

WFM-KOSTEN, WINST, SERVICE LEVEL EN KLANTTEVREDENHEID

IN DE VORIGE AFLEVERINGEN VAN DEZE SERIE OVER WORKFORCEMANAGEMENT ZAG JE DAT FLEXIBILITEIT – SCHAALGROOTTE EN ROBUUSTHEID, ZO MIN MOGELIJK SKILLS DOOR DE JUISTE ONDERSTEUNING MET KENNIS, FLEXIBELE ROOSTERS EN INZETMOGELIJKHEDEN E.D. – EIGENLIJK BELANGRIJKER ZIJN OM WFM-PROBLEMEN OP TE LOSSEN DAN DE ALGORITMES ZELF. DIT LAATSTE ARTIKEL IN DEZE SERIE GAAT IN OP HET SERVICE LEVEL EN MOGELIJKE ALTERNATIEVEN DAARVOOR VANUIT DE THEORIE EN DE PRAKTIJK WAARBIJ KLANTTEVREDENHEID IN DE DISCUSSIE WORDT BETROKKEN EN HET RELATIEVE BELANG VAN HET SERVICE LEVEL WORDT TOEGELICHT.

Door prof.dr. Ger Koole en drs. Annemiek van Moorst

Al eerder zag je dat er weinig zeker is in callcenters: grootheden zoals het aanbod fluctueren op een onvoorspelbare manier, je kunt zelfs niet met zekerheid onder- en bovengrenzen aangeven. Flexibiliteit is dan geboden, maar die heeft zijn grenzen en ook zijn prijs. En die prijs kan hoog oplopen. Dit komt doordat je in alle omstandigheden het gewenste service level wilt bieden. Maar waarom wil je dat eigenlijk en wat is gewenst? Om daar antwoord op te kunnen geven moet je eerst wat beter nadenken over de interpretatie van het begrip service level, meestal gedefinieerd als het percentage calls dat in een bepaalde tijdsinterval (meestal 30 minuten) binnen 20 seconden wordt beantwoord. Vaak ben je tevreden met een percentage van 70 of 80. Maar waarom 70? Geef je dan niet om die 30% die langer dan 20 seconden moet wachten? Waarom dus niet 100%? Het antwoord hierop is bekend: dit is onmogelijk. Door

fluctuaties in het aanbod kan het altijd voorkomen dat de klant langer dan 20 seconden wacht, hoeveel agents je ook inzet. Ook zegt bijvoorbeeld een 80/20 service level iets over de 'staart' van de wachttijd: als slechts 20% langer dan 20 seconden moet wachten, dan hoeft bijvoorbeeld maar 15% langer dan 40 seconden

te wachten. Hoe groter het callcenter, hoe lager dit getal.

De constatering dat een 80/20 service level ook iets zegt over het percentage dat langer dan 20 seconden moet wachten is op zich relevant en het is belangrijk om te constateren hoe langzaam dit percentage

Relatie grootte callcenter, vertraagde calls en snelheid service bieden aan klant

Beschouw een callcenter met een AHT van 5 minuten en slechts 10 agents. Bij een aanbod van 1.43 is er volgens de Erlang C formule exact een 80/20 service level en ook een 83.2/40 service level. Bij 20 agents is bij een 80/20 service level het aanbod 3.21 (meer dan 2 keer 1.43 vanwege de schaalvoordelen) en de staart gaat sneller omlaag namelijk naar 84.6/40. Bij 100 agents is het aanbod voor 80/20 18.48 en is het service level voor de staart 87.9/40. Het omgekeerde geldt voor een wachttijd van 0 seconden. Dus in een groot callcenter worden meer calls vertraagd maar worden klanten relatief snel geholpen. Deze getallen zijn berekend met de ErlangC-calculator [1].

omlaag gaat naarmate de calls langer wachten. De vraag waarom je op 80% mikt en niet zo hoog mogelijk, blijft daarom belangrijk. Om hier antwoord op te geven moet je naar de interpretatie van dit getal: het is de fractie van een groot aantal calls die binnen 20 seconden worden beantwoord. Wat betekent dit nu voor een individuele beller? Als een beller vaak belt naar een callcenter met een constant service level van 80/20, dan zal hij ruwweg in 80% van de gevallen minder dan 20 seconden wachten. Ruwweg, want hij (of zij) kan pech hebben. Maar als je maar vaak genoeg belt, komt iedereen bij 80% uit. Dus een 80/20 service level heeft voornamelijk zin voor een callcenter waar klanten vaak terugbellen: dan kun je in 4 van de 5 gevallen een redelijk service level 'garanderen'. Stel je nu voor dat het service level vaak fluctueert, soms 70% en soms 90%, maar gemiddeld komt het op 80% uit. Wat is dan de ervaring van de beller? Dit hangt dan af van zijn gedrag: belt hij op wisselende tijdstippen, of zit er een zeker patroon in zodat de beller toevallig vaak op een laag service level stuit en daardoor ook een laag gemiddeld service level ervaart? Laten we een paar voorbeelden geven van zulke patronen. Beschouw een callcenter dat gedurende de ochtend en middag een bovengemiddeld service level heeft, maar dat tijdens de lunchperiode door de lunch van de agents inzakt. Een klant die regelmatig van zijn eigen lunchpauze gebruikmaakt om te bellen, zal dan vaak een laag service level aantreffen. Een ander voorbeeld is een klant die ophangt vanwege de lange wachttijd. Analyse van callcentergegevens leert dat klanten die ophangen vaak snel terugbellen: als het service level laag is, is de kans groot dat er weer een negatieve ervaring volgt. Was de lange wachttijd het gevolg van fluctuaties maar is het service level in orde, dan zal de klant nu waarschijnlijk wel snel aan de beurt zijn. Het gedrag van klanten ten aanzien van redials is dus zeer belangrijk. Hoe dit er voor een willekeurige callcenter uitziet, volgt uit een analyse van de data. Bij een dergelijke analyse is het van essentieel belang dat de identiteit van afhakers bekend is, bijvoorbeeld door Calling Line Identification of het intoetsen van een klantnummer.

WAAROM EEN 80/20 SERVICE LEVEL?

Waarom wordt er meestal gekozen voor een 80/20 service level? In de meeste callcenters weet je niet wat de opbrengst van een call is. Vaak wordt de klant bediend om negatieve publiciteit te voorkomen of om minder klanten te verliezen of om in de toekomst nog eens een product te kunnen verkopen. Daarnaast zijn er callcenters waar je een indicatie hebt van de opbrengsten van een call. Bijvoorbeeld in sales callcenters zoals een beursorderlijn of een postorder callcenter.

In het eerste soort callcenters, waar je geen idee hebt van de opbrengst van een call (of de beperking van een toekomstig verlies, of beperking van verlies door het voorkomen van negatieve publiciteit), wordt veelal gekozen voor een 80/20 service level. De reden hiervoor is simpel en heeft met kosten te maken. Werkgevers hebben liever geen personeel dat zit te niksen en bij een 80/20 servicelevel weet je dat agents een gezond deel van de tijd aan de telefoon zitten, terwijl een groot deel van de klanten binnen 20 seconden wordt geholpen. De meeste bellers wachten enige

ook benutten om marketinginformatie te laten horen: on hold marketing. Voor beide situaties geldt dat de boodschap of de muziek niet langer dan 30 seconden mag duren omdat anders een negatief effect bereikt wordt.

Het motiveren van een 80/20 service level doelstelling is moeilijker in het geval van eenmalige of eerste contacten waarbij de beller die afhaakt niet terugbelt. Als de tijd tot afhaken op 20 seconden wordt geschat, dan betekent een 80/20 service level dat ongeveer 20% van de bellers afhaakt en daarmee verloren gaan. Of 20 seconden correct is, hangt af van het callcenter. Het zal waarschijnlijk hoger liggen en moet volgen uit een data-analyse. Belangrijker is de keuze voor 80%. Waarom 80%? Waarom accepteer je dat je 20% van je klanten kwijtraakt? Antwoord: dat doe je niet, je gaat intuïtief al op een hoger service level zitten. Maar hoe onderbouw je dat?

MAXIMALE WINST EN SERVICE LEVEL BIJ INBOUND SALES

In een (inbound) sales callcenter waar je een indicatie hebt van de opbrengsten van een gemiddelde call, kun je het service le-

Model voor bepalen van de winst

Veronderstel een callcenter met een aanbod van 10 calls per minuut, een AHT van 5 minuten, gemiddelde tijd tot afhaken 1 minuut, prijs per agent per uur 12 euro en winst per call gemiddeld 2 euro. Met behulp van een software tool, gebouwd door drs. Auke Pot van de Vrije Universiteit, kun je het optimale aantal agents en de winst bepalen [2]. In figuur 1 zie je hoe je het tool in moet stellen en tevens vind je het optimale aantal agents: 51. Het bijbehorende service level is 90% voor bellers die bereid zijn te wachten tot ze een agent aan de lijn krijgen. Het is interessant om zelf hiermee te experimenteren. Bijvoorbeeld het aantal agents als invoer te nemen en deze te variëren: dan zie je hoe ongevoelig dit model hiervoor is. Ook het variëren van de opbrengst per call is interessant.

tijd afhankelijk van hun vraag of bellen op een ander tijdstip terug. Uit allerlei onderzoeken – o.a. van de Erasmus Universiteit, Universiteit van Wageningen en ander vooral psychologisch onderzoek uit Amerika – blijkt bijvoorbeeld het volgende. Telefoonkosten hebben een negatieve invloed op de perceptie van de wachttijd. Muziek kan afhankelijk van het type en de lengte de waarneming positief of negatief beïnvloeden. Daarnaast kun je de wachttijd

vel beter afstemmen, namelijk door uit te rekenen bij welk service level de winst maximaal is. Hiervoor heb je dan nog wel wat informatie nodig. Je moet namelijk weten na hoeveel tijd klanten ophangen als ze niet worden geholpen. Dit vergt een uiterst complexe berekening, want er is een percentage dat na 40 seconden ophangt, een percentage dat na 50 seconden ophangt, et cetera. Je gaat dus eigenlijk kijken wat een tevreden klant oplevert,

Figuur 1

wat een extra agent kost en wat hij/zij aan calls verwerkt en dus oplevert en je invalshoek verandert van een cost center naar een profit center. De wet van de afnemende meeropbrengsten geldt ook in dit geval: er is een optimaal aantal agents en ook een service level dat daarmee correspondeert. Dat kan heel wel 80% zijn maar dat hoeft absoluut niet. Een rekenvoorbeeld staat in het kader Model voor bepalen van de winst.

Voor het vinden van de optimale oplossing met behulp van bijvoorbeeld een mo-

del als hierboven beschreven, is het noodzakelijk dat alle kerngetallen bekend zijn. Alleen dan kan routinematig voor elke interval het beste aantal agents worden bepaald. Dit bewijst eens te meer het nut van een uitgebreide data-analyse.

Bereikbaarheid speelt een bijrol in klantcontact, zo blijkt weer eens uit de samenvoeging van een aantal Europese studies over post- en pre-sales klantcontacten zowel in de b2c als in de b2b markt door een niet met name genoemd instituut waar Van der Poel onlangs over rapporteerde [3]. Het gaat om inhoud en kennis van zaken. Hoe minder tevreden met de oplossing, hoe minder tevreden over de wachttijd. Als je geen antwoord hebt gekregen, ben je even ontevreden, of je nu kort of lang hebt moeten wachten. Bereikbaarheid gaat pas een rol van enige betekenis spelen wanneer er een bevredigend antwoord is gegeven. Het eindresul-

taat van het gesprek kleurt dus de herinnering aan de verschillende aspecten van het gesprek. Met andere woorden, First Contact Resolution (FCR) is dus de belangrijkste prestatie-indicator in een contactcenter [4]. Het service level is een zogenaamde hygiënefactor die in verschillende omstandigheden (inbound service, inbound sales) en afhankelijk van de kwaliteit (FCR!) aan verschillende eisen moet voldoen om bij te dragen aan de klanttevredenheid. Een adequaat besturingsmodel – of in de tegenwoordig gangbare terminologie Business Process Management – met situationeel bepaalde en uitgewerkte beslisregels is essentieel om kwaliteit en snelheid of kwantiteit in balans te brengen. Ook outsourcers worden in toenemende mate op kwaliteit afgerekend. Diepgaande kennis over de besturing van contactcenters en helpdesks is vereist om hierin de juiste keuzes te maken zoals de ervaring uitwijst [4]. **CCM**

WAT BEPAALT KLANTTEVREDENHEID?

[1] Op <http://www.math.vu.nl/~koole/ccmath/ErlangC.php>.
 [2] Zie <http://www.math.vu.nl/~sapot/software/ErlangProfit>.
 [3] Van der Poel, R. Eind goed al goed. *Telecommerce Magazine*, nr. 4 2005.
 [4] Van Moorst, A.M. *Strategische sourcing van customer care. Kerncompetentie delen!* BBP, 2004.

[1] Op <http://www.math.vu.nl/~koole/ccmath/ErlangC.php>.

[2] Zie <http://www.math.vu.nl/~sapot/software/ErlangProfit>.

[3] Van der Poel, R. Eind goed al goed. *Telecommerce Magazine*, nr. 4 2005.

[4] Van Moorst, A.M. *Strategische sourcing van customer care. Kerncompetentie delen!* BBP, 2004.

Annemiek van Moorst is directeur Tote-m business architects – annemiek@tote-m.com – en Ger Koole is hoogleraar Optimalisatie van Bedrijfsprocessen aan de Vrije Universiteit Amsterdam – koole@few.vu.nl

Een wiskundige blik op het service level

In een cost center is een service level definitie gebruikelijk, zoals 80% van de calls moet binnen 20 seconden worden beantwoord. Maar is dit wel een verstandige keuze? Natuurlijk, naast dit percentage moet ook de abandonment rate bekeken worden, maar het is zeer de vraag of iets als 80/20 de juiste maat is voor de wachttijd. Beschouw de volgende hypothetische situatie. Een outsourcer wordt puur afgerekend op zijn service level: 80/20. In dit geval heeft hij er geen enkel belang bij calls die langer dan 20 seconden wachten te helpen. Die bedien je niet tot ze vanzelf afhaken. Ook in minder absurde situaties, waar bijvoorbeeld sprake is van meerdere skills en een verschillend service level per skill, kan de gebruikelijk service level norm tot vreemde situaties leiden. Een alternatief is de ASA, maar deze is niet gevoelig voor fluctuaties in de wachttijd. Een goed alternatief zou zijn de 'Average Excess Time' (AET): de tijd die calls langer dan een bepaalde tijd (bijvoorbeeld 20 seconden) in de wachtrij doorbrengen. Neem bijvoorbeeld 5 calls met wachttijden 0, 15, 25, 30 en 50 seconden. Dan is het service level 40% < 20 sec., de ASA 24 sec. en de AET 7. De AET stimuleert tot het behandelen van lang wachtende calls en is, in tegenstelling tot de ASA, wel gevoelig voor fluctuaties. Neem bijvoorbeeld als wachttijden 5 maal 24 seconden. Dan is het service level 0% < 20 sec., de ASA blijft gelijk op 24 en de AET is 4. Intuïtief heb je liever geen uitschieters in wachttijden, hetgeen pleit voor de AET.