Toistuvien geometristen hahmojen louhinta (alustava tiivistelmä)

Taneli Mielikäinen

Spatiaalisen tiedon louhinta -seminaari
Kevät 2003

Eräs tiedon louhinnan keskeisiä tavoitteita on löytää mielenkiintoisia hahmoja datasta [Han02, Man02]. Eräs hahmon mielenkiintoisuuden mitta on sen tyypillisyys datassa. Eräs hahmon tyypillisyyden mitta on sen frekvenssi datassa. Suurin osa mielenkiintoisien hahmojen louhinnan tutkimuksesta on keskittynyt juuri toistuvien eli korkeafrekvenssisten hahmojen etsintään. Toistuvien hahmojen etsintään keskitytään myös tässä kirjoitelmassa.

Toistuvien hahmojen louhinta voidaan muotoilla laskennalliseksi ongelmaksi seuraavasti:

Syöte: Aineisto $D \in \mathcal{D}$, hahmokokoelma $\mathcal{P}$, frekvenssifunktio

$$fr : \mathcal{P} \to \mathcal{D} \to [0, 1]$$

ja frekvenssiraja $\sigma \in [0, 1]$.

Tuloste: Toistuvien hahmojen joukko

$$\mathcal{P}_\sigma = \{P \in \mathcal{P} : fr(P, D) \geq \sigma\}$$

ja toistuvien hahmojen frekvenssit $fr_\sigma : \mathcal{P}_\sigma \to [0, 1]$.

(Hahmokokoelma $\mathcal{P}$ ja frekvenssifunktio fr esitetään yleensä implisiittisesti (uniformeilla algoritmeilla). Hahmokokoelmat ja frekvenssikuvaukset voivat olla eksponentiaalisesti aineistoa suurempia.) Hahmoa $P \in \mathcal{P}$ kutsutaan toistuvaksi, jos $fr(P, D) \geq \sigma$.

Tunnetuin toistuvien hahmojen etsintä on toistuvien joukkojen etsintä. Olkoon R jokin äärellinen joukko ja merkitään sen potenssijoukkoa $2^R = \{X : X \subseteq R\}$. Toistuvien joukkojen etsinnässä mahdollisia aineistoja D ovat joukon R osajoukkojen (äärelliset) jonot

$$D = D_n = \left(2^R\right)^n = X_1 \dots X_n.$$

(Yleensä osajoukkojen järjestystä jonossa ei kuitenkaan huomioida, jolloin jonon sijasta voitaisiin tarkastella monijoukkoa.) Hahmokokoelmana $\mathcal{P}$ toimii joukon R potenssijoukko 2^R. Joukon $X \in \mathcal{P}$ frekvenssi on

$$fr(X, D) = fr(X, D_n) = \frac{|\{i \in [n] : X \subseteq X_i\}|}{n} \in [0, 1].$$

Suoraviivainen toistuvien joukkojen löytäminen tapahtuu laskemalla kaikkien joukkojen $X \in 2^R$ frekvenssit $fr(X, D)$ ja tulostamalla ne joukot, joiden frevenssi on yli σ.

Vuonna 1994 Agarwal ja Srikant sekä Mannila, Toivonen ja Verkamo esittivät toisistaan riippumatta tasoittain etenevän etsintäalgoritmin, joka tunnetaan nimellä Apriori-algoritmi [AMS+96]:

1. $C_1 := \{\{x\} : x \in R\}$; $i := 1$; $F = \emptyset$
2. while $C_i \neq \emptyset$ or $F = 2^R$
   (a) $F_i :=$ findFrequent(C_i, D)
   (b) $C_{i+1} :=$ generateCandidates(F_i)
   (c) $F := F \cup F_i$; $i := i + 1$

   return$(F, fr(F))$

Algoritmi etenee siis tasoittain tarkistaen edellisen kierroksen toistuvista joukoista laskettujen potentiaalisten toistuvien joukkojen frekvenssit. Aliohjelma FINDFREQUENT voidaan toteuttaa niin, että aineisto D käydään läpi kerran kullakin aliohjelman kutsulla ja kerrallaan aineistosta pitää olla muistissa vain yksi joukko. Tehokasta frekvenssien laskemista tärkeämpää olisi kuitenkin karsia pois ne joukot, jotka eivät voi olla toistuvia. Tämä onnistuu hyvin joukkojen frekvenssien antimonotonisuuden avulla: Joukon frekvenssi ei voi olla suurempi kuin pienin sen osajoukkojen frekvensseistä. Siispä joukko ei voi olla toistuva, jos jokin sen osajoukoista ei ole toistuva.

Tarkemmin katsottuna Apriorin tasoittainen etsintä yleistyy osittainjärjestetyille hahmokokoelmille ja antimonotonisille mielenkiintoisuusmitoille [MT97]. Tästä huolimatta erilaisille hahmoluokille on esitetty lukuisia Apriori-algoritmin muunnelmia eli olennaisesti mukautettu aliohjelmat FINDFREQUENT ja GENERATECANDIDATES eri hahmoluokille.

Näin voidaan menetellä myös geometristen hahmojen – esimerkiksi pisteparvien, tieverkostojen, aikasarjojen tai proteiinirakenteiden – tapauksessa. Geometriset hahmot eroavat monista muista kuitenkin sillä merkittävällä tavalla, että niiden tapauksessa ollaan kiinnostuneita likimääräisistä esiintymistä. Tämä näyttää merkittävästi vaikeuttavan esimerkiksi potentiaalisten hahmojen generointia. Osittaisia ratkaisuja toistuvien geometristen hahmojen louhintaan on esitetty artikkelissa [KK02].

Viitteet

[AMS+96] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen ja A. Inkeri Verkamo. Fast discovery of association rules. 307-328. Teoksessa *Advances in Knowledge Discovery and Data Mining*, toimittanut U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth ja R. Uthurusamy, luku 12, sivut 307 – 328. AAAI/MIT Press, 1996.

[Han02] David J. Hand. Pattern detection and discovery. Teoksessa *Pattern Detection and Discovery*, toimittanut D.J. Hand, N.M. Adams ja R.J. Bolton, osa 2447 sarjasta *Lecture Notes in Artificial Intelligence*, sivut 1 – 12. Springer-Verlag, 2002.
URL `http://link.springer.de/link/service/series/0558/tocs/t2447.htm`

[KK02] Michihiro Kurakochi ja George Karypis. Discovering frequent geometric subgraphs. Teoksessa *Proceedings of the 2002 IEEE International Conference on Data Mining*. IEEE Computer Society, 2002.

[Man02] Heikki Mannila. Local and global methods in data mining: Basic techniques and open problems. Teoksessa *Automata, Languages and Programming*, toimittanut P. Widmayer, F. Triguero, R. Morales, M. Hennessy, S. Eidenbenz ja R. Conejo, osa 2380 sarjasta *Lecture Notes in Computer Science*, sivut 57 – 68. Springer-Verlag, 2002.
URL `http://link.springer.de/link/service/series/0558/tocs/t2380.htm`

[MT97] Heikki Mannila ja Hannu Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241 – 258, 1997.