. The watermark is of type number.

g. Changes one byte of the sound file.

Ministry of Higher Education and Scientific
Iraqi committee of Computers and Informatics

Computers Journal

A Refereed Scientific Journal

NO . 40

2002