交通系ICカードデータによる鉄道利用者行動のバスケット分析*

Basket Analysis of Railway Passenger Behavior with Smart Card Transaction Data *

日下部貴彦**・朝倉康夫***
By Takahiko KUSAKABE**・Yasuo ASAKURA***

１．はじめに

近年，鉄道・バスなどの公共交通機関で，交通系ICカードの普及が進んでいる．これらの交通系ICカードは，よりスムーズな料金収受を目的としている一方で，その処理データ（以下ICデータと呼ぶ）には，交通行動分析に応用可能だと考えられる特徴がある．その特徴には，①改札通過時刻を１分単位と詳細に取得できる，②長期にわたる定点観測データが収集できる，③カード毎のIDを識別することで利用者を識別した上での分析ができることなどが上げられる．①の特徴は，時間解像度の高い分析が可能となることを示唆しており，②は長期間の変動に着目した分析が可能であることを示している．③の特徴により，各カードに割り当てられたID毎の行動変化を考慮した分析や，ID毎の利用駅や利用時間等の組み合わせを考慮した分析が可能となると期待できる．

ICデータの特徴を活かした発見的なアプローチよるこれまでの研究では，日下部ら[1]は可視化技術を応用して，長期間の旅客流動の変動に着目した分析を行っている．Agardら[2]は，データマイニングの一手法であるクラスター分析を行うことで，バス利用者の利用パターンを分析している．Morencyら[3]は，バス利用者毎の利用したバス停数に着目し，累積図を用いて利用者の行動の空間的な広がりを分析し，考察している．

交通行動データとしてICデータを捉えた場合には，他の交通機関などの鉄道会社以外での利用者の行動が取得できないことや，交通目的や目的地を取得できないという短所もある．しかし，はじめに述べたような長所があるとともに，料金収受に伴って自動的に取得可能であり，観測のための新たなコストがほとんど発生しないことから，継続的な観測が可能であるという長所もある．したがって，それらの長所を活かせる分析手法を構築することが求められており，データマイニングなどの発見的な手法を用いたデータ解析による知見の積み重ねが必要であるといえる．

本研究では、各利用者が利用した駅の組み合わせに着目した分析を行う．ICデータに記録されている空間的な情報は，各利用者が乗降時に利用した駅であり，これらを分析することで，Within-dayの鉄道を利用したトリップチェインや，各利用者の利用駅の空間的な広がりについての知見が得られると期待できるからである．こうした組み合わせを考慮した発見的な分析を行う上で，最も単純な方法は，すべての駅の組み合わせについて，利用された回数を集計することであろう．しかし，単純に組み合わせを数え上げることは，数億トリップに及ぶ大量のデータに対しては，困難であることが多い．例えば，50駅を対象として起終点の組み合わせを分析する場合には，2500個の組み合わせを考慮する必要である．さらに利用日や時間帯など組み合わせを考慮する場合には，膨大な量の組み合わせを考慮しなければならないことになる．このような組み合わせを効率的に数え上げ，特徴的な組み合わせのみを抽出する手法として，データマイニングの一分析手法であるマーケットバスケット分析（以下バスケット分析と呼ぶ）がある[4]．これまでの交通分野での研究でも，観測された行動の組み合わせを分析するためにバスケット分析が応用されている．出水ら[5]はプローブパーソン調査によるデータを用いて，特定の施設滞在後にどういう移動ー活動をしたのかを分析し，交通行動の行動文脈の解析への適用可能性を示している．Yamashitaら[6]による研究では，アンケート調査のデータをもととして，都市鉄道と都市間鉄道の結節駅での駅内での滞在施設の利用に関する特徴を分析している．

本研究では，鉄道利用者が一日の間に利用する駅の組み合わせについての知見を得ることを目的とし，バスケット分析を応用する．第二章では，データの仕様及び分析対象のデータについて述べる．第三章では，鉄道利用者の利用駅の組み合わせに関する基礎的分析を行う．第四章では，鉄道利用者が一日の間に利用する駅の組み合わせに着目したバスケット分析を行い，五章でまとめについて述べる．

*キーワーズ：交通系ICカードデータ、マーケットバスケット分析、交通行動、旅客流動
**学生員，工修，神戸大学大学院工学研究科・日本学術振興会特別研究員DC
（神戸市灘区六甲台町１－１，
TEL078-803-6360，FAX078-803-6360）
***正会員，工博，神戸大学大学院工学研究科

２．ICデータの概要

本研究で使用するICデータの形式を図１に示す．このデータ形式では，レコード毎に一回の鉄道利用が記録されている．各レコードには，「使用年月日(営業日)，入場駅，入場時刻，出場駅，出場時刻，カードID」が記録されている．

本研究では，都市部に路線をもつ鉄道会社 A で観測された IC データを用いる．対象路線の駅数は，41 駅である．対象の期間は 2008 年 4 月 1 日～ 5 月 31 日であり，この間の平日は 41 日，休日（土曜・日曜祝日）は 20 日である．この間に観測されたトリップ数は，3,505,436 トリップである．なお，本研究では，鉄道利用１回を１トリップとして数えることとする．したがって，本来の同一トリップで２度，対象路線を利用した場合には２トリップと数えることとなる．

ICデータ

使用年月日	入場駅	入場時刻	出場駅	出場時刻	カードID
2007年10月13日	A駅	7:10	C駅	7:23	A25687DK
2007年10月13日	A駅	7:11	C駅	7:23	B68677DS
2007年10月13日	A駅	7:11	B駅	7:18	B67732RR
⋮					
2007年10月13日	B駅	17:57	A駅	18:09	B67732RR
2007年10月13日	C駅	18:00	A駅	18:17	B89751RR
⋮					

図１　ICデータのデータ形式の例

３．組み合わせに着目した基礎分析

本章では，バスケット分析を行うための予備的な知見を得ることを目的として，各鉄道利用者の利用駅の広がりに着目した分析を行う．バスケット分析のようなデータマイニングの手法による発見的な分析では，予見可能な行動パターンばかりが抽出されることがあり得る．こうした予見可能な行動パターンの抽出を防ぐためには，事前に予見される行動パターンについて分析を行い，マイニング時には予見可能なパターンが抽出されないような工夫を行う必要がある．したがって，本章の分析では，予見可能な行動パターンとして規範パターンを設定し，規範パターンと観測値のずれに着目して分析を行う．

（1）規範パターン

都市内の鉄道利用者の最も単純な一日の行動パターンを考えたとき，自宅から出発し，自宅近くの乗車駅，目的地近くの降車駅，目的地の順に移動し，目的地で何らかのアクティビティを行った後に，目的地へ向かった時の経路を逆順にたどって自宅へと移動する行動パターンが考えられる．本研究では，この行動パターンを規範パターンと呼ぶ．また，規範パターンでは，日々の利用で一日の初めの乗車駅は変化しないものとする．規範パターンに従う利用者は，一日に二回のトリップを行い，一日の初めの乗車駅と最後の降車駅は一致する．

（2）分析方法

鉄道利用者の行動は，利用者の職業，年齢，居住地などの個人属性によって異なると考えられる．しかし，IC データでは，一般にこれらの項目が観測されているとは限らない．したがって，IC データで観測されている項目から，利用者の属性を考慮する必要がある．本研究では，各鉄道利用者の利用頻度を利用者の属性として考える．鉄道の利用頻度は，利用者の属性が反映された結果，行動として顕示され，観測されたものであると捉えることができるためである．そこで，期間中の利用者の鉄道を利用した日数についてのヒストグラムを分析する．

実際の鉄道利用者の行動では，複数のアクティビティがある利用者や，一日の初めの乗車駅と一日の終わりの降車駅が一致しない利用者，一日の初めの乗車駅が日によって異なる利用者など規範パターンと一致しない利用者も数多く観測されているはずである．また，規範パターンを行っている利用者であっても，日々の目的地が異なる場合もあり，目的地に空間的な広がりをもっていることも想定される．そこで，分析では，以下の四つの点に着目した集計分析を行う．

a）利用駅数

利用駅数の分析では，目的地に空間的な広がりをみることを目的としている．観測期間内に各利用者が利用した降車駅数を集計し，その集計値について各利用頻度帯の利用者毎にヒストグラムを作成する．

b）往復率

一日の初めの乗車駅と最後の降車駅の一致率を分析する．利用者毎に一日の初めの乗車駅と最後の降車駅が一致している日数を集計し，利用日数で除したものを往復率とする．往復率について各利用頻度帯の利用者毎にヒストグラムを作成する．

c)ベース駅利用率

一日の初めの乗車駅が日によって異なるかを分析する．利用者毎に一日の初めの乗車駅として最も多く観測された駅をベース駅とする．ベース駅を一日の初めの乗車駅として利用している日数を利用者毎に集計し，各利用頻度帯の利用者毎にヒストグラムを作成する．

d)複数回のトリップ

各利用者の一日あたりの平均トリップ回数を利用頻度階級毎にヒストグラムを作成し，トリップ回数の分布を分析する．

（3）分析結果

図2は，ICカード利用者が分析期間中に鉄道を利用した日数をヒストグラムに表したものである．横軸が利用日数であり，縦軸が横軸の利用日数に対応する鉄道利用を行った利用者数の全利用者に対する構成比である．また，折れ線は対応する利用日数以下の累積人数の構成比を示している．この図より，利用日数が1日もしくは2日の利用者が62%を占めていることがわかる．9割の利用者が20日以下の利用日数であり，大多数の利用者が低頻度の利用を行っていることがわかる．また，平日の日数と同じ41日付近になだらかな山があることも確認できる．

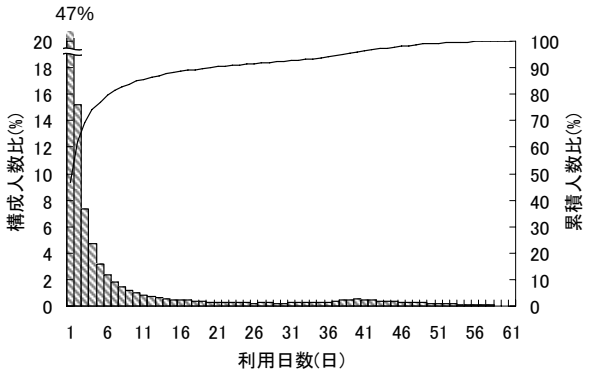

図2　利用日数のヒストグラム

a）利用駅数

図3は，観測期間内に各利用者が利用した降車駅数を集計し，各利用日数の利用者毎に示したものである．横軸が利用日数，縦軸は各利用者が利用した降車駅数を示している．色は，対応する降車駅数の利用者数を各利用日数の階級内での構成比を示している．濃色の部分がわずかに右肩上がりであり，利用日数が増えると利用する駅数も増加する傾向にあることがわかる．平均値でみると，利用日数が 1 日の利用者では，1.55 駅であったのに対し，41 日の利用者では，4.43 駅であった．

b）往復率

図4は，往復率について a)と同様の方法で図に示したものである．利用日数が約 30 日以上の利用者に着目すると，一日の初めの乗車駅と最後の降車駅が一致する鉄道利用を 90%以上の利用日で行っている利用者が多くいることがわかる．利用回数が 41 日の利用者では，43%の利用者が 90%以上の往復率を示している．一方で，利用日数が少なくなるにつれて，往復率が高い利用者が少なくなる傾向にあることがわかる．この結果は，高頻度の利用者と低頻度の利用者では，鉄道利用の仕方が異なることを示唆している．また，高頻度の利用者では，往復率の観点では規範パターンに近い利用を行っていることがわかる．

c)ベース駅利用率

図5はベース駅利用率を b)までと同様の方法で示したものである．往復率と同様に利用日数が多い利用者ほどベース駅を一日の初めの鉄道利用での乗車駅として選択している傾向があり，規範パターンに近い行動をしている利用者が多いことがわかる．利用回数が 41 日の利用者では，90%以上のベース駅利用率の利用者が 76%となっている．

d)複数回のトリップ

図6は一日の平均トリップ回数に関して a)～c)と同様の方法で示したものである．トリップ回数においても，利用頻度が高い利用者ほど，一日あたり2回に近いトリップを行う利用者が多い．利用回数が 41 日の利用者では，90%の利用者が，1.5 トリップ以上 2.5 トリップ未満の平均トリップ回数をもっている．高頻度の利用者ほど規範パターンに近い行動を行う傾向があることがわかる．

4．バスケット分析

本章では，鉄道利用者の Within-day での複数の目的地への移動に着目する．利用される駅の組み合わせについての規則性についての知見を得るために，バスケット分析を用いた発見的なアプローチによる分析をおこなう．

（1）バスケット分析

バスケット分析は，同じバスケットの中に入っているアイテムの組み合わせ（アイテムセット）について規則性を見つけるためのデータマイニングの手法である．例えば，鉄道の駅をアイテムとし，ある利用者の一日の行動をバスケットと見立てた場合には，一人の利用者の一日の行動の内に利用される駅の組み合わせに関するルールを見つけることとなる．つまり，「A 駅を利用する利用者は B 駅も利用する傾向にある」というようなルールをデータから抽出するための手法である．バスケット分析では，有効なルールが抽出されたかを判断する指標として，サポート値，確信度，リフト値が用いられる．アイテム A と B の組み合わせに対するサポート値は，

$$P(A \cap B) = \frac{N(A \cap B)}{N} \qquad (1)$$

ただし，

$N(S)$ は組み合わせ S の観測回数．

N は全観測数

で表される．サポート値は，観測全体のうちにその組み合わせが占める割合を示すものである．

確信度は，アイテム A が観測されている場合に，アイテム B が観測される割合を示している． 確信度は

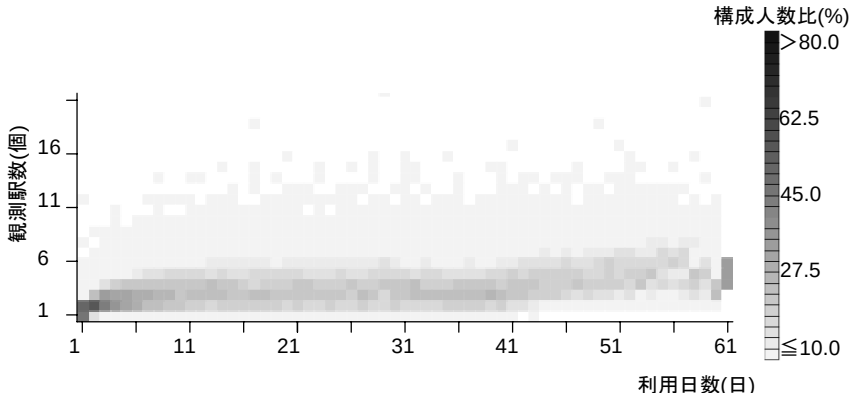

図３　利用日数の階級別利用駅数

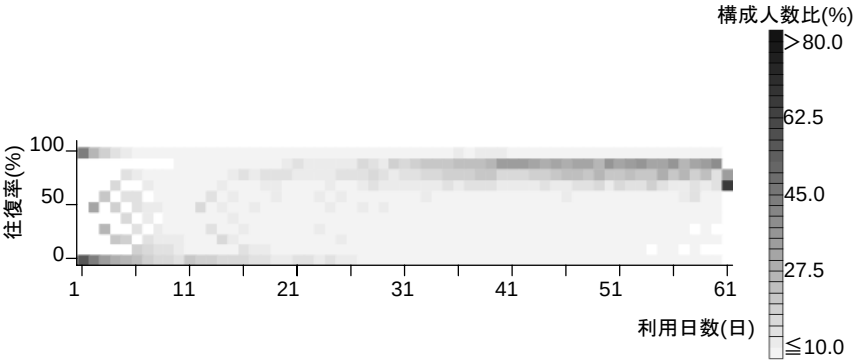

図４　利用日数の階級別往復率

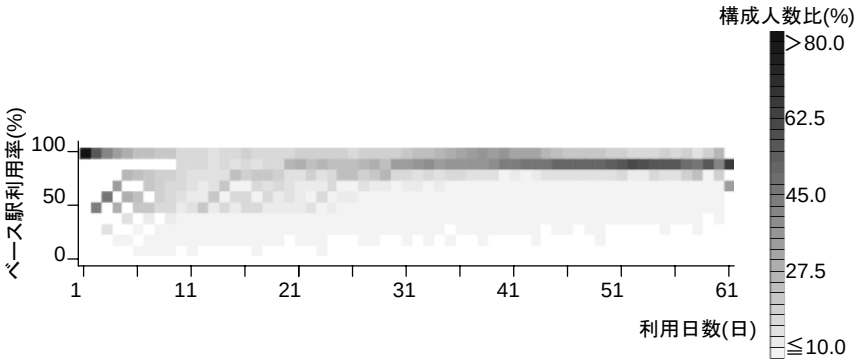

図５　利用日数の階級別ベース駅出発率

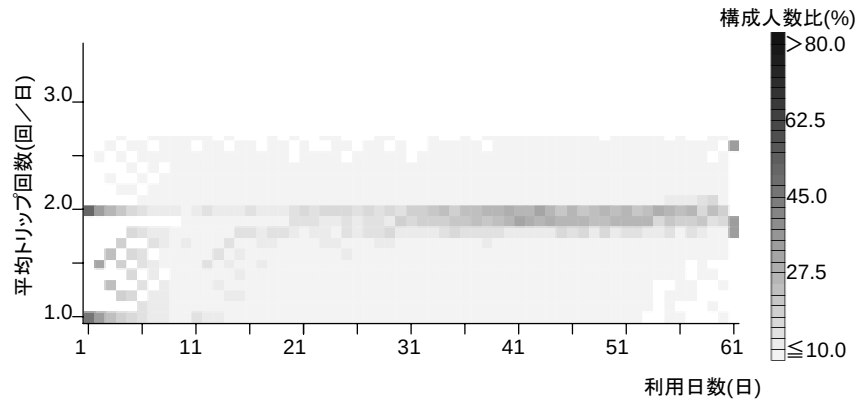

図６　利用日数の階級別平均トリップ回数

$$P(A \cap B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad (2)$$

と表すことができる．また，この関係を「$A \rightarrow B$」と表す．

リフト値は，アイテムBが観測される割合が，アイテムAが観測されている場合と，アイテムAの観測が確認されていない場合でどちらが大きいかを比較するための指標である．これは，

$$L(A \cap B \mid A) = \frac{P(A \cap B \mid A)}{P(B)} \qquad (3)$$

と表される．リフト値が 1 以上であれば，アイテムAが観測されれば$A \cap B$が観測されるという関係が，アイテムAが観測されていない場合よりも強いことを示している．

以上に示したような指標値をすべてのアイテムの組み合わせについて算出することは一般に計算上の困難になりやすいと言える．すべての組み合わせを考慮して，数え上げた場合には，アイテム数 n に応じて，2^n の組み合わせを考慮しなければならないからである．しかし，有効なアイテムセットは，サポート値やリフト値といった指標が一定の基準を満たしていることが条件となることから，少ないアイテム数のアイテムセット(通常は 1 アイテム)から数え上げた上で，各指標が一定の規準を満たしたアイテム同士の組み合わせのみを数えあげることで，計算時間を節約する手法がとられる．本研究では，Brin ら[7]による，Dynamic Itemset Counting(DIC)を用いて計算を行う．

（2）分析方法

Within-day での鉄道を利用した行動による行動範囲の規則性に着目するために，利用駅の組み合わせを分析する．分析では，各降車駅をアイテムとして設定し，バスケットは，各利用者の各一日の行動とする．

３章の基礎分析では，規範パターンを行う利用者が一定数いることが示唆された．バスケット分析の結果に帰宅時の駅選択による規則性が含まれてしまうことを防ぐために，３章で定義したベース駅をバスケットから取り除いたものをアイテムセットとした．各利用者のベース駅以外の駅への降車が複数回ある利用日が分析対象となる．

DIC による分析では，アイテムセットの生成の際の基準となるサポート値，確信度，リフト値を設定する必要がある．本分析では，それぞれ 0.0001， 0.2， 1.0 と設定した．分析の対象日は，平日と休日に分けて分析を行う．

表－１　バスケット分析の結果(平日)

順位	アイテムAの駅番号	アイテムBの駅番号	サポート値	確信度A→B	確信度B→A	リフト値
1	2	8	0.0212	0.1208	0.2598	1.4820
2	3	10	0.0190	0.1059	0.2536	1.4152
3	5	12	0.0081	0.0522	0.2361	1.5294
4	3	24	0.0071	0.0398	0.2713	1.5145
5	3	26	0.0067	0.0374	0.3894	2.1736
6	1	25	0.0059	0.0142	0.4770	1.1575
7	3	29	0.0054	0.0301	0.2413	1.3471
8	3	27	0.0052	0.0288	0.3237	1.8068
9	3	35	0.0037	0.0208	0.2923	1.6315
10	1	32	0.0036	0.0087	0.4404	1.0687

表－２　バスケット分析の結果(休日)

順位	アイテムAの駅番号	アイテムBの駅番号	サポート値	確信度A→B	確信度B→A	リフト値
1	2	8	0.0278	0.1247	0.2983	1.3373
2	3	10	0.0254	0.1308	0.2795	1.4425
3	1	9	0.0229	0.0574	0.4037	1.0142
4	1	13	0.0102	0.0257	0.4556	1.1446
5	5	12	0.0086	0.0617	0.2209	1.5854
6	3	11	0.0071	0.0364	0.2044	1.0550
7	1	19	0.0070	0.0177	0.4812	1.2088
8	1	25	0.0059	0.0149	0.4483	1.1263
9	3	24	0.0057	0.0293	0.2755	1.4216
10	1	32	0.0057	0.0142	0.5502	1.3821

（3）分析結果

平日の分析では，171,664 個のバスケットが分析対象となり，バスケット分析によってアイテム数が 2 個のアイテムセットが 39 個抽出された．休日では，56,210 個のバスケットが対象となり，アイテム数が 2 個のアイテムセットが 18 個抽出された．表－1，表－2は，それぞれで抽出されたアイテムセットについて，サポート値の順に上位 10 個のアイテムセットを示したものである．なお，表中の駅番号は，分析期間中の各駅の降車人数の順位が高いものから順に割り当てたものである．駅番号 1，3，8，11 は他社線との接続駅となっている．

表によると，サポート値が上位２位までのアイテムセットは，平日，休日で同じものが抽出されている．これらの組み合わせは，降車人数が比較的上位にある駅の組み合わせとなっている．３位～１０位のアイテムセットでは，平日と休日で大きく順位が異なることが確認できる．３位以下のアイテムセットでは，降車人数が比較的上位の駅と下位の駅との組み合わせが多く見られる．また，このような組み合わせでは，「下位の駅→上位の駅」の確信度と「上位の駅→下位の駅」の確信度との差が上位２位までと比べて大きいことがわかる．このことは，上位の駅を利用した人が下位の駅を利用することは多くないが，下位の駅を利用した人が上位の駅も利用することは多いと言う事を示している．

５．おわりに

本研究では，各利用者が Within-day の行動で利用した駅の組み合わせに着目してバスケット分析を行った．

３章の基礎分析では，利用頻度による鉄道利用者の特性の違いを分析した．分析では，鉄道利用者の行動の規範パターンを定義し，このパターンとのずれに着目して考察を行った．その結果，利用頻度が高い利用者ほど規範パターンに近いパターンでの行動が多く見られることが示唆された．

４章のバスケット分析では，Within-day での複数回のトリップを行う利用者の降車駅に着目した分析を行った．抽出された駅の組み合わせについて，平日と休日で利用のされ方が変わるものと変わらないものがあるなど，特徴がある組み合わせが確認された．しかし，それらの要因を明らかにするためには，より詳細な分析が必要である．その際には，アンケートなど追加的なデータの取得も視野に入れる必要がある．よって，要因に関する分析は今後の課題としたい．

参考文献

1）日下部貴彦，中島良樹，朝倉康夫: 可視化技術をもちいた交通系ICカードデータの分析，土木計画学研究・講演集，CD-ROM，Vol.39，徳島，2009

2）Agard, B., Morency, C. and Trépanier, M.: Mining Public Transport User Behaviour from Smart Card Data, 12th IFAC Symposium on Information Control Problems in Manufacturing –INCOM 2006, Saint-Etienne, France, 2006.

3）Morency, C., Trépanier, M. and Agard, B.: Measuring Transit Use Variability with Smart-card Data, Transport Policy, Vol.14 (3), pp. 193–203, 2007

4）Agrawal, R., Imielinski, T. and Swami, A.: Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, 1993

5）出水浩介 and 羽藤英二: プローブパーソンデータを用いた移動 - 活動パターンのバスケット分析，土木計画学研究・講演集，CD-ROM，Vol.30，2007

6）Yamashita, Y., Hibino, N. and Uchiyama, H.: A Behavioral Analysis of Passengers' Railway Station Facilities Visiting Characteristics, Journal of the Eastern Asia Society for Transportation Studies, Vol.7, pp.808-816, 2007

7）Brin, S., Motwani, R., Ullman, J.D. and Tsur, S.: Dynamic Itemset Counting and Implication Rules for Market Basket Data, SIGMOD Record, Vol. 6 (2), pp. 255-264, 1997