Eksploracja zasobów internetowych

Wykład 7

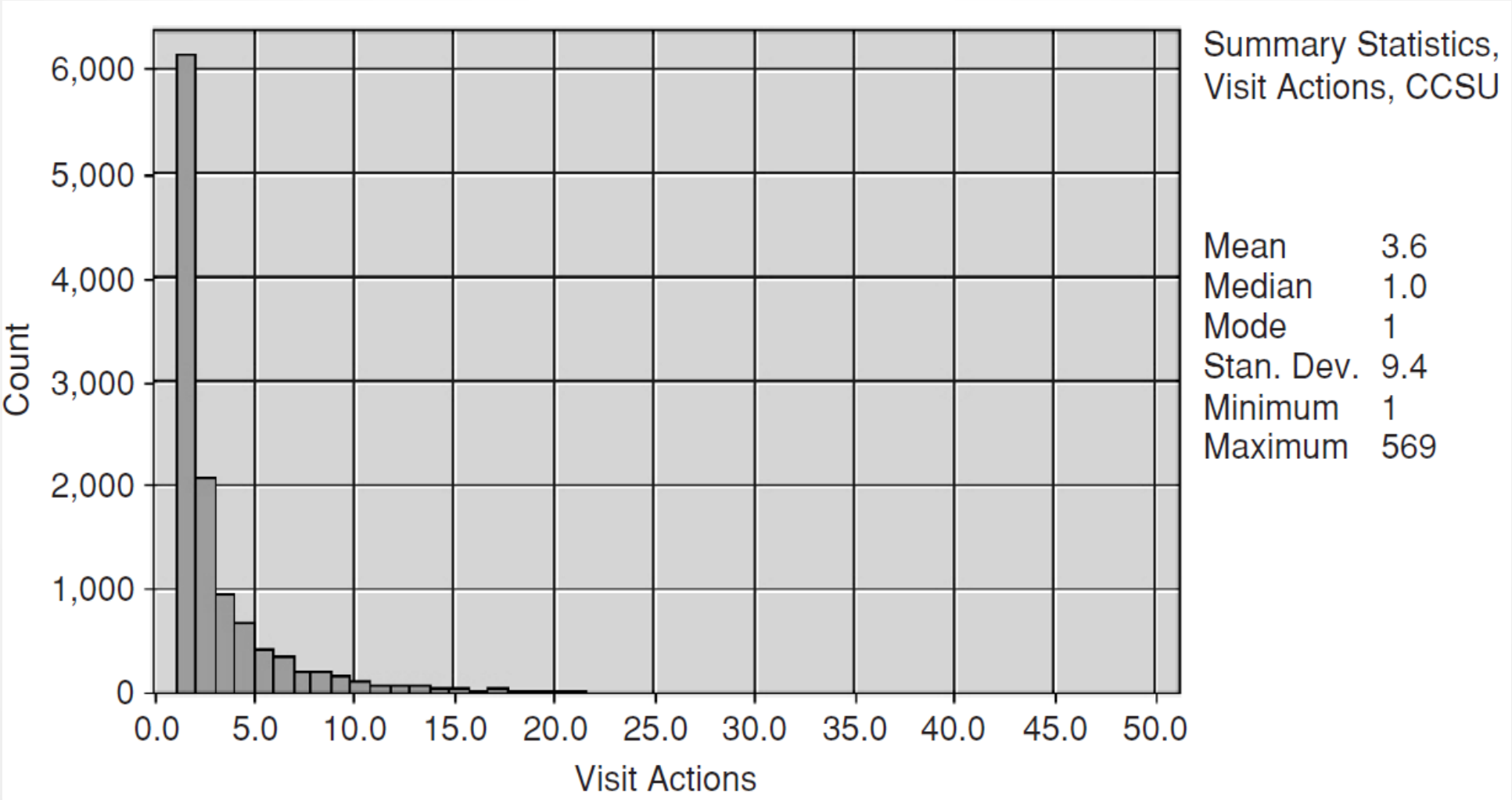
Eksploracja wykorzystania sieci WWW – zastosowanie analizy statystycznej i metod eksploracji danych

Piotr Hońko
Wydział Informatyki
Politechnika Białostocka

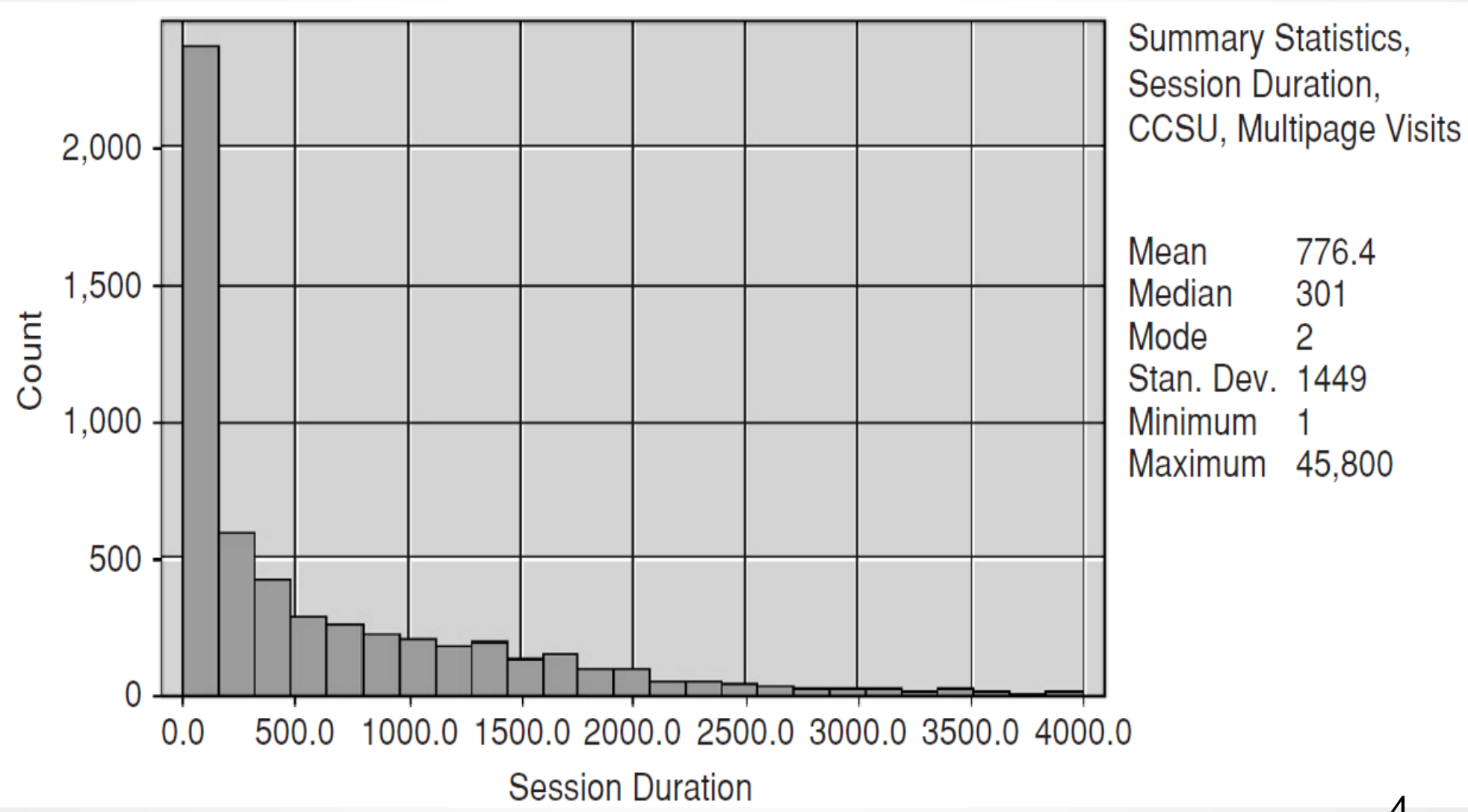
Metody eksploracji wykorzystania sieci WWW – Standardowa analiza

- W celu określenia charakterystyki wizyt stron danego serwisu mogą zostać wykorzystane narzędzia statystyczne.
- Za pomocą takich narzędzi można określić m. in:
  - Statystykę dotyczącą czas trwania sesji;
  - Statystykę dotyczącą liczby akcji użytkownika (tj. liczba zgłoszonych żądań).
  - Korelację pomiędzy liczbą akcji użytkownika a czasem trwania sesji.

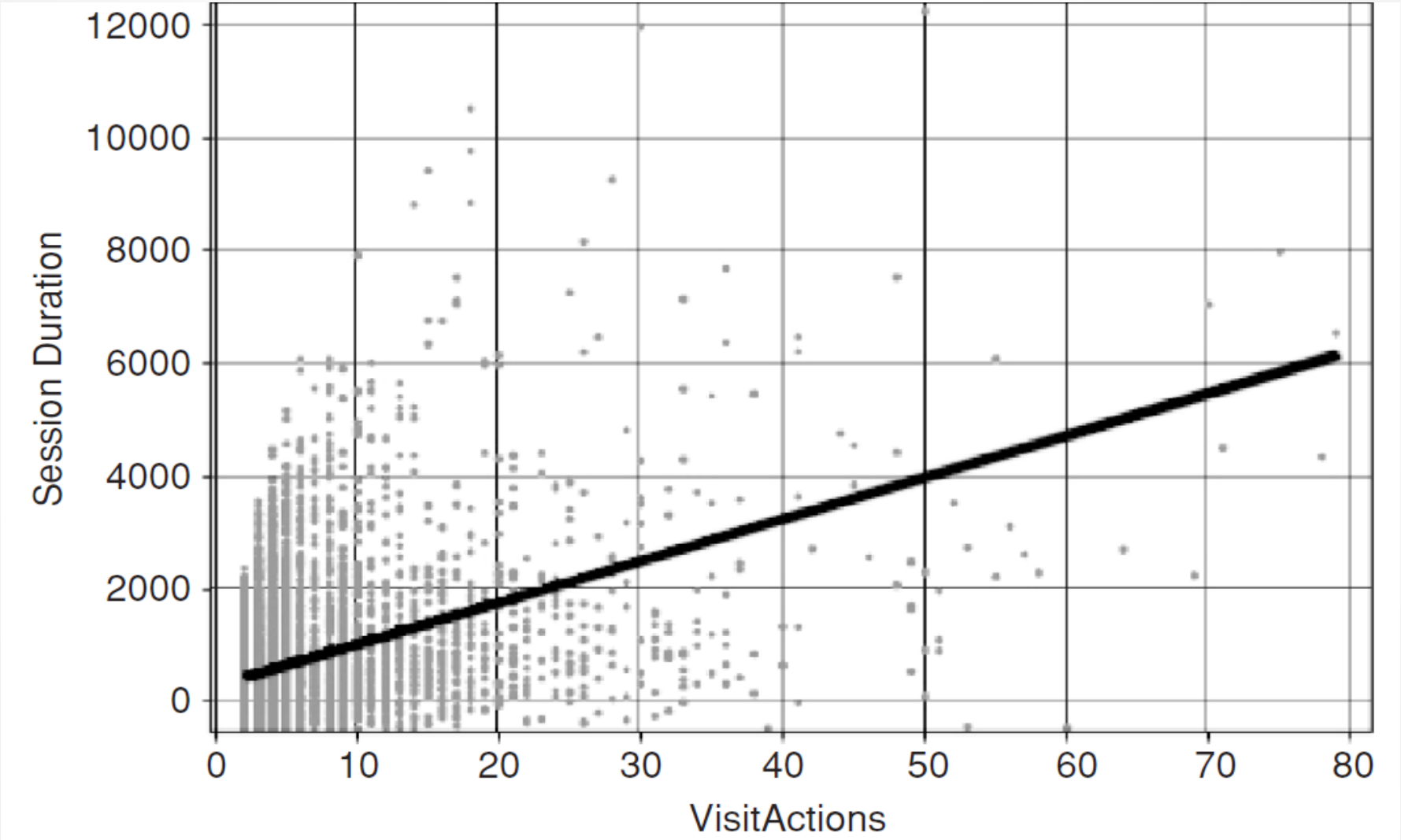
Rozkład dotyczący akcji użytkowników – przykład



Rozkład dotyczący trwania sesji – przykład



Korelacja pomiędzy liczbą akcji użytkownika a czasem trwania sesji – przykład



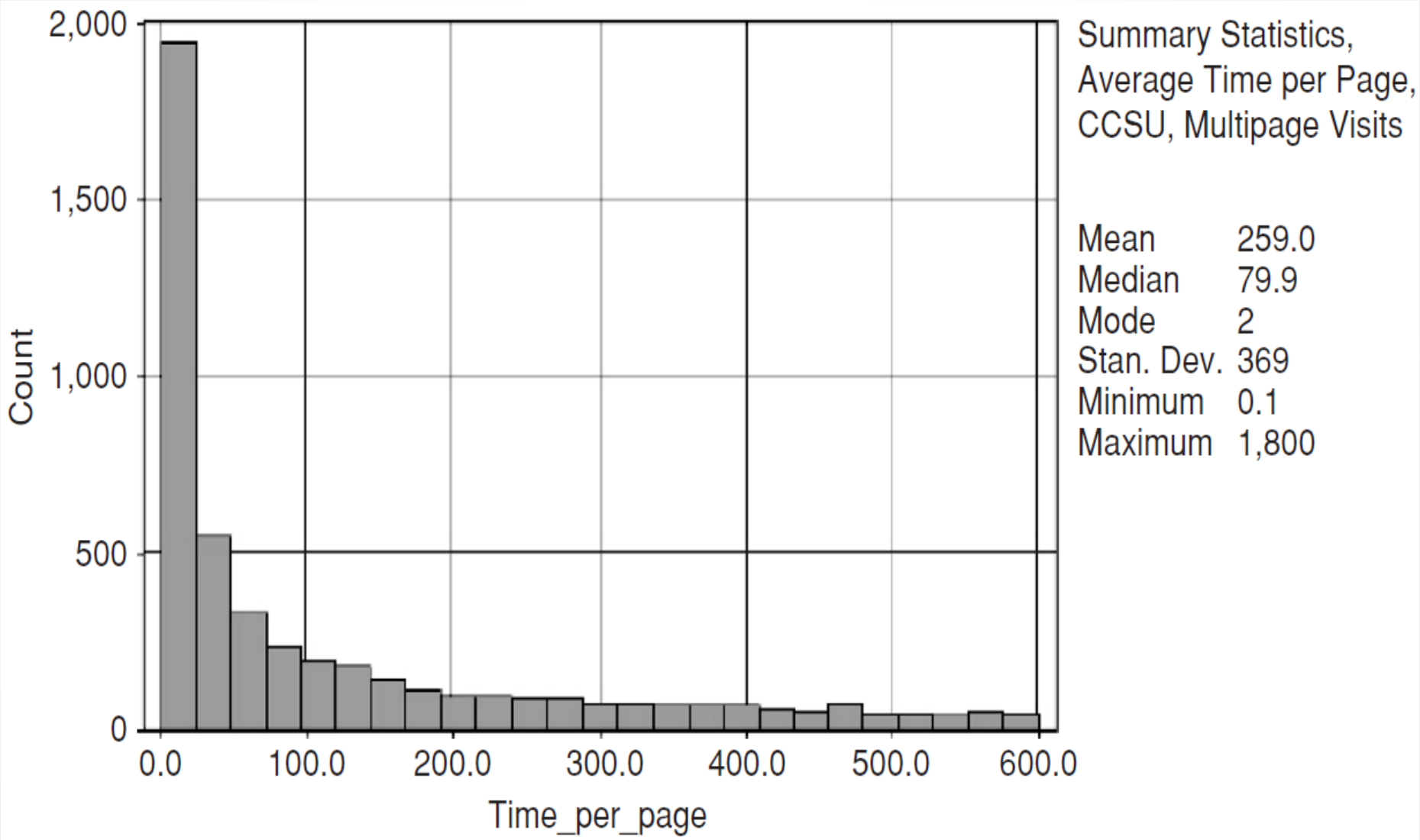
Szacowanie długości trwania sesji na podstawie liczby akcji użytkownika – przykład

- Na podstawie analizy regresyjnej parametry prostej określonej równaniem $y = ax + b$ przyjmują wartości $a = 73.7$, $b = 310.5$.
- Przykładowo szacowany czas trwania sesji przy 10 akcjach użytkownika wynosi

$$(73.7 \cdot 10 + 310.5)\,s = 1047.5\,s = 17.5\,min$$

- Dla każdej dodatkowej akcji szacowane wydłużenie trwania sesji wynosi 73.7 s.

Czas przeglądania strony – przykład



Czy wszystkie strony przeglądane w ramach sesji są tak samo istotne?

- Większość sesji użytkowników kończy się po przejrzeniu kilku stron.
- Statystyka dotycząca pierwszych stron może być istotniejsza od statystyki pozostałych stron.
- Kluczowym zadaniem jest zdolność określania profilu użytkowników na podstawie pierwszy odwiedzanych stron.

Charakterystyka dwóch pierwszych stron odwiedzanych w sesji – przykład

Page 1 Duration			Page 2 Duration		
Overall	Navigation	Content	Overall	Navigation	Content
277.5	239.7	278.3	241.8	178.4	243.8

Metody eksploracji wykorzystania sieci WWW – Grupowanie danych

- Grupowanie danych jest często wykonywane jako wstępny krok do dalszej eksploracji danych.
- Otrzymane grupy stanowią dane wejściowe dla właściwych metod stosowanych do analizy/eksploracji danych.
- Dzięki zastosowaniu grupowania może zostać zredukowana przestrzeń poszukiwań.

Grupowanie danych dotyczących sposobów wykorzystania sieci WWW

- Na podstawie danych dotyczących sposobów wykorzystania sieci WWW tworzone są grupy.
- Każda z grup określa jeden ze sposobów wykorzystania sieci przez użytkowników.
- Liczba sposobów wykorzystania sieci nie jest znana z góry, zatem stosowany jest algorytm, który nie wymaga określenia na starcie liczby grup lub liczba ta jest ustalana eksperymentalnie.

Grupy określające sposoby wykorzystania sieci przez użytkowników – przykład

Variable	Cluster 1	Cluster 2
Page_/Index/	0.6949	0.0073
Page_/Index/C.htm	0.1908	0.0042
Page_/Index/Default.htm	0.6961	0.0022
Page_/search/	0.3663	0.0006
Page_/search/Default.htm	0.3757	0.0006
First_dir/search/	0.3934	0.0053
First_dir/Index/	0.7256	0.0273
Session Actions	13.8	5.0
Session Duration	983	741
Average Time per Page	79	290
Page 1 duration	119	290
Page 2 duration	65	146

Grupy określające sposoby wykorzystania sieci przez użytkowników – przykład c.d.

- Podział danych dotyczących sesji na grupy.

Cluster	1	2
count	937 (16%)	4857 (84%)

Analiza grup określających sposoby wykorzystania sieci przez użytkowników – przykład

- Strony katalogów *index* i *search* są wielokrotnie częściej odwiedzane przez użytkowników z grupy 1.
- Przeciętna liczba akcji w trakcie jeden sesji jest ponad dwukrotnie wyższa w grupie 1.
- Przeciętny czas trwania sesji jest wyraźnie wyższy w grupie 1.
- Przeciętny czas przeglądania strony jest kilkukrotnie wyższy w grupie 2.
- Przeciętny czas przeglądania stron 1 i 2 jest ponad dwukrotnie dłuższy w grupie 2.

Metody eksploracji wykorzystania sieci WWW – klasyfikacja

- Zadanie klasyfikacji może zostać wykonane w przypadku wykorzystania sieci WWW w celu identyfikacji istotnych cech (np. krótki czas trwania sesji) dotyczących sposobów korzystania z danego serwisu.

Metody eksploracji wykorzystania sieci WWW – klasyfikacja c.d.

- Klasyfikacja może również być wykonywana w czasie rzeczywistym:
  - Określenie grupy (tj. profilu) użytkownika na podstawie sposobu wykorzystania przez niego sieci (np. serwis o charakterze handlowym).
  - Dopasowywanie treści stron do profilu użytkownika (np. sugerowanie produktów do zakupu).

Transformacja danych sieciowych dla celów eksploracji

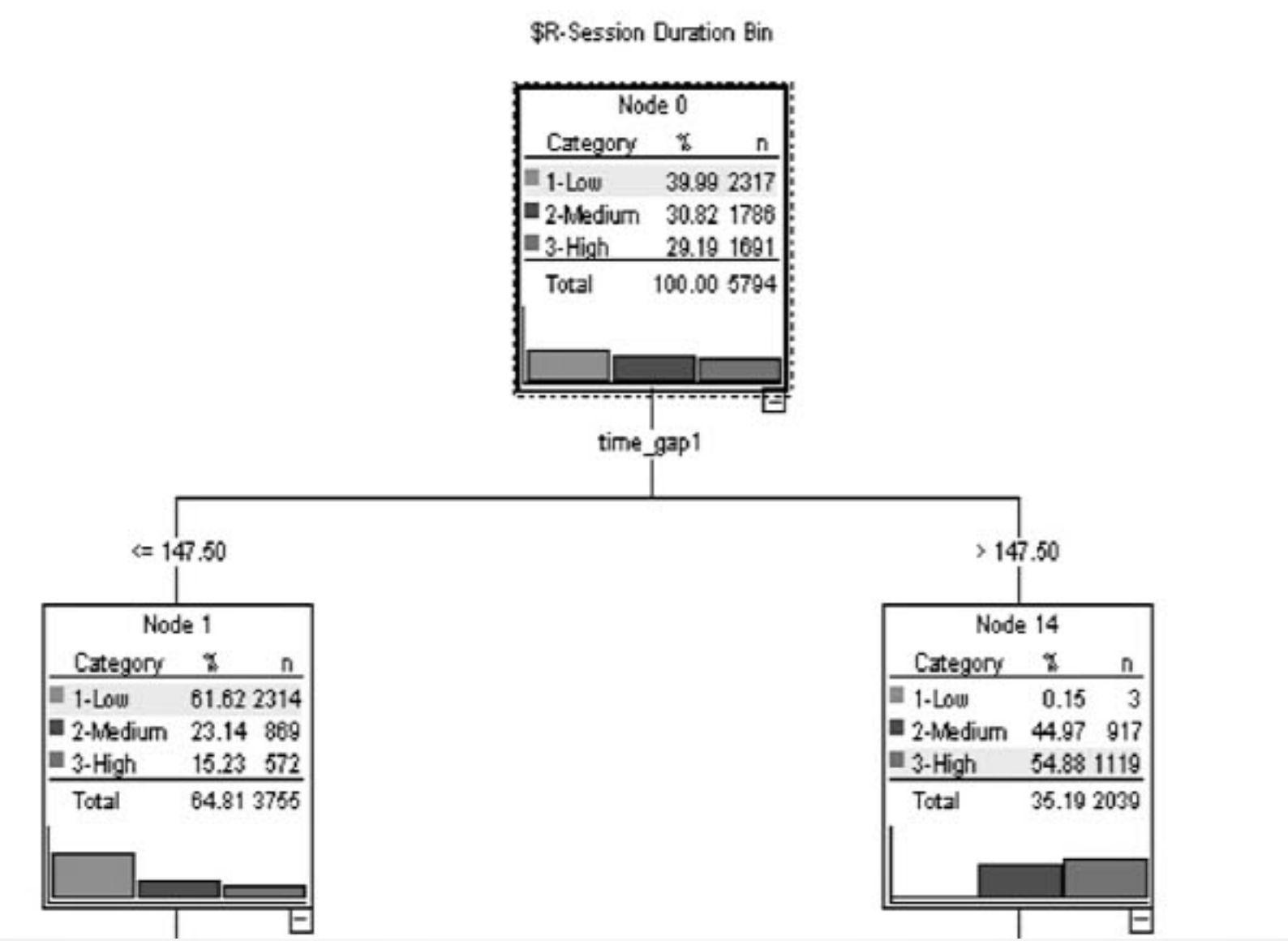
- Atrybuty numeryczne poddawane są dyskretyzacji.

$$\text{Session Actions Bin} = \begin{cases} \text{Low} & \text{if Session Actions} \leq 2 \\ \text{Medium} & \text{if } 3 \leq \text{Session Actions} < 7 \\ \text{High} & \text{if Session Actions} \geq 7 \end{cases}$$

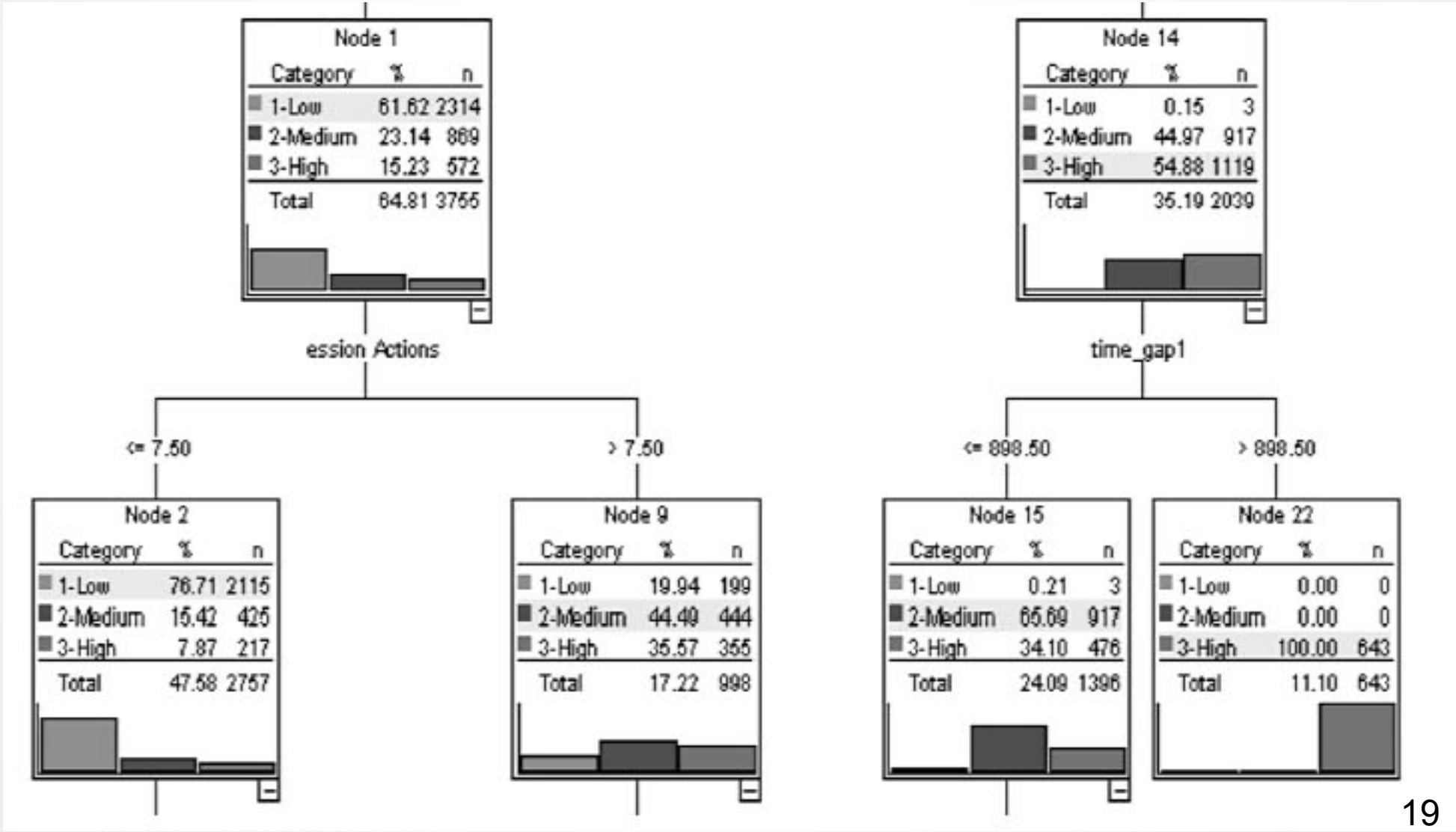
$$\text{Session Duration Bin} = \begin{cases} \text{Low} & \text{if Session Duration} < 150 \\ \text{Medium} & \text{if } 150 \leq \text{Session Duration} \leq 900 \\ \text{High} & \text{if Session Duration} > 900 \end{cases}$$

$$\text{Time per Page Bin} = \begin{cases} \text{Low} & \text{if Time per Page} < 68 \\ \text{Medium} & \text{if } 68 \leq \text{Time per Page} < 504 \\ \text{High} & \text{if Time per Page} \geq 504 \end{cases}$$

Drzewo klasyfikacyjne – przykład



Drzewo klasyfikacyjne – przykład c.d.



Reguły klasyfikacyjne utworzone na podstawie drzewa klasyfikacyjnego – przykład

- Na podstawie węzłów 0, 14 i 22 otrzymujemy następującą regułę:

$$Time\_gap1 > 157.50 \wedge Time\_gap1 > 898.50 \Rightarrow Session\ Duration = High$$

$$Time\_gap1 > 898.50 \Rightarrow Session\ Duration = High$$

Odkrywanie asocjacji

- Szukanie istotnych zależności wstępujących pomiędzy danymi rozpatrywanej bazy danych.
- Odkrywanie modelu skojarzeniowego przedstawiającego współwystępowanie rożnych elementów w danym przypadku.

Odkrywanie asocjacji c.d.

- Odkrywanie asocjacji jest zadaniem deskryptywny składającym się zwykle z dwóch kroków:

1. Wyznaczenie zbiorów (wzorców) częstych;
2. Konstrukcja reguł asocjacyjnych na podstawie wzorców częstych.

Reprezentacja asocjacji: Zbiory częste

- Zbiór elementów (ang. *itemset*)
  - Kolekcja dowolnej liczby elementów (ang. *items*).
- Wsparcie (częstotliwość) zbioru elementów
  - Częstotliwość występowania zbioru elementów, mierzona jako stosunek liczby wystąpień danego zbioru elementów do liczby wszystkich przypadków.
- Zbiór częsty
  - Zbiór, którego wsparcie jest nie mniejsze niż zadany próg (minsup).

Zbiory częste – przykład

Customer

id	*Age*	*Gender*	*Income*	*GoodCustomer*
1	30	male	1800	no
2	33	female	2500	yes
3	30	male	2000	no
4	30	female	1800	yes
5	26	female	2500	yes
6	30	male	3000	yes
7	30	female	1600	no

Niech *minsup*=0.5

Zbiory elementów:

$\{30\},\{female\},\{1800\},\{[0,2300]\}\ldots,\{30,male\},\{30,1800\},$
$\{male,1800\},\ldots,\{30,1800,male\},\ldots$

Zbiory częste:

$\{30\}(wsparcie=5/7),\{female\}(wsparcie=4/7)\ldots,$
$\{30,[0,2300]\}(wsparcie=4/7),\ldots$

Od zbiorów elementów do wzorców

- Zbiór elementów może być zapisany w postaci wzorca następującej postaci:

$$a_1 = v_1 \wedge a_2 = v_2 \wedge \ldots \wedge a_k = v_k$$

gdzie $a_1, a_2, \ldots, a_k$ są atrybutami określającymi rodzaj danych elementów, a $v_1, v_2, \ldots, v_k$ są wartościami reprezentującymi te elementy.

- Wzorzec może zawierać również warunki skonstruowane przy użyciu innych relacji niż relacja równości, np. relacja należenia elementu do zbioru.

Od zbiorów elementów do wzorców – przykład

- Zbiory elementów z_1, z_2 :

$$z_1 = \{male\},$$
$$z_2 = \{30, [0,2300]\}$$

mogą być zapisane w postaci następujących odpowiadających im wzorców w_1, w_2 :

$$w_1 : Gender = male,$$
$$w_2 : Age = 30 \wedge Income \in [0,2300]$$

Reprezentacja asocjacji: Reguły asocjacyjne

- Reguła asocjacyjna
  - Wrażenie postaci $X \rightarrow Y$, gdzie *X* i *Y* są (częstymi) zbiorami elementów, takimi że $X \cap Y = \emptyset$.
- Wsparcie reguły asocjacyjnej (ang. *support*)
  - Stosunek liczby jednoczesnych wystąpień zbiorów *X* i *Y* do liczby wszystkich przypadków.
- Zaufanie (wiarygodność) reguły (ang. *confidence*)
  - Stosunek liczby jednoczesnych wystąpień zbiorów *X* i *Y* do liczby wystąpień zbioru *X*.

Zadanie odkrywania reguł asocjacyjnych

- Znalezienie w podanej bazie danych wszystkich reguł asocjacyjnych taki, że:
  - Wsparcie każdej reguły jest nie mniejsze niż ustalony próg *minsup;*
  - Zaufanie każdej reguły jest nie mniejsze niż ustalony próg *minconf;*

Reguły asocjacyjne – przykład

Customer

id	*Age*	*Gender*	*Income*	*GoodCustomer*
1	*30*	*male*	*1800*	*no*
2	*33*	*female*	*2500*	*yes*
3	*30*	*male*	*2000*	*no*
4	*30*	*female*	*1800*	*yes*
5	*26*	*female*	*2500*	*yes*
6	*30*	*male*	*3000*	*yes*
7	*30*	*female*	*1600*	*no*

Zbiory częste

$\{female\}$ (wsparcie=5/7)

$\{30, [0,2300]\}$ (wsparcie=4/7)

Reguła asocjacyjna skonstruowana za pomocą zbiorów częstych

$\{30, [0,2300]\} \rightarrow \{female\}$ (wsparcie=2/7, zaufanie=1/2)

Zapis reguł asocjacyjnych z wykorzystaniem wzorców

- Regułę asocjacyjną można zapisać w następującej postaci:

$$a_1 = v_1 \wedge a_2 = v_2 \wedge \ldots \wedge a_i = v_i \rightarrow$$
$$a_{i+1} = v_{i+1} \wedge a_{i+2} = v_{i+2} \wedge \ldots \wedge a_k = v_k$$

gdzie $a_1, a_2, \ldots, a_k$ są atrybutami określającymi rodzaj danych elementów, a $v_1, v_2, \ldots, v_k$ są wartościami reprezentującymi te elementy.

Zapis reguł asocjacyjnych z wykorzystaniem wzorców – przykład

- Reguła asocjacyjna

$$\{30,[0,2300]\}\rightarrow\{female\}$$

mogą być zapisane w następującej postaci

$$Age=30\wedge Income\in[2000,3000]\rightarrow Gender=female$$

Interpretacja reguł

- Reguła

$$Age = 30 \wedge Income \in [2000, 3000] \rightarrow Gender = female$$

o zaufaniu 1/2 może być zinterpretowana w następujący sposób:

Prawdopodobieństwo, że osoba w wieku trzydziestu lat i dochodzie w przedziale od 2000 do 3000 jest kobietą wynosi 1/2.

Ocena użyteczności reguł asocjacyjnych

- Czy wysokie zaufanie reguły oznacza, że reguła jest użyteczna/interesująca?
- Reguła może zostać oceniona jak użyteczna, jeżeli prawdopodobieństwo spełnienia wniosku reguły jest wyższe niż prawdopodobieństwo spełnienia wniosku przed skonstruowaniem reguły.
- Użyteczność reguły jest rozumiana jako zwiększenie trafności (zaufanie) opisu (część warunkowa) danej sytuacji (cześć wnioskowa).

Ocena użyteczności reguł asocjacyjnych – przykład

Customer

id	*Age*	*Gender*	*Income*	*GoodCustomer*
1	30	male	1800	no
2	33	female	2500	yes
3	30	male	2000	no
4	30	female	1800	yes
5	26	female	2500	yes
6	30	male	3000	yes
7	30	female	1600	no

Które z reguł asocjacyjnych są użyteczne?

$Income < 3000 \rightarrow GoodCustomer = yes$ (zaufanie=0.5)

$Age \geq 30 \wedge Geneder = female \rightarrow GooodCustomer = yes$ (zaufanie=0.67)

Prawdopodobieństwo, że dowolny klient jest dobrym klientem, tj. zachodzi warunek $GooodCustomer = yes$, wynosi 0.57.

Metody eksploracji wykorzystania sieci WWW – analiza asocjacyjna

- Analiza asocjacyjna wykorzystania sieci WWW ma na celu określenie jakie cechy dotyczące danego sposobu wykorzystania sieci
  - występują często razem (wzorce częste)

    lub
  - są od siebie zależne (reguły asocjacyjne).

Wzorce częste – przykład

Pattern		Support %
BIRCH Cluster = cluster-2	Page _/Default.htm	69.468
Page _/Default.htm	Time Per Page Bin = 1-Low	38.367
Time Per Page Bin = 1-Low	Session Duration Bin = 1-Low	36.607
BIRCH Cluster = cluster-2	Session Actions Bin = 2-Medium	36.227
Page _/Default.htm	Session Actions Bin = 2-Medium	35.054
BIRCH Cluster = cluster-2	Session Duration Bin = 1-Low	33.621
BIRCH Cluster = cluster-2	Session Actions Bin = 1-Low	32.637
BIRCH Cluster = cluster-2	Session Actions Bin = 2-Medium Page _/Default.htm	30.290

Jaką wiedzę niosą ze sobą wzorce częste? – przykład

- Na podstawie jednego wzorca:

$$Cluster = 2 \wedge \text{Page\_/Default.htm} \quad (wsparcie = 69\%)$$

Blisko 70% sesji dotyczy użytkowników z grupy 2, którzy odwiedzali stronę Page_/Default.htm.

- Na podstawie więcej niż jednego wzorca:

$$Cluster = 2 \wedge Session\ Actions = Low \quad (wsparcie = 32.6\%)$$
$$Cluster = 2 \wedge Session\ Actions = Medium \quad (wsparcie = 30.3\%)$$

Blisko 63% sesji dotyczy użytkowników z grupy 2, których liczba akcji nie była wysoka.

Generowanie reguł asocjacyjnych

- Na podstawie danych dotyczących sposobu wykorzystania sieci generowane są reguły o wysokim wsparciu i/lub zaufaniu.
- Spośród otrzymanych reguł wybierane są, te które można określić jako użyteczne.

Reguły asocjacyjne dotyczące sposobów wykorzystania sieci WWW – przykład

Consequent	Antecedent	Confidence %	Rule Support %
Page _/Default.htm	BIRCH Cluster = cluster-2	82.870	69.468
BIRCH Cluster = cluster-2	Page _/Default.htm	82.564	69.468
Page _/Default.htm	Time Per Page Bin = 1-Low	80.573	38.367
Time Per Page Bin = 1-Low	Session Duration Bin = 1-Low	91.541	36.607
BIRCH Cluster = cluster-2	Session Actions Bin = 2-Medium	87.277	36.227
Page _/Default.htm	Session Actions Bin = 2-Medium	84.449	35.054
BIRCH Cluster = cluster-2	Session Duration Bin = 1-Low	84.074	33.621
BIRCH Cluster = cluster-2	Session Actions Bin = 1-Low	98.592	32.637
BIRCH Cluster = cluster-2	Session Actions Bin = 2-Medium Page _/Default.htm	86.411	30.290

Jaką wiedzę niosą ze sobą reguły asocjacyjne? – przykład

- Na podstawie jednej reguły:

$$Session\,Duration = Low \Rightarrow Time\,Per\,Page = Low$$
$$(zaufanie = 91.5\,\%)$$

Jeżeli sesja trwa krótko, to prawdopodobieństwo, że czas przeglądania strony jest też krótki wynosi 91.5%.

Jaką wiedzę niosą ze sobą reguły asocjacyjne? – przykład c.d.

- Na podstawie więcej niż jednej reguły:

$$Session\,Duration = Low \wedge Cluster = 2 \Rightarrow Time\,Per\,Page = Low \quad (zaufanie = 90\%)$$

$$Session\,Duration = Low \wedge \text{Page\_/Default.htm} \Rightarrow Time\,Per\,Page = Low \quad (zaufanie = 91.6\%)$$

W przypadku krótko trwających sesji prawdopodobieństwo, że czas przeglądania strony jest też krótki jest większe o 1.6 pkt proc. w przypadku odwiedzin strony Page_/Default niż w przypadku sesji dotyczącej użytkownika z grupy 2.

Ocena użyteczności reguł asocjacyjnych – przykład

- Czy reguła $\text{Page\_/Default.html} \Rightarrow \text{cluster } 2$ o zaufaniu 82.6% jest użyteczna?
- Prawdopodobieństwo, że dowolna sesja dotyczy użytkownika grupy 2 wynosi 84%.

  Zatem reguła z „pustym” warunkiem, tj. $1 \Rightarrow \text{cluster } 2$ posiada zaufanie 84%.

Reguły asocjacyjne o dużej użyteczności – przykład

Consequent	Antecedent	Confidence %	Rule Support %
Time Per Page Bin = 1-Low	Session Duration Bin = 1-Low	91.541	36.607
BIRCH Cluster = cluster-2	Session Actions Bin = 1-Low	98.592	32.637
Time Per Page Bin = 1-Low	Session Duration Bin = 1-Low BIRCH Cluster = cluster-2	89.990	30.255
Session Duration Bin = 1-Low	Time Per Page Bin = 1-Low BIRCH Cluster = cluster-2	82.146	30.255
Time Per Page Bin = 1-Low	Session Duration Bin = 1-Low Page _/Default.htm	91.548	28.788
BIRCH Cluster = cluster-2	Session Actions Bin = 1-Low Page _/Default.htm	100.000	26.355
Time Per Page Bin = 1-Low	Session Duration Bin = 1-Low BIRCH Cluster = cluster-2 Page _/Default.htm	89.637	22.989
Session Duration Bin = 1-Low	Time Per Page Bin = 1-Low BIRCH Cluster = cluster-2 Page _/Default.htm	80.825	22.989

Określanie istotnych cech danego serwisu WWW na podstawie reguł asocjacyjnych

- Aby na podstawie reguły asocjacyjnej określić istotną cechę serwisu WWW, wniosek takiej reguły powinien dotyczyć tej cechy.
- Na podstawie zbioru wszystkich reguł, których wniosek dotyczy danej cechy serwisu można, rozpatrując warunki reguły, określić przyczyny występowania tej cechy.

Określanie istotnej cechy serwisu WWW na podstawie reguł asocjacyjnych – przykład

Rule	Confidence	Support
Time per Page = Low and Cluster 2 ⇒ Session Duration = Low	82.146%	30.255%
Time per Page = Low and Cluster 2 and Page_/Default.htm ⇒ Session Duration = Low	80.825%	22.989%
Time per Page = Low and Session Actions = Medium ⇒ Session Duration = Low	90.694%	17.829%

- Krótki czas trwania sesji
  - każdorazowo wiąże się z krótkim czasem przeglądania stron.
  - często ma miejsce, gdy użytkownik jest z grupy 2.