Data Mining

Association Rule Mining

Definisi

- Bagian dari *rule-based machine learning*
- Metode untuk menemukan hubungan yang menarik antara variabel dalam database besar.

Rule-based machine learning

- Istilah dalam ilmu komputer yang dimaksudkan untuk mencakup metode pembelajaran mesin apa pun yang mengidentifikasi, mempelajari, atau mengembangkan 'aturan' untuk menyimpan, memanipulasi, atau menerapkan aturan tersebut.

Tujuan

- Bertujuan untuk mengidentifikasi aturan-aturan kuat yang ditemukan dalam basis data menggunakan beberapa ukuran ketertarikan.
- Pendekatan berbasis aturan ini juga menghasilkan aturan baru karena menganalisa lebih banyak data.

Rule

- {tua, merokok} → {kangker}
- {harga murah, ukuran kecil} → {snack, bumbu}
- {sumber tidak jelas, bahasa provokatif} → {hoax}
- {penduduk sedikit, akses jalan terbatas} → {tidak ada rumah sakit}
- {kata-kata kasar, kekecewaan} → {sentimen negatif}

Konsep Association Rule Mining

- Support
- Confidence
- Lift
- Conviction

Support

- Dukungan adalah indikasi seberapa sering itemset muncul di dataset.
- Dukungan X sehubungan dengan T didefinisikan sebagai proporsi transaksi t dalam dataset yang berisi itemset X.

$$\operatorname{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

Confidence

- Keyakinan adalah indikasi seberapa sering aturan itu terbukti benar.

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

Lift

- Jika aturan itu memiliki angkat 1, itu akan menyiratkan bahwa kemungkinan terjadinya anteseden dan yang dari akibatnya adalah independen satu sama lain.
- Ketika dua peristiwa terpisah satu sama lain, tidak ada aturan yang dapat ditarik yang melibatkan kedua peristiwa tersebut.

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)}$$

Kondisi Lift

- Jika lift > 1, menunjukkan sejauh mana kedua kejadian tersebut bergantung satu sama lain, dan membuat aturan-aturan tersebut berpotensi berguna untuk memprediksi konsekuensi dalam kumpulan data masa depan.
- Jika lift <1, menunjukkan kejadian tersebut saling menggantikan satu sama lain. Ini berarti bahwa kehadiran satu kejadian memiliki efek negatif pada kehadiran kejadian lain dan sebaliknya.
- Nilai dari lift adalah bahwa ia menganggap baik kepercayaan aturan dan keseluruhan kumpulan data.

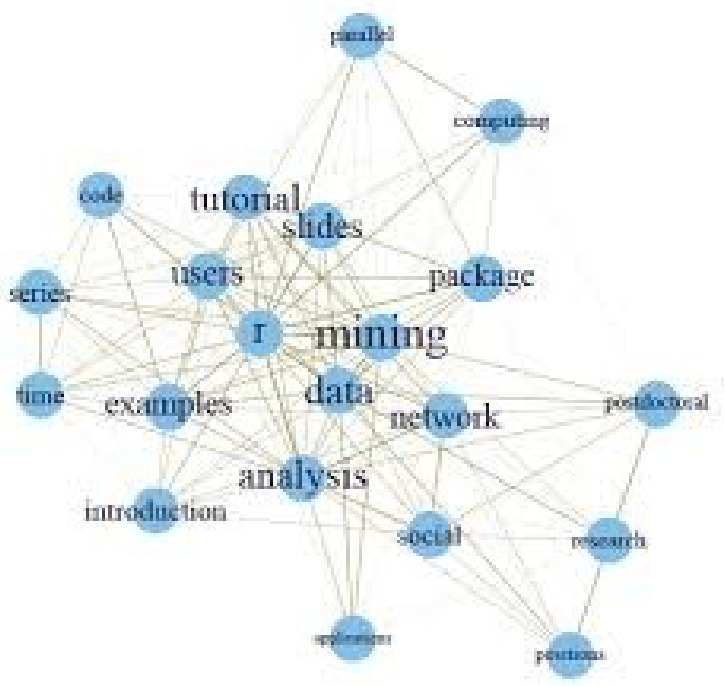
Conviction

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$$

- Dapat diartikan sebagai rasio dari frekuensi yang diharapkan X terjadi tanpa Y (artinya, frekuensi bahwa aturan tersebut membuat prediksi yang salah) jika X dan Y independen dibagi oleh frekuensi yang diamati dari prediksi yang salah.

Co-occurrence network

- Umumnya digunakan untuk menyediakan visualisasi grafis dari hubungan potensial antara orang, organisasi, konsep, organisme biologis seperti bakteri atau entitas lain yang diwakili dalam bahan tertulis.



Konsep Co-occurrence Network

- Interkoneksi kolektif dari istilah berdasarkan kehadiran pasangan mereka dalam unit teks tertentu.
- Jaringan dihasilkan dengan menghubungkan pasangan istilah menggunakan sekumpulan kriteria yang mendefinisikan kejadian bersama.

Contoh Konsep Co-Occurrence

- Istilah A dan B dapat dikatakan "terjadi bersamaan" jika keduanya muncul dalam artikel tertentu.
- Artikel lain mungkin mengandung istilah B dan C.
- Menghubungkan A ke B dan B ke C menciptakan jaringan co-kejadian dari ketiga istilah ini

Sequential pattern mining

- Topik Data mining yang berkaitan dengan menemukan pola yang relevan secara statistik di antara kumpulan data di mana data disampaikan secara berurutan.
- Pengenalan pola yang berurutan.
  - Contoh pola transaksi dalam satu tahun

Asumsi pada Sequential Pattern Mining

- Umumnya diasumsikan sebagai nilai diskrit, berdasarkan satuan waktu tertentu.
  - Meningkat dalam satu hari→peningkatan dengan perbandingan hitungan jam dalam 1 hari
- Penambangan pola berurutan adalah kasus khusus data mining dengan urutan terstruktur.

Repeat-related problems

- Pengenalan pola berdasarkan pengulangan kejadian dalam satuan waktu tertentu
  - Sering terjadi → ditemukan 100 transaksi dalam satu hari
  - Sering terjadi secara berurutan → panas dilanjutkan dengan hujan deras terjadi puluhan kali dalam satu tahun

Alignment problems

- Pengenalan pola berdasarkan urutan kejadian dalam satuan waktu tertentu :
  - Pola perubahan suhu dalam satu hari → walau berbeda-beda tetapi siang lebih panas dari malam.
  - Pola pergantian musim kemarau dan hujan → walau tidak pasti tetapi pasti ada pergantian.