UPPSALA
UNIVERSITET

UPTEC STS 16 014
Examensarbete 30 hp
Juni 2016

Mötet mellan automatiserade informationsinsamlare och användarkrav

En studie om webbspindlars funktion vid utveckling av en eventkalender

Åsa Bengtson
Karl Söderlund

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

The Encounter Between Web Crawlers and User Requirements

Åsa Bengtson & Karl Söderlund

Web crawlers constitutes an important component in the ability of performing Internet searches today. Their aim is to collect and categorize web based information in order to make the information easier to search for. Therefore, individuals and professionals can access information more efficiently with the help of web crawlers. The ability to access relevant information today is still limited due to a missing mutual structure concerning how information is presented on the Internet. If it would be possible to compile and present relevant information in a more efficient way, the advantages would be many. One example concerns companies who want to promote capacity building among staff by participation in conferences, seminars and other industry specific events.
The aim of this study was to illuminate how well a web crawler can collect information and in what extent this information would correspond with what a company's users are seeking. The aim was also to find how the distribution of information can be made more efficient on the Internet and be adapted to suit the users better. The aim was met by implementation of interviews, a survey and the development of an event calendar prototype consisting of a web crawler and a web page where extracted information is presented.
The study showed that the most requested attributes were the hardest to extract, and that the inconsistent structure of web pages prevented the web crawler from performing optimally. Three solutions were proposed; change of user requirements, altering of the web crawling technique and improvement of the methods used to present information on the web. The latter is concluded to have the greatest impact on the overall performance of web crawlers and distribution of web based information.

Handledare: Jonnie Hedqvist & Jessica Nylund
Ämnesgranskare: Mats Lind
Examinator: Elísabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 16 014

Förord

Följande rapport är ett av flera resultat som detta examensarbete utmynnat i. Examensarbetet är genomfört inom civilingenjörsprogrammet System i Teknik och Samhälle (STS) vid Uppsala universitet. Arbetet är utfört i samarbete med ÅF AB i Solna under tidsperioden januari 2016 till juni 2016.

Ett annat resultat av detta examensarbete är den eventkalender som ÅF:s medarbetare i framtiden ska kunna använda för att finna relevanta evenemang som ett steg i sin kompetensutveckling. Examensarbetet har även resulterat i ökad inblick i livet som konsult, lärdomar kring mjukvaruprogrammering och rapportskrivning samt resulterat i en givande sista termin som studenter.

Slutligen är det ett antal personer som förtjänar att lyftas fram och tackas för deras hjälp med detta examensarbete. Först och främst är det Jessica Nylund och Jonnie Hedqvist som tillsammans har handlett examensarbetet från början till slut. Denna handledning har bland annat inneburit tekniskt stöd i utvecklingen av eventkalendern och erfarenheter kring hur saker fungerar på ÅF, men framförallt har det inneburit stöd och gemenskap. Tor Ericsson bör också tackas för sitt stöd, sina råd och erfarenheter gällande kompetensutveckling och branschinsikt. Slutligen bör uppmärksamhet riktas mot ämnesgranskare Mats Lind som funnits som bollplank, bidragit med råd och väglett oss framåt i alla skeden av detta examensarbete.

Åsa Bengtson och Karl Söderlund.

Solna, juni 2016.

Populärvetenskaplig sammanfattning

En viktig komponent som möjliggör sökning på internet idag är de program som kallas för webbspindlar. Deras uppgift är att samla in och kategorisera webbaserad information för att göra den möjlig att söka efter. Tack vare dem är det möjligt för privatpersoner och yrkesverksamma att få tillgång till information enkelt och effektivt. Möjligheten att få relevant information presenterad för sig utan att aktivt söka efter den är dock fortfarande begränsad. Detta som en följd av frånvaron av gemensam struktur som information presenteras i på internet. Skulle det gå att sammanställa relevant information på ett effektivt sätt skulle fördelarna vara många. Ett exempel är för företag som vill främja kompetensutveckling bland sina anställda genom att delta i konferenser, mässor eller andra branschspecifika evenemang i större utsträckning.

Denna studie genomfördes i samarbete med ÅF AB. För anställda på ÅF innebär det en viss tidsinvestering att söka efter relevanta evenemang, vilket innebär att många potentiellt intresserade avstår från att delta i dem. Studien gjordes för att undersöka hur väl en egenutvecklad webbspindel kan extrahera webbaserad information och presentera den på ett intuitivt vis. Syftet var att undersöka skillnaden mellan den information som webbspindeln kunde extrahera med den information som anställda var intresserade av, samt hur den eventuella skillnaden kunde minimeras.

För att undersöka behovet bland anställda på ÅF genomfördes både djupintervjuer och en omfattande enkätundersökning, där anställda fick besvara vilken typ av evenemang de var intresserade av att delta i och vilken information de ville få presenterad för sig. Dessutom utvecklades ett antal webbspindlar som programmerades att extrahera data från på förhand givna webbsidor. Den data som extraherades bearbetades och lagrades i en databas för att sedan göras tillgänglig på en webbsida i form av en eventkalender.

Intervjuerna och enkäterna visade att det fanns ett intresse för att öka deltagande på externa evenemang, men att många anställda prioriterade bort deltagande på grund av tidsåtgången. Utvecklingen av den prototyp som gjordes i denna studie visade att frånvaro av gemensam struktur på internet hämmade webbspindelns förmåga att extrahera information på ett optimalt sätt. Vidare visade det sig att mycket av den information som anställda ansåg vara viktigast i en sammanställning även var svårast att extrahera från internet.

Studien visar att de tekniska hjälpmedel som finns för att strukturera information på internet inte används i praktiken i den utsträckning som de kan, vilket som följd gör webbspindlar och liknande tekniska lösningar mer ineffektiva. Vinsten av att använda en gemensam struktur på nätet skulle, som följd av de resultat studien visar, bidra till mer effektiv spridning av relevant information och mer skalbara lösningar för att extrahera webbaserad information.

Innehållsförteckning

1. Inledning

1.1 Introduktion

1995 hade mindre än en procent av jordens befolkning tillgång till internet (Internet Live Stats, 2016). Idag har drygt 40% av jordens befolkning internetuppkoppling och världens internetanvändare fortsätter öka i mycket snabb takt (Internet Live Stats, 2016). Det finns fortfarande en stor klyfta i användning av internet mellan olika länders befolkning där Sverige tillsammans med Norge och Island ligger i topp med 95% av befolkningen uppkopplad på internet jämfört med Bangladesh där bara 6% av befolkningen är uppkopplad (Carter, 2015). Kombinationen av att internetanvändning och antal uppkopplade användare ökar, samt att det dynamiska nätverk som består av cirka en miljard sidor (Internet Live Stats, 2016) också växer, ger oss människor större möjlighet till framgångsrikt och effektivt användande av internet. Med stora mängder tillgänglig information ökar också möjligheterna till olika användningsområden. Eftersom innehållet på internet saknar gemensam struktur har det utvecklats viss typ av programvara för att underlätta för användare att finna relevant information (Ahuja et al. 2014). En typ av program som utvecklats för att ge mer struktur åt den ständigt växande informationsmängden är så kallade webbspindlar. Dessa programs syfte är att söka igenom webbsidors innehåll och spara särskild typ av information för att kategorisera denna information och göra den mer lätthanterlig.

1.2 Problemformulering

Att finna viss information på internet är i dagsläget sällan ett problem. Tjänster som exempelvis Google gör det enkelt att snabbt få tillgång till miljontals källor och genom väl valda sökord nå önskad information inom loppet av några sekunder. För att finna rätt information handlar det i de flesta fall om en aktiv process där användaren måste ställa en fråga till en sökmotor för att hitta den rätta informationen. Det är dock inte alltid så att användaren vet vilken information den söker utan istället vill få förslag presenterade för sig baserat på individuella preferenser.

Explosiv tillväxt av sidor på internet har resulterat i invecklade och sammanvävda system av webbsidor som kräver särskilda färdigheter av användaren samt avancerade verktyg för att hjälpa användaren att finna önskad information (Shirgave & Kulkarni, 2013). Att finna önskad information på internet har blivit en kritisk ingrediens i det vardagliga livet privat, inom utbildning och professionellt. På grund av efterfrågan av “rätt information” finns även ökad efterfrågan av sofistikerade verktyg för att hjälpa användaren navigera webbsidor och finna den efterfrågade informationen. Användarna måste få information och tjänster presenterade för sig som stämmer överens med deras specifika behov snarare

än en odifferentierad massa av information. För att upptäcka intressanta och frekventa navigeringsmönster från webbloggar har många nya webbutvinningsverktyg utvecklats (Shirgave & Kulkarni, 2013).

Det faktum att användare inte aktivt kan söka efter något de inte vet finns innebär potentiellt förlorad information. Samma problem gäller för organisationer som vill ligga i framkant inom sitt verksamhetsområde. Är de anställda inte medvetna om utbildningar, konferenser eller mässor går de miste om möjligheter till individuell och kollektiv kompetensutveckling. Det omvända gäller för personerna bakom nämnda evenemang; vet de inte vilka som potentiellt är intresserade av att delta är det svårt att marknadsföra sig via tillgängliga kanaler.

Skulle ett företag enkelt kunna utveckla en programvara som möjliggör automatisk inhämtning av relevanta evenemang skulle det innebära flera fördelar. Dels för arrangörer som skulle ges ökad exponering utan att aktivt behöva marknadsföra sig. Dels för användare som skulle få aktuella evenemang presenterade för sig som de annars inte hade känt till. Webbspindlar är ett möjligt verktyg för att både samla in och tillgängliggöra information om evenemang till intresserade parter. För att detta ska fungera på ett tillfredsställande sätt krävs dock att en webbspindel både kan lokalisera relevanta evenemang, och sedan extrahera den information som potentiella användare anser vara viktig.

1.3 Syfte och frågeställning

Denna studie ämnar utreda hur väl en automatiserad informationsinsamlare på internet, så kallad webbspindel, kan samla in textbaserad information och i vilken utsträckning denna information överensstämmer med vad ett företags användare efterfrågar. Detta i syfte att ta reda på hur informationsspridning kan effektiviseras och anpassas efter användare.

Målet med studien är att besvara frågan: *I vilken utsträckning skiljer sig informationen som kan hämtas via en webbspindel från den information som ett företags användare efterfrågar, och hur kan denna skillnad förminskas?*

1.4 Avgränsningar

Studien är utförd i samarbete med ÅF AB, ett svenskt konsultföretag med stor internationell närvaro. Studien kommer enbart att undersöka verksamheten i Sverige och baseras på de krav som framkommer från ÅF:s anställda i Solna.

För studien har en webbspindel utvecklats. Denna har inte utvecklats från grunden, utan baserats på öppen källkod, det vill säga redan uppsatta ramverk av programmeringskod som anpassats efter studiens ändamål. Detta val gjordes för att studiens begränsade

tidsram medförde att utveckling från grunden ansågs vara för omfattande för att vara försvarbart.

Informationen som webbspindlarna programmeras att samla in rör branschspecifika evenemang, exempelvis mässor, föreläsningar och seminarier, där endast textbaserad information berörs. Andra publika evenemang undersöks inte i denna studie, även om vissa slutsatser går att generalisera för fler typer av information på internet.

1.5 Rapportens struktur

Rapporten inleds med ett kapitel där en genomgång görs av relaterad forskning med fokus på informationspresentation på internet, webbspindlar och viss kritik mot dessa. Även visst användarperspektiv tas upp. I kapitel 3 behandlas studiens metodval. Därefter följer en genomgång av studiens resultat med avsnitt som presenterar intervjuresultat och enkätresultat för att belysa den del av frågeställningen som gäller användarna. I kapitel 5 kommer en prototyp presenteras och gås igenom som utvecklats för att belysa den andra delen av frågeställningen gällande vilken information webbspindlar kan utvinna. Senare i rapporten presenteras ett analyskapitel och senare ett diskussionskapitel där diskrepansen mellan vad användare efterfrågar och vad webbspindlar kan utvinna diskuteras, samt hur denna diskrepans kan förminskas. Slutligen dras slutsatser och ett avsnitt gällande fortsatta studier presenteras.

1.6 Presentation av fallföretaget

ÅF AB är ett ingenjörs- och konsultföretag med cirka 7500 anställda som har uppdrag inom energi, industri och infrastruktur (ÅF, 2016). Enligt ÅF:s årsredovisning 2014 har företaget en nettoomsättning på cirka 9 miljarder kronor och har kontor i mer än 30 länder. Bland ÅF:s kunder återfinns företag inom ett stort antal industribranscher samt privata och offentliga verksamheter. Några av ÅF:s största kunder är Volvo, EON, Ericsson, Försvarets Materialverk och Oslo Lufthavn. Under 2014 genomfördes projekt i drygt 90 länder.

Enligt ÅF:s hemsida grundades Sveriges första industriföretag, Ångpanneföreningen, 1895 i Malmö. Företagets uppgift var att bevaka ägarna till ångpannor och andra tryckkärls intressen och med återkommande besiktningar kontrollera säkerheten för att förhindra olyckor. 1977 förstatligades Ångpanneföreningens besiktningsverksamhet men konsulttjänsterna blev kvar i företaget. 1986 noterades Ångpanneföreningen på Stockholms Fondbörs och bytte år 2008 namn till ÅF. 2010 såldes all besiktningsverksamhet ut och ÅF befäste ytterligare inriktning som ett tekniskt konsultföretag.

ÅF:s årsredovisning (2014) visar att företaget består av de fyra divisionerna Industry, Infrastructure, Technology och International. ÅF AB är ett svenskt publikt aktiebolag och

mellan aktieägarna, styrelsen, verkställande direktören och företagsledningen fördelas styrning, ledning och kontroll av företaget. Bolagsstämman är företagets högsta beslutande organ och den största aktieägaren år 2014 var Stiftelsen ÅForsk med 37% av rösterna. Stiftelsen ÅForsk bildades 1985 som en arvtagare till Ångpanneföreningen och är en forsknings- och utvecklingsorganisation som bland annat verkar inom miljö, infrastruktur och energi (Stiftelsen ÅForsk, 2016).

2. Relaterad forskning

I detta kapitel kommer tidigare forskning relaterat till studiens övergripande ämne presenteras. Först presenteras evenemang ur ett marknadsföringssyfte, sedan följer en genomgång av hur information presenteras på webbsidor följt av information bland annat kring den semantiska webben, dynamiskt webbinnehåll, webbspindlar och kritiken mot dessa. Avslutningsvis behandlas viss forskning kring användarförväntningar.

2.1 Evenemang ur marknadsföringssyfte

Brand experience är ett koncept som växt i betydelse under de senaste åren och bygger på att stärka varumärken och kundupplevelsen framkallade av varumärkesrelaterade stimuli (Tafesse, 2016). Tafesse (2016) beskriver att inom brand experience har eventmarknadsföring etablerat sig som en form av marknadsföringskommunikation. Eventmarknadsföring ses som ett kommunikationsverktyg vars syfte är att sprida ett företags marknadsföringsmeddelande genom att inkludera målgruppen i interaktiv och experimentell aktivitet. Dessa typer av evenemang i marknadsföringssyfte växer i popularitet som ett alternativt kampanjverktyg och marknadsförare investerar intensivt i dessa. Denna ökning kan kopplas till arrangörers växande medvetenhet av eventmarknadsförings effektivitet (Tafesse, 2016).

2.2 Hur information presenteras på webbsidor

Sirsap (2014) beskriver att informationen som presenteras på internet idag mestadels är uppbyggd av semistrukturerad text som är representerad i HTML-format. Detta format är en standard som är utformad för att dels vara läsbara för människor, men även för datorer, genom att tilldela viss information attribut för att kunna ge struktur åt löpande text (Richards, 2006).

Trots att en webbsida kan tyckas bestå av endast text och bilder finns det i den bakomliggande källkoden ett ramverk för att namnge olika delar av sidans innehåll (Buxton, Melton, 2011). I HTML-kod skrivs allt innehåll inom så kallade elementmarkeringar, exempelvis:

```
<h1> Rubrik </h1>
```

Texten i markeringarna, i detta fall “h1”, ger innehållet mellan elementmarkeringarna särskilda egenskaper, till exempel textstorlek eller fetstil. All text inom elementmarkeringar behöver dock inte ge innehållet en egenskap, utan kan istället användas för att identifiera textinnehållet (Buxton & Melton, 2011). Att kategorisera innehållet genom elementmarkeringar istället för genom textinnehåll är vad som ger webbspindeln möjlighet att söka efter en viss typ av information från en webbsida (Richards, 2006). Det saknas dock en gemensam praxis kring hur elementmarkeringar

benämns, vilket medför svårigheter för att söka igenom flera webbplatser med samma metod (Buxton & Melton, 2011).

För att kunna lokalisera specifik information i en webbsidas källkod finns ett programspråk som heter Xpath. Xpath är ett språk som inte används separat, utan tillsammans med andra språk för att navigera i exempelvis HTML-dokument, antingen genom att söka efter namn på elementmarkeringar, eller genom specifika positioner i dokumentet (Richards, 2006).

Det finns tre sätt att organisera data på internet; strukturerad, ostrukturerad och semistrukturerad data. Skillnaden mellan dessa kan vara svårtydd (Rouse, 2014). Rouse menar att ostrukturerad data inte har organiserats för att passa ett visst format. Väldigt lite data är idag helt ostrukturerad eftersom även data som tycks vara helt oorganiserad, så som bilder och vissa dokument, ändå bygger på någon sorts struktur. Strukturerad data är motsatsen till ostrukturerad data eftersom den blivit formaterad så att dess element är organiserade i en struktur som lätt kan kommas åt, kombineras med annat och användas för önskvärt ändamål. Strukturerad data förekommer vanligtvis i databaser. Semistrukturerad data ligger mellan dessa två ytterligheter och är vanligast. Den är inte organiserad i komplexa strukturer som gör sofistikerad analys möjlig, men har ändå viss information kopplad till elementmarkeringar och ämneskategorier (Rouse, 2014).

Vidare förklarar Sirsap (2014) att information på semistrukturerade sidor oftast saknar formaterad dokumentstruktur och att i vilken utsträckning textdokumenten är strukturerade eller inte bland annat beror på syftet och storleken av den information som presenteras. Det finns ingen fullständig, generell semantisk struktur som textdokument anpassas till när de skrivs i olika format på webbsidor, vilket gör det svårt för automatiska informationsinsamlare att fungera effektivt (Sirsap, 2014). Även om mycket forskning gjorts kring hur strukturer och arkitekturer av webbsidor kan förbättras är dessa förändringar i stor utsträckning dolda för användarna som inte upplever förbättringarna (Sidiropoulos et al., 2008). Från användarens perspektiv finns två distinkta skillnader mellan äldre och nuvarande webbsidor; den ökande dynamiken och specialanpassningen av innehållet, samt ökningen av komplexiteten kring designen av webbsidan (Domenech et al., 2010). Dessa skillnader påverkar även hur framgångsrika automatiska informationsinsamlare kan vara eftersom de får allt större svårighet att samla in information allteftersom informationen presenteras mer dynamiskt och komplext på webbsidorna (Domenech et al., 2010).

2.3 Semistrukturerad information på webben

Serrano et al. (2007) menar att den fullständiga bilden av internets struktur och dess karaktärsdrag är helt beroende av utformningen av de verktyg som används för att observera den. De menar att det finns stora svårigheter i att karaktärisera webben när den är så omfattande och har innehåll som både kan expandera och bli inaktuell i hög fart. De

beskriver att trots att det idag finns en ungefärlig bild av webben och dess innehåll tack vare automatiserade informationsinsamlare saknas fortfarande en definitiv bild av webbens egenskaper och arkitektur. Detta kan också påverka hur effektivt vi kan navigera, söka, indexera och utvinna information. Deras förslag är att innan större slutsatser kan dras bör webben och dess struktur undersökas ytterligare (Serrano et al., 2007).

När det gäller HTML-sidor och deras struktur lägger designen av HTML stor vikt vid presentationen av data snarare än manipulationen av data, så att den passar ändamålet och organisationen "bakom skärmen" (Dong et al., 2013). Bristande organisation av data på webben skapar stora svårigheter vid integration och insamling av data (Dong et al., 2013). Alvarez et al. (2008) skriver att när det kommer till webbspindlar finns begränsningar eftersom den presenterade informationen sällan är strukturerad och på samma form webbsidor emellan. De förklarar att för det första måste flera webbsidor vara på samma form och använda samma mall sinsemellan för att informationsinsamlarna ska fungera automatiskt. För det andra måste antagandet göras att sidorna håller samma struktur genomgående (Alvarez et al., 2008). Alla informationsinsamlare som används i dagsläget kräver mänskligt ingripande för att skapa och konfigurera webbspindlarna för att fungera för extrahering. När de olika webbkällorna inte är kända på förhand fungerar inte denna metod (Alvarez et al., 2008).

2.4 Den semantiska webben

I förhållande till den semistrukturerade webben ligger begreppet "semantisk webb" relativt nära. Zulqurnan et al. (2016) beskriver att den semantiska webben väcker allt mer intresse på grund av sin potential. Semantisk webb gör presenterad information begriplig för maskiner och maskiner tillåts alltså tolka data som publicerats av webben i maskinbegriplig form. Detta borde gynna både privatpersoner och företag, förklarar Zulqurnan et al. vidare. Många företag rör sig exempelvis mot e-handel och skapar e-handelssidor, men det finns problem med de innevarande e-handelssystemen på grund av brist på ordentlig standardisering av webben (Zulqurnan, 2016). Dagligen söker tusentals människor efter information eller produkter som de vill ha men på grund av ineffektiviteten av nuvarande system ödslas mycket tid och resurser av användaren. Semantisk webb har möjligheten att övervinna flera av dessa problem och kan accelerera företag till ytterligare högre nivåer där bland annat e-handelssidor kommer spela en viktig roll (Zulqurnan, 2016).

Shirgave och Kulkarni (2013) argumenterar för att metoder bestående av enbart användarbaserad teknik kan förbättras avsevärt genom att integrera även det som finns på webbsidan. Genom att kombinera webbsidans struktur och innehåll med användarbaserad informationsutvinning kan processerna effektiviseras. Författarna förespråkar metoder där traditionell webbanvändarutvinning semantiskt berikas och där fälten i

webbanvändarutvinning kombineras med karaktär av den semantiska webben. I de föreslagna metoderna är den ostrukturerade informationen som fås ut från användardata berikad med semantisk information som extraherats från webbsidor och webbsidestruktur. De anser, med sina vetenskapliga artiklar i bakgrunden, att denna typ av informationsutvinning är mycket användbar och mer träffsäker än många av de andra metoder som finns i dagsläget.

2.5 Dynamiskt webbinnehåll

Bland alla hundratals miljoner webbsidor på internet (Kim et al., 2012) finns en hel del sidor som innehåller dynamiskt presenterad information, det vill säga information som förändras varje gång en sida efterfrågas och laddas (Computer Sweden, "IT-ord"). Denna typ av dynamiska webbsidor är vanligt förekommande och försvårar för webbspindlar ytterligare (Kim et al., 2012). Vidare förklarar Kim et al. (2012) hur dynamiskt webbinnehåll skiljer sig från statiska webbsidor som mer liknar innehåll från bibliotek och onlinedatabaser. Först och främst är dynamiska sidor istället sammansatta av komplexa, ickehierarkiska strukturer. För det andra har de mycket korta cykler för skapande och sedan förintelse av information vilket medför att sidorna förändras ofta och inte har tydliga fysiska gränser. För att kunna söka efter information och spara ner den med hjälp av en webbspindel från liknande sidor krävs därför komplicerade sökprocesser och strategier (Kim et al., 2012).

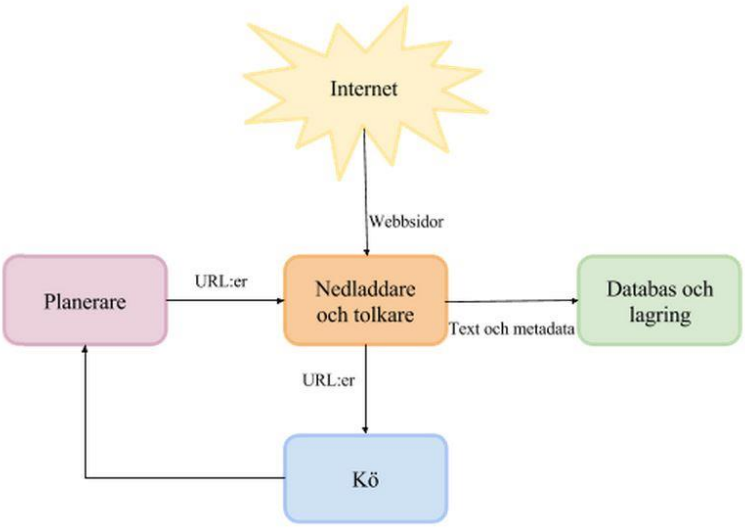
2.6 Standardiserad struktur

Semantisk informationsutvinning har problem med att automatiskt utvinna relevant information från en ostrukturerad webbsida. Chooralil et al. (2015) förklarar att även om det i många fall finns ett rättframt angreppssätt för automatisk utvinning från stora mängder information, återstår problemet fortfarande i att för datorn automatiskt känna igen relaterad information av andra eller högre ordningens förbindelse. Resource Description Framework (RDF) skapar gemensam arkitektur och gemensamma system för information på den semantiska webben genom en extra mall utöver och ovanpå bland annat HTML (Duval et al., 2002). RDF är utformat för att stödja återanvändning av vokabulär inom elementmarkeringar genomgående i hela koden (Duval et al., 2002). På detta sätt skulle exempelvis alla sidor som presenterar evenemang vara strikt styrda till att spara titel under elementmarkeringen "title", alla datum under "date" etcetera för att alltid ha rätt attribut på rätt plats och underlätta informationsutvinning av denna typ. Duval et al. (2002) förklarar att RDF även kan tänkas på i termer av en framgångsrikt konstruerad byggnad. RDF är då ett sätt att organisera och skapa byggnadsdelar för att lättare kunna passa när de senare sammanfogas till stora formationer och konstruktioner. Är delarna inte av samma karaktär redan när de börjar byggas och därför inte kan passa ihop i ett senare skede är det mycket svårare att få en fungerande byggnad i slutändan (Duval et al., 2002).

2.7 Webbspindelns funktion

Batsakis et al. (2009) beskriver webbspindlar som verktyg vars uppgift är att samla in webbinnehåll och spara detta innehåll lokalt. Webbspindlar har utvecklats för att tillfredsställa individers och organisationers behov att skapa och bibehålla olika typer av ämnesspecifika webbportaler lokalt. En webbspindel kan även anpassas till att lokalisera komplex, specialiserad information som inte en vanlig webbsökning kan återge. Typiska krav som ställs på tekniken rörande webbspindlar är att dessa ska tillfredsställa användarnas behov för högkvalitativa och aktuella resultat medan mängden resurser som går åt för att genomföra sökningen, i form av bland annat tid och utrymme, minimeras (Batsakis et al., 2009).

Vidare förklarar Batsakis et al. (2009) att så kallade fokuserade webbspindlar försöker ladda ner så många webbsidor som möjligt som är relevanta för det efterfrågade ämnet medan de försöker hålla antalet icke-relevanta sidor till ett minimum. Webbspindlar ges ett antal startsidor från internet som deras input (representerat som ”Internet” i det gula fältet i figur 1), extraherar länkar som är synliga från dessa startsidor och beslutar vilka länkar de ska följa härnäst baserat på vissa förbestämda kriterier. Härifrån laddas den eftersökta informationen ned och tolkas (representerat som ”Nedladdare och tolkare” i det orange fältet i figur 1) för att sedan lagras i en databas (representerat som ”Databas och lagring” i det gröna fältet i figur 1). De webbsidor som dessa länkar hänvisar till laddas ner och de som uppfyller de särskilda relevanskriterierna sparas i en kö på en lokal förvaringsplats (representerat som ”kö” i det blåa fältet i figur 1). Webbspindeln fortsätter att besöka webbsidor tills dess att ett särskilt antal sidor blivit besökta (representerat som ”Planerare” i det rosa fältet i figur 1) eller tills de lokala resurserna, så som lagringsutrymmet där den utvunna informationen sparats, inte längre räcker till.



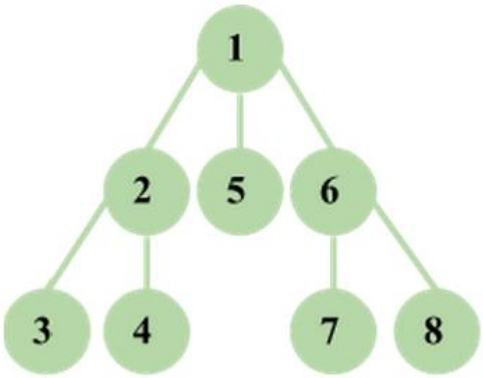
Figur 1: Illustration av webbspindelns arkitektur.

2.8 Ett urval av olika typer av webbspindlar

Det finns olika typer av webbspindlar som lämpar sig olika väl beroende på hur en given webbsida är uppbyggd och vilken typ av information som efterfrågas. Nedan följer ett urval av olika webbspindlar.

2.8.1 Djup-först-spindel

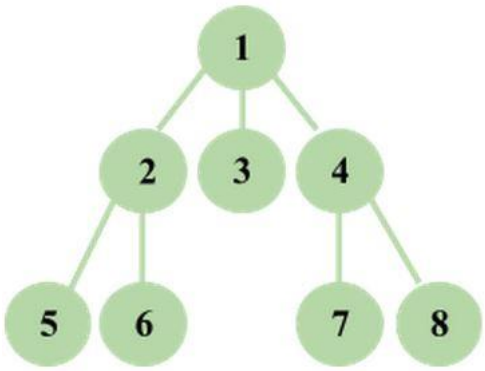
Melhorn & Sanders (2008) beskriver att en spindel av denna typ använder en djup-först-sökning när den söker igenom en webbsidas länkar, som börjar söka utifrån roten av trädet och jämför sedan de yttersta noderna i trädet hela vägen till botten innan den går vidare till en parallell nod. Den går alltså djupt ner i trädet på en sida först innan den går vidare. De förklarar även att en djup-först-spindel är en relativt oförsiktig spindel; när den finner en ny nod så går den omedelbart vidare och utforskar utifrån den. Den går tillbaka till en tidigare utforskad nod endast om den inte har några andra alternativ. Även om djup-först-sökning leder till ett generellt sett ganska obalanserat träd jämfört med exempelvis en bredd-först-sökning, gör kombinationen av ivrig utforskning och perfekt minne av det den sett, en djup-först-spindel väldigt användbar (Mehlhorn & Sanders, 2008).



Figur 2: Illustration av ordningen vid en djup-först-sökning.

2.8.2 Bredd-först-spindel

Spindeln börjar med en grupp bestående av ett fåtal webbsidor och undersöker sedan andra webbsidor genom att följa länkar i följden av en bredd-först-sökning (Ahuja et al., 2014). En bredd-först-sökning börjar från roten av trädet, jämför alla noderna i samma nivå och går sedan vidare och jämför noderna på nivån nedanför och så vidare (Beamer et al., 2013). Bredd-först-sökning kan, jämfört med djup-först-sökningen, istället ses som en försiktig, konservativ strategi för systematisk utforskning som tittar på kända saker innan den går vidare till outforskade miljöer (Mehlhorn & Sanders, 2008).



Figur 3: Illustration av ordningen vid en bredd-först-sökning.

2.8.3 Inkrementell spindel

Enligt Ahuja et al. (2014) uppdaterar en inkrementell spindel den existerande gruppen av webbsidor istället för att starta om genomsökningen från gång till gång. Dessa spindlars metod involverar sätt att bestämma om en sida har förändrats sedan förra gången den var genomsökt av spindeln. Spindeln består av en anpassningsbar metod där data från föregående genomsökningar bestämmer vilka sidor som ska genomsökas och därför blir materialet färskt och resulterar i låg toppbelastning. Denna typ av spindel lämpar sig när antalet källor som spindeln ska söka igenom är stort.

2.8.4 Gömd spindel

Eftersom mycket av den data som finns synlig på webben befinner sig i databaser måste denna data hämtas med hjälp av gömda webbspindlar genom att dessa anropar databasen med rätt typ av frågor eller fyller i formulär på webben som en skenanvändare (Ahuja et al., 2014). Andra spindlar söker endast genom sådana webbsidor som kan nås genom att följa länkar och på så sätt ignorerar de sidor och formulär som kräver viss befogenhet eller tidigare registrering. Dessa gömda spindlar tar sig alltså längre in bland svårtillgänglig information (Ahuja et al., 2014).

2.9 Kritik mot webbspindlar

I och med att webbspindlar anses vara kraftfulla finns också en del etiska aspekter och kritik som bör tas i beaktande. Det finns olika typer av webbspindlar som används till olika ändamål, av olika typer av användare. Hela spektrat, från sökmotorer som söker av hela webben ner till små program utvecklade av hobbyprogrammerare som laddar ner ett mycket begränsat dataset, förekommer i webbspindelsammanhang (Thelwall, 2006). Thelwall menar att det är viktigt att belysa att alla webbspindlar står under sin utvecklares kontroll. Antingen bestämmer utvecklaren att spindeln ska ladda ner hela webben, någon datapunkt här och där, eller allt som oftast någonting däremellan. Det finns också aspekter

som ligger utanför utvecklarens kontroll, nämligen hur sidorna ser ut som spindeln arbetar på, hur stor kapacitet dessa sidor har och hur stor bandbredden är som påverkar maxfarten som sidorna kan laddas ner med.

Thelwall (2006) argumenterar för att det finns fyra problem som webbspindlar kan skapa för samhället eller individen i form av överbelastning, kostnad, integritet och upphovsrätt. Med överbelastning menar Thelwall att om en webbsidas server är upptagen med att svara till webbspindlars begäran kan den vara långsam på att svara till andra användare som då upplever sidan som långsam och dålig. Här kan dock designen av webbsidan underlätta en del för att minska risken för så kallade spindelfällor, där webbspindlar trasslar in sig i ett stort antal länkar som leder till liknande sidor om och om igen. Med kostnad menar Thelwall att webbspindeln kan orsaka webbsidornas ägare stora kostnader. Ägare av webbdomäner har olika sätt att ta betalt för de som använder sidor hos dem, och ett vanligt förekommande sätt är att ta betalt för hur mycket bandbredd som används, och att det kan kosta mycket om en viss gräns överskrids. Gällande integritet råder det delade meningar. Vissa argument menar att webbspindlar har fritt tillträde eftersom allting på internet är inom den publika domänen. Dock finns en del motargument. Spindlar kan inskränka integriteten då de inhämtar stora mängder information från ett stort antal användare utan deras tillåtelse, exempelvis samlar in e-postadresser för utskick av spam. Slutligen kan upphovsrätt diskuteras. Webbspindlar gör till synes något olagligt: de skapar permanenta kopior av upphovsrättskyddat material utan ägarens tillåtelse (Thelwall, 2006). I slutändan är det viktigt att den som skapar webbspindeln inte utvecklar ett program som överträder upphovsrättsliga lagar eller inskränker andras integritet.

2.10 Användarförväntningar

Bachrach och McKean (2014) menar att användare idag förväntar sig att kunna finna vad de är ute efter snabbt och effektivt, eftersom det idag finns förutsättningar att söka efter information på ett annat sätt än vad som tidigare varit möjligt. Vidare hävdar de att användares förtroende för företag minskat, medan förtroendet för andra användare ökat. Den stora tillgången till information gör att användare har större förutsättningar att granska varor och tjänster kritiskt än vad som tidigare varit möjligt, eftersom det ofta går att ta reda på hur tidigare användare upplevt dem (Bachrach & McKean, 2014).

Att konsumenter själva kan ta reda på information om vad de vill påverkar också hur mottagliga de är av traditionell reklam (Bachrach & McKean, 2014). Allt fler konsumenter blockerar så mycket direktreklam som möjligt både i sin e-post och i den fysiska brevlådan. Författarna argumenterar därför för de potentiella vinster som finns för företag i att sträva efter en transparent verksamhet och dela information frikostigt, eftersom detta gör det lättare för konsumenter att få tillgång till erbjudanden i den mån som de själva vill (Bachrach & McKean, 2014).

3. Metod

I kapitlet nedan kommer studiens metodval behandlas och motiveras. Inledningsvis introduceras flermetodsforskning med denna studies metodbeståndsdelar i form av genomförda intervjuer och enkätstudie. Metodkapitlet avslutas sedan med en genomgång av valen kring utformningen av prototypen med val av webbspindel, databas och grafiskt interface.

3.1 Flermetodsforskning

Eftersom studiens syfte är att undersöka den eventuella diskrepans som råder mellan vilken information som webbspindlar kan hämta och vilken information som användare efterfrågar krävs metod som fångar in relevant data på vilken prototypen kan baseras. Studien kommer därför bestå av både kvalitativa och kvantitativa metoder, i form av intervjuer och enkäter, i så kallad flermetodsforskning. Flermetodsforskning har valts eftersom metoderna kompletterar varandra och väger upp för varandras bevisade svagheter så som intervjumetodens begränsning i att intervjuare måste tolka respondentens svar och enkätmetodens begränsning i att respondenter får förbestämda svarsalternativ. Flermetodsforskning skapar en mer komplett studie i form av både djup och bredd (Bryman, 2011).

3.1.1 Intervjuer

Intervjuer har genomförts med personer på ÅF i Solna för att få en djupare bild av hur behovet kring en eventkalender ser ut. Intervjuerna avsågs underlätta utformningen av enkätstudien för att kunna ställa så effektiva frågor som möjligt. Intervjuerna med dessa personer gjordes också för att få förståelse för hur en prototyp av programvaran kan utformas och implementeras, och vilka typer av evenemang som är mest relevanta för dessa personers respektive verksamhetsområden.

För denna studie har semistrukturerade intervjuer valts, vilket enligt Saunders et al. (2015) innebär att frågor på förhand skrivs ner men att intervjun inte är bunden till den strukturen. Istället ges respondenten möjligheten att reflektera fritt kring frågorna, vilket innebär att intervjuerna kan ge olika utfall trots att materialet de baseras på är detsamma. Bryman och Bell (2013) beskriver det som fördelaktigt med semistrukturerade intervjuer när respondentens åsikt är relevant och inte bara deras områdesexpertis.

Intervjuer genomfördes i sju fall av åtta i personliga möten eftersom det ger större utrymme för att tolka kontexten i vilket påståenden sägs, något som kan vara en felkälla vid exempelvis intervjuer över telefon (Saunders et al., 2015). En av intervjuerna genomfördes via Skype på grund av inställt möte av respondenten men Skypeintervjun anses i sammanhanget vara likvärdig ett personligt möte.

Inför intervjuerna genomfördes förarbete. I detta förarbete sammanställdes genomtänkta frågor av öppen och semistrukturerad karaktär för att få ut så mycket som möjligt av respondenternas egna tankar, erfarenheter och åsikter. Frågorna delades in i underrubriker för att underlätta för intervjuarna att behålla struktur i intervjuerna, men även för att respondenterna skulle uppleva intervjuerna som genomtänkta och kunna besvara frågorna i en naturlig ordning. Under utformningen av frågorna togs även hänsyn till att respondenternas svar kunde komma att spegla personliga åsikter, vilket i denna intervjustudie var att föredra. Frågorna utformades även med tanken att respondenternas svar senare skulle vara relativt enkla att utvärdera och till största möjliga mån vara tydliga för att undvika missförstånd och oklarheter.

Efter genomförda intervjuer transkriberades allt material som sedan sammanställdes och delades in i underrubriker. Detta gjordes främst för att skapa en överblick av vad som sagts och för att säkerställa att allt material kommit med. Författarna anser att noggrann genomarbetning av intervjuerna i närtid minskar risken för missad information och att delar glöms bort.

3.1.1.1 Urval av respondenter

Intervjurespondenterna är personer i olika roller inom ÅF och dess tre huvuddivisioner.

Tabell 1. Sammanställning av intervjurespondenter.

Namn	Avdelning	Position	Datum
Respondent 1	Technology	Sektionschef	2016-02-09
Respondent 2	Technology	HR-chef	2016-02-11
Respondent 3	Technology	Konsult	2016-02-16
Respondent 4	Technology	Försäljningschef	2016-02-17
Respondent 5	Industry	Projektledare	2016-02-18
Respondent 6	Industry	Sektionschef	2016-02-22
Respondent 7	Infrastructure	Utvecklingschef	2016-03-07
Respondent 8	Infrastructure	Konsult	2016-03-08

Respondent 1 är sektionschef för en av avdelningarna på ÅF Technology i Solna och har cirka 10 konsulter under sig. Arbetsuppgifterna för denne person består främst av tre delar i form av personal, rekrytering och sälj. Denne person intervjuades för att få en bild av hur eventkalendern skulle kunna användas av chefer inom divisionen Technology på ÅF.

Respondent 2 jobbar som chef inom HR på ÅF Technology i Solna. Bland arbetsuppgifterna återfinns bland annat ansvar för personal, rekrytering och kompetensförsörjning. Respondent 2 intervjuades för att inkludera hur evenemang och en eventkalender skulle kunna kopplas till kompetensutveckling inom företaget.

Respondent 3 arbetar som konsult på ÅF:s in-house byrå för IT på ÅF:s Solnakontor. Denna byrå har framförallt hand om olika interna portaler såsom intranätet och olika HR-system. Respondent 3 intervjuades för att inkludera en konsults perspektiv gällande synen på evenemang.

Respondent 4 arbetar inom divisionen Technology på Solnakontoret. Där innehar denne en chefsposition och är ansvarig för försäljning och affärsutveckling inom divisionen. Respondent 4 inkluderades eftersom åsikterna och erfarenheterna kan återspegla en chefs syn på evenemang i marknadsföringssyfte för ÅF.

Respondent 5 arbetar på divisionen Industry i Solna och Nynäshamn. Respondent 5 arbetar som konsult till 85% och resten av tiden som projektledare där denne bygger och underhåller system. Respondent 5 intervjuades för att bidra till konsultperspektivet från ytterligare en avdelning samt för att bidra med en projektledares perspektiv.

Respondent 6 arbetar som sektionschef för en avdelning inom divisionen Industry. Respondent 6 är framförallt stationerad i Nynäshamn men jobbar delvis i Solna. Inom avdelningen arbetar denne med produktionsstyrande system till industrin där det bland annat hanteras planering och kvalitetssäkring på industrianläggningar. Respondent 6 intervjuades för att komplettera chefsperspektivet med ytterligare en sektionschef från en annan division.

Respondent 7 tillhör divisionen Infrastructure, sitter på kontoret i Solna och innehar en roll som innebär stöttning av ÅF:s olika affärsområden och dess chefer i affärsutveckling med fokus på förvärv. Respondent 7 inkluderades för att bidra med erfarenhet av branschen och erfarenheter kring att representera ÅF på olika evenemang i marknadsföringssyfte.

Respondent 8 arbetar som systemutvecklare för järnvägssystem på ÅF Infrastructure och har inget personalansvar utan ansvarar istället för systemdrift och optimering. Respondent 8 bidrar i och med sin medverkan till ännu en konsults erfarenheter kring evenemang.

3.1.2 Enkäter

För att komma i kontakt med fler tilltänkta användare av programvaran valdes en enkätstudie som lämplig metod. Syftet med enkäten var att utreda de krav respondenterna har på en tjänst för att de ska vara beredda att använda den. Saunders et al. (2015) menar att enkätstudier är effektiva vid jämförelser och sammanställning av svar från många respondenter. Genom att utforma icke ledande frågor och få svar från tillräckligt stort urval med hög svarsfrekvens kan slutsatser dras från enkäten då svaren kan anses vara representativa för populationen (Saunders et al., 2015).

Efter genomförda intervjuer hade ett underlag skapats och utifrån denna information utformades ett antal enkätfrågor. Dessa frågor togs fram för att reflektera ämnen som främst gällde användarnas förväntningar på eventkalendern. Enkätfrågorna innehöll olika svarsalternativ eftersom detta skulle göra enkäten enklare att genomföra för respondenterna. Nackdelen med att bistå med svarsalternativ kan vara att respondenterna känner sig manade att svara i viss riktning eller att de leds in på ett visst tankespår eftersom svaren är standardiserade (Brace, 2008). Fördelen med standardiserade svarsalternativ är att svarsfrekvensen går upp om det är lätt och går fort för respondenten att svara på frågorna (Brace, 2008). I denna studie gjordes valet att det viktigaste är att nå många på företaget och att svarsfrekvensen blir så hög som möjligt för att få en representativ bild av förväntningarna. Är svarsfrekvensen hög ökar också studiens trovärdighet och ger styrka åt resultaten (Brace, 2008). Brace menar vidare att frågor måste vara entydiga för att kunna tolkas. På grund av detta gjordes ett utkast på enkäten som reviderades flera gånger. Dessutom genomfördes en pilotundersökning där några anställda valdes ut och fick testa enkäten innan den distribuerades till övriga.

3.1.2.1. Genomförande av enkätundersökning

För att erhålla hög svarsfrekvens delades enkäterna ut i pappersformat. Pappersenkäter antas ge en mer direkt kontakt med respondenterna vilket kan ge högre svarsfrekvens än om de delas ut i elektronisk form, exempelvis genom e-post (Kongsved et al., 2007). Enkäterna delades ut på ÅFs kontor i Solna. Hit hör cirka 1500 konsulter men där majoriteten sitter på uppdrag ute hos kund. Respondenter valdes ut slumpmässigt vid sina arbetsstationer och hade möjlighet att avstå om de önskade. Enkäten innehöll all nödvändig information för att kunna besvaras, vilket innebar att alla respondenter fick samma förutsättningar att svara på frågorna.

Enkätundersökningen utfördes vid två tillfällen då enkäter delades ut på kontorets samtliga våningsplan. Enkäterna delades främst ut till de anställda som satt vid sina arbetsstationer och som inte var upptagna i möte eller telefonsamtal. Eftersom alla anställda inte sitter vid sina arbetsstationer hela tiden ansågs urvalet vara slumpmässigt. För att få högre svarsfrekvens gavs respondenterna cirka två timmar att genomföra frågeformuläret innan de samlades in. De respondenter som inte ville delta i undersökningen kunde avstå, vilket noterades för att senare föras in i sammanställningen.

Om en respondent gav ofullständiga svar eller inte följde de givna instruktionerna kasserades svaret och noterades som icke deltagande i sammanställningen.

Efter att enkätsvaren samlats in skapades en digital enkät med hjälp av webbverktyget Google Forms där alla svar fördes in manuellt. Detta gjordes för att Google Forms sedan sammanställer enkätsvaren i lättbegripliga diagram med tydliga etiketter och sammanfattningar (se bilaga C).

3.2 Utformning av prototyp

3.2.1 Val av spindel

För att avgöra vilken spindel som var lämplig i denna typ av studie fanns ett antal faktorer som togs i beaktande. Dessa faktorer fungerade som avgränsningar när valet av lämplig extraheringsmetod gjordes. Informationen som söktes i denna studie var öppen, det behövdes med andra ord inte en spindel som klarade av att hantera formulär för att nå särskilda sidor. Detta faktum innebar att tidigare nämnda gömda spindlar kunde bortses från. Antalet sidor som i detta projekt genomsöktes var relativt litet, vilket medförde att behovet av en inkrementell spindel också var litet. Kvar återstod spindlar som var antingen djup-först eller bredd-först. Urvalet av denna typ av spindlar är stort, så andra faktorer fick spela in vid valet av teknik.

De två viktigaste faktorer som fick avgöra vilken teknik som skulle användas i denna studie var vilket språk som spindeln skrevs i och vilken data som eftersöktes. Eftersom det endast var textbaserad information som var relevant för denna studie, och att utvecklingstiden var begränsad valdes programmeringsspråket Python. Python är ett lättviktigt språk, vilket innebär att det är relativt enkelt att förstå för utomstående. Dessutom krävs inte så mycket kod för att utföra vanliga kommandon.

Även inom Python finns ett urval av spindlar att välja mellan. Valet gjordes därför med hänsyn till vilken teknik som är vanligast och har störst användarbas. Detta medför en enklare felsökningsprocess och en mer utförlig dokumentation av programvaran. Valet av program för att utveckla spindeln för denna studie föll således på ett program som heter Scrapy. I sitt standardutförande använder Scrapy en djup-först-algoritm för att söka igenom webbsidor och länkar, vilket i detta fall lämpar sig väl. Anledningen till detta diskuteras närmare i avsnitt 5.2.

3.2.2 Databasstruktur

För att kunna lagra och hantera extraherad data måste en databas användas. Det finns en rad databaser på marknaden, där den dominerande typen under många år varit relationsdatabaser, exempelvis MySQL (Győrödi et al., 2015). De senaste åren har dock flera stora företag gått ifrån denna typ av databas till förmån av så kallade icke-

relationsbaserade databaser, där den mest använda är MongoDB. Icke-relationsbaserade databaser hanterar stora datamängder snabbare och kräver inte, till skillnad från relationsbaserade databaser, en på förhand bestämd struktur för att lagra data, vilket gör dem mer flexibla (Győrödi et al., 2015).

För denna studie valdes MongoDB av ytterligare anledningar. Den extraherade datan sparades enligt en struktur som gjorde relationsbaserade databaser överflödiga. Dessutom är Scrapy väl anpassat efter att fungera tillsammans med MongoDB (Scrapy). Detta innebar en tidsbesparing för studien och ansågs vara av stor vikt vid valet av databas.

3.2.3 Grafiskt interface

För att koppla den insamlade informationen som ligger i databasen till webbsidans interface användes PHP (Hypertext Preprocessor), vilket är den vanligaste tekniken som används för att skapa dynamiska hemsidor och koppla en databas till en webbsida (Powers, 2010). Den grafiska webbsida som sedan presenterar den extraherade informationen är skriven i HTML och CSS, vilka är verktyg för att visa information i en webbläsare.

HTML står för Hypertext Markup Language och är det språk nästan alla sidor på webben är uppbyggd av (Schultz & Cook, 2007). HTML är ett relativt lätt språk och de grundläggande reglerna är lätta att lära sig och använda. HTML är inte komplett som det är utan fungerar som bäst tillsammans med CSS. CSS står för Cascading Style Sheets och är ett språk som beskriver hur webbdokument presenteras visuellt (Schultz & Cook, 2007). CSS är kraftfullt, flexibelt och komplext och kan presentera all typ av information på ett tydligt sätt i rätt kombination med HTML (Schultz & Cook, 2007). För att ytterligare spara tid och utnyttja redan välfungerande mallar valde författarna till denna studie att använda Bootstrap, vilket är en open-sourcelösning baserad på CSS för att lättare kunna bygga välfungerande grafik och webbsidor (Bootstrap).

4. Resultat: användarbehov och förväntningar

I detta kapitel presenteras resultaten från intervju- och enkätstudien. Dessa resultat berör den del av syftet som gäller användarnas efterfrågningar. Inledningsvis presenteras intervjuresultaten under avsnitt som berör evenemangs relevans, hur anställda på ÅF ser på vikten av att medverka i evenemang och vad det finns för önskad funktionalitet kring en eventkalender. Efter detta presenteras resultaten från enkätundersökningen.

4.1 Evenemangs relevans

Inom företag är det viktigt med kompetensutveckling och framförallt inom ett så stort företag som ÅF är det viktigt att ha tydliga sätt att dela kunskap sinsemellan, menar Respondent 2. Respondent 2 som innehar en chefsposition förklarar att ÅF har stora möjligheter att anställa kompetenta konsulter men att dessa också måste utvecklas över tid för att kunna fortsätta vara de bästa inom sina respektive branscher och kunskapsområden. Kompetensutveckling och kunskapsutbyte mellan kollegor är något som även gör att medarbetare känner att de utvecklas och känner ett ytterligare värde i att gå till jobbet (Intervju, Respondent 2).

Respondent 4 som även denne innehar en chefsposition beskriver att det på ÅF anordnas många evenemang och att det finns stora möjligheter att dela kunskap sinsemellan men att det stora hindret ligger i att anställda ofta inte vet vad som pågår runt omkring sig själva och sina avdelningar. Är den anställda dessutom konsult och sitter största delen ute hos kund är det viktigt att kunna bibehålla kontakt med sina kollegor på ÅF (Intervju, Respondent 4). Respondent 4 menar att en stor anledning till att evenemang är viktiga är att de anställda knyter kontakter och bibehåller relationer till sina kollegor som de kanske inte träffar så ofta annars.

Respondent 5 är konsult på ÅF och lyfter inspiration som en viktig aspekt av att gå på evenemang. Denne tycker evenemang både är roliga och viktiga samt bidrar till att konsulter håller sig uppdaterade och inspirerade till att lära mer (Intervju, Respondent 5). Respondent 3, som också jobbar som konsult, delar samma synsätt som Respondent 5. Denne belyser även att kulturen att gå på evenemang för utvecklare är mycket utbredd inom andra företag och att det finns stora möjligheter till kompetensutveckling för konsulter på detta sätt även inom ÅF om anställda bara finner rätt information, kommer in i vanan att lära från evenemang och tar sig tiden att gå på dem (Intervju, Respondent 3). Respondent 6, som är sektionschef, beskriver att själva innehållet i sig inte alltid är det viktigaste utan att det mest intressanta istället är att kunna bygga nätverk inom sina respektive branscher.

Respondent 5 tror att det i slutändan mest är chefer som kommer att gå på evenemang och utnyttja ett verktyg som belyser vilka evenemang som pågår. Respondent 5 förklarar

vidare att personer i chefspositioner är mer intresserade av att ÅF ska synas på mässor och stora branschdagar jämfört med vad konsulterna är, vilket även Respondent 1, 3 och 4 instämmer med. Respondent 1 menar att chefer på ÅF eftersträvar att ÅF ska vara ett företag som folk känner till som att vara i framkant och därför vill ÅF vara närvarande vid stora evenemang för många deltagare. ÅF vill gärna delta vid mässor som riktar sig mot studenter för att säkerställa att duktiga personer söker sig till företaget (Intervju, Respondent 1). Respondent 4 hävdar att för folk i högre positioner är trendspaning något mycket centralt. Denne menar att trendspaning är viktigt för alla delar inom ÅF eftersom stora trender påverkar alla kunder ÅF har, inom alla branscher och inom alla ämnesområden (Intervju, Respondent 4). Respondent 8 arbetar som systemutvecklare och blir sällan inbjuden till evenemang men hade gärna velat delta om möjligheten gavs (Intervju, Respondent 8). Detta resonemang delar även Respondent 3 som menar att tillfälle måste ges från cheferna för att konsulterna ska kunna komma in i vanan att söka upp och gå på evenemang (Intervju, Respondent 3).

4.2 Traditioner kring evenemang på ÅF

Även om ambitionen från ÅF är att hålla sin personal i utvecklingens framkant är deltagande i externa evenemang mycket varierat, menar Respondent 5. Orsaken är en kombination av faktorer; bland annat frånvaro av tydliga ramverk för hur uppföljning ska gå till efter att någon deltagit i en kompetensutvecklande aktivitet, och att det är svårt att få tillgång till information om evenemang som är relevanta för ens egna arbetsuppgifter (Intervju, Respondent 5).

Respondent 2 beskriver att det är svårt att avgöra hur kompetensutveckling ska gå till när många konsulter har debiteringskrav. Det innebär att en kompetensfrämjande aktivitet måste utgöra en investering som väger tyngre än att under samma tid fakturera mot kund (Intervju, Respondent 2). Respondent 2 ser dock en möjlighet att bättre utnyttja tiden hos de konsulter som är mellan uppdrag. Istället för att vänta på att nya uppdrag ska börja menar denne att dessa hade varit utmärkta tillfällen att utveckla sin kompetens. Tyvärr finns det i dagsläget ingen plattform för att ta reda på vilka evenemang som eventuellt skulle vara relevanta vilket leder till att vissa inte utnyttjar sin tid optimalt, något som även Respondent 7 och 8 bekräftar (Intervju, Respondent 2, 7 och 8).

Flera personer som intervjuats i denna studie uppger att de idag använder tjänster som Google för att ta reda på vad som pågår inom sina yrkesområden, men att de då måste veta vad de eftersöker på förhand gör det ineffektivt (Intervju, Respondent 1, 3, 4 och 7). Andra sätt att hålla sig uppdaterad är genom tips från kollegor eller inbjudningar från arrangörer. Båda dessa kräver dock att någon typ av intresse har visats sedan tidigare för att kunna få ta del av informationen. Det krävs med andra ord en aktiv strävan efter att delta i evenemang för att fler möjligheter ska visa sig (Intervju, Respondent 4, 5 och 6).

I dagsläget tycks det vara evenemang som ÅF själva anordnar som folk är mest medvetna om. Respondent 3 menar att ett stort företag som ÅF har större möjlighet att anordna organiserade evenemang för att kompetensutveckla stora grupper åt gången, men att många går miste om de mindre, informella evenemangen där kreativitet och spontanitet står i centrum (Intervju, Respondent 3).

Chefer inom ÅF som intervjuats för denna studie har olika syn på både anställdas möjlighet att gå på olika kompetensutvecklande evenemang och på hur denna kompetens ska föras vidare inom organisationen. Respondent 1 och 6, som bägge är sektionschefer, menar att de gärna skickar anställda på externa evenemang om det är välmotiverat och passar in i de anställdas roll och schema (Intervju, Respondent 1 och 6). Spridning av kunskap från dessa evenemang, menar de, kan ske på de interna sektionssammanträden som inträffar ungefär en gång per kvartal. Respondent 2 och 4 uppger att de föredrar att få en skriven rapport från den anställde som deltagit på evenemang, så att informationen blir tillgänglig för alla inom ÅF. Konsulterna som intervjuats (Respondent 3,5 och 8) lyfter dock en viss problematik kring detta. Dels upplevs det som att det inte finns tid att lägga på att sammanställa en rapport om arbetstid redan har tillbringats på externa evenemang. Vidare kan kravet att skriva en rapport eller berätta om sina erfarenheter i ett öppet forum avskräcka folk från att faktiskt delta i evenemang (Intervju, Respondent 3). Respondent 8 beskriver att avdelningen denne jobbar på inte prioriterar att anställda ska delta på evenemang trots att det finns ett gemensamt intresse för att få upplysningar om detta (Intervju, Respondent 8).

4.3 Önskad funktionalitet

Respondent 1 förklarar att i egenskap av sektionschef, i en eventkalender helst skulle vilja se vilken typ av evenemang det är som presenteras i kalendern. Om det är ett evenemang företaget är med på för att synas, som exempelvis en studentmässa, eller om det är ett evenemang som är menat att fördjupa anställdas kunskaper inom ett specifikt ämne, exempelvis ett seminarium (Intervju, Respondent 1). Respondent 1, 4 och 6 är alla mest intresserade av evenemang som är arrangerade av stora företag och organisationer, exempelvis branschdagar dit de största talarna kommer och där de nyaste teknikrönen avhandlas. De påpekar även att de är mycket intresserade av studentdagar och evenemang arrangerade av Arbetsförmedlingen eftersom ÅF där kan finna nya potentiella medarbetare (Respondent 1, 4 och 6). Även Respondent 3 intygar att intresset främst ligger i evenemang arrangerade av stora företag inom IT som exempelvis Ebay och Google. För denne är det just arrangör, typ av evenemang och datum som är viktigast medan priset beskrivs som mindre viktigt eftersom det alltid är chefen som bestämmer om priset är godkänt eller inte (Intervju, Respondent 3).

Det är av stor vikt att ett verktyg som en eventkalender på sikt kan inkluderas i ÅF:s intranät menar Respondent 3. Denne ser gärna att evenemangen kan illustreras i ett

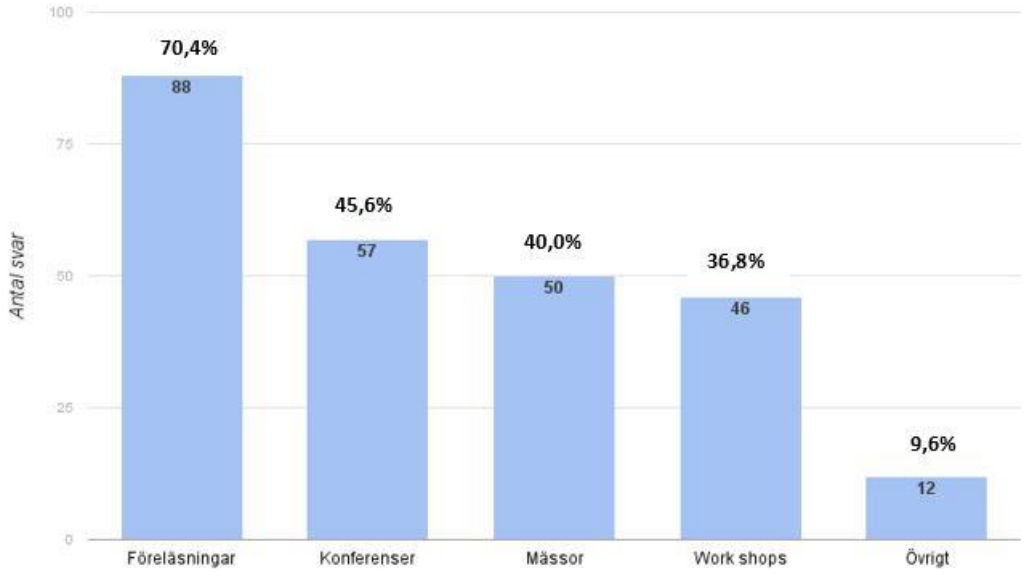
nyhetsflöde på intranätets förstasida. Vidare menar Respondent 3 att det hade varit uppskattat att få personligt anpassade förslag på evenemang presenterade för sig, liksom många andra moderna webbtjänster erbjuder (Intervju, Respondent 3). Respondent 3 skulle också gärna se vilka evenemang vänner och kollegor ska gå på och tror att en skulle bli mer frekvent deltagare om denne fick gå tillsammans med sina kollegor. Då lockas också de kollegor som sitter på heltid ute hos kund eftersom de kan träffa sina kollegor från ÅF som de annars kan tappa kontakten med, något som även Respondent 7 och 8 instämmer i (Intervju, Respondent 3, 7 och 8)

Respondent 4 beskriver att denne vill se en eventkalender kopplad till sin egen kalender där det ska vara möjligt att prenumerera på vissa typer av evenemang med hjälp av en filterfunktion. Respondent 4 säger också att det skulle vara önskvärt att kunna få påminnelser längre fram i tiden gällande nya evenemang som kommer upp, men även påminnelser gällande när sista anmälningsdag närmar sig (Intervju, Respondent 4). Denne lyfter det faktum att tjänsten måste vara lättanvänd och tydlig för att medarbetarna på ÅF verkligen ska ta del av den och inkludera sökningar i eventkalendern i sina dagliga vanor.

4.4 Resultat från enkätundersökning

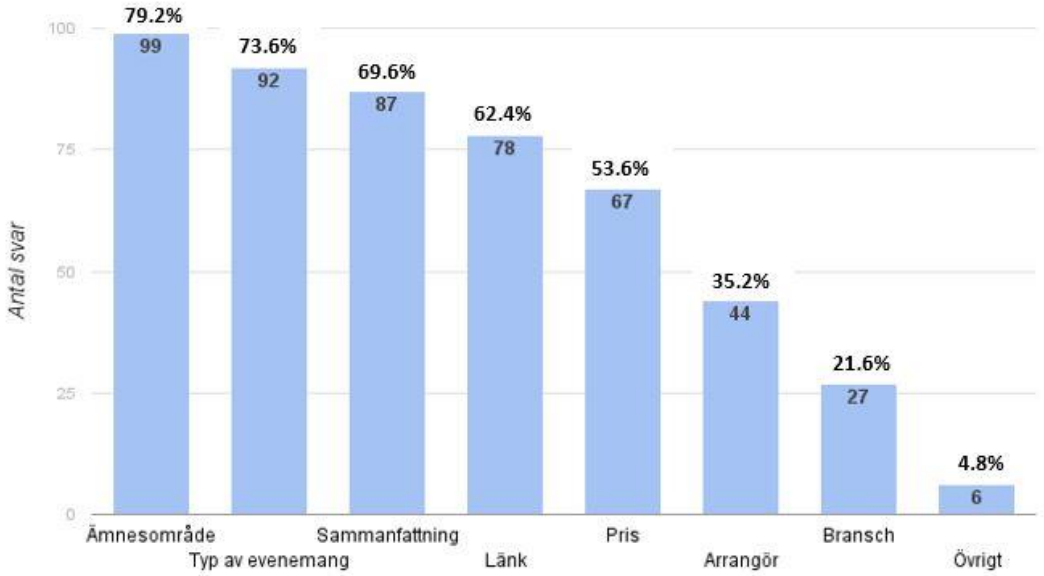
I och med enkätundersökningen har författarna till denna studie kommit i kontakt med 125 medarbetare på ÅF i Solna. Sju personer valde att inte delta i undersökningen, vilket innebar en svarsfrekvens på cirka 94% av de tillfrågade. Av enkätstudien framgick att respondenternas ålder är relativt jämnt fördelad, samt att över 80% av de tillfrågade tillbringar minst 75% av sin arbetstid på ÅF:s kontor i Solna. Se bilaga C för en fullständig sammanfattning av enkätsvaren.

Respondenterna ombads att välja två typer av evenemang som de ansåg vara mest relevanta att delta i. 70% uppgav att de är mest intresserade av att gå på föreläsningar. Mässor, work shops och konferenser hade mellan 40% och 45% vardera. Några andra typer av evenemang som deltagarna i undersökningen föreslog under alternativet “Övrigt” var bland annat kurser och middagar. Även på frågan om vilka evenemang deltagarna skulle vilja få presenterade för sig i en kalender var föreläsningar det mest populära alternativet. Undersökningen visar dock att respondenterna vill få fler typer av evenemang presenterade för sig i eventkalendern, även om de inte är lika intresserade av dem som de är av föreläsningar.



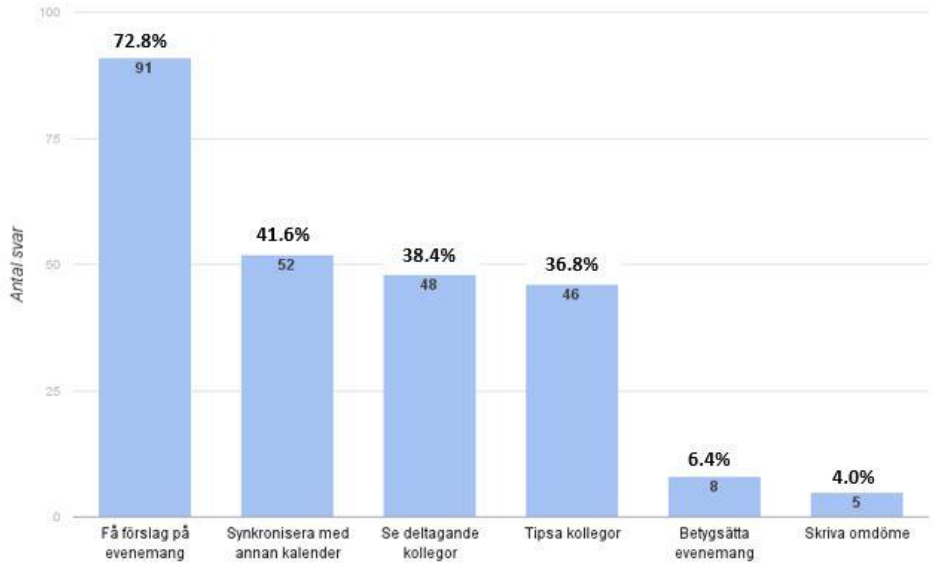
Figur 5: De populäraste typer av evenemang att gå på där respondenten fick välja två alternativ.

När respondenterna skulle svara på frågan om vilka fyra attribut som de ansåg vara viktigast att få presenterade för sig svarade cirka 80% att "Ämnesområde" är ett av de viktigaste attributen. Även "Typ av evenemang" och "Sammanfattande text om eventet" hade höga andelar runt 70%. Förslag som nämndes under "Övrigt" var framförallt datum som på förhand tagits bort som svarsalternativ eftersom det varit en självklarhet från början att datumattributet skulle finnas med i eventkalendern.



Figur 6: De viktigaste attributen i en eventkalender där respondenten fick välja fyra alternativ.

I enkätundersökningen fick respondenterna även svara på vilka två ytterligare funktioner de skulle vilja ha tillgång till i en eventkalender. Den funktion som de allra flesta valde (73%) var “Att kunna få förslag på passande evenemang för just dig”. Därefter kom alternativen “Att kunna se vilka kollegor som deltar i vilka evenemang”, “Att kunna synkronisera med annan kalender” och “Att kunna tipsa en kollega om ett evenemang” som vardera fick cirka 40% som valde de alternativen som en av två ytterligare funktioner att ha i kalendern.



Figur 7: Önskvärda funktioner i en eventkalender där respondenten fick välja två alternativ.

5. Resultat: prototyp

I detta kapitel presenteras prototypen som utvecklats för att kunna svara på den del av frågeställningen som gäller vad en webbspindel kan utvinna. Prototypen som utvecklats är gjord i syfte att visualisera hur väl ett relativt enkelt automatiserat program kan utvinna webbaserad information i förhållande till de krav som användarna ställer. Först görs en genomgång av resultatet från webbspindeln för att sedan presentera den bakomliggande tekniken, ett exempel kring hur extraheringsprocessen kan se ut och till sist göra en genomgång av den slutliga eventkalendern.

5.1 Resultat av webbspindeln

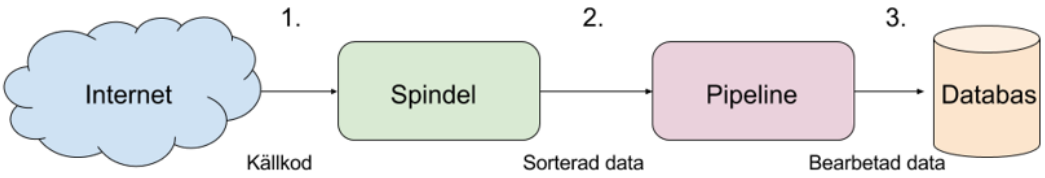
Målsättningen med webbspindeln var att extrahera så många relevanta evenemang som möjligt. På grund av de problem som diskuterats av bland andra Alvarez et al. (2008), nämligen den frånvaro av gemensam struktur som råder på internet, behövdes en del val göras. Vissa evenemang presenterades exempelvis i bildformat istället för i textformat. Detta omöjliggjorde extrahering av information med vald teknik, vilket gjorde att dessa evenemang fick bortses från. Andelen evenemang som presenterades på detta sätt var lyckligtvis liten vilket medförde att webbspindelns övergripande funktionsduglighet därmed inte berördes nämnvärt.

För denna studie utvecklades sex webbspindlar som extraherar information från lika många webbportaler. Antalet evenemang som finns tillgängliga i projektets databas varierar, men i regel lyckas de sex webbspindlarna extrahera och kategorisera mellan 110 och 120 evenemang som presenteras på projektets webbsida. Denna webbsida uppdateras en gång per dygn nattetid. För dessa drygt 110 samlade evenemang kunde webbspindeln sammanställa och presentera följande attribut: titel, datum, plats, sammanfattande text, nyckelord och slutligen en länk till evenemanget i fråga.

Varje webbspindel anpassades efter den webbsida den letade evenemang på. Det format som portalerna presenterade information på skiljde sig från fall till fall. Eftersom målsättningen med projektet var att sammanställa informationen på ett generiskt sätt krävdes både vissa anpassningar och avgränsningar. Vissa portaler erbjöd exempelvis inte en beskrivning av evenemanget, något som inte bara påverkade webbspindelns funktionalitet, utan även gjorde det svårare att kategorisera evenemanget under bearbetningsprocessen.

5.2 Tekniken bakom webbspindeln

Nedan beskrivs hela förloppet webbspindeln genomgår för att hämta och spara information enligt Scrapys dokumentation (Scrapy, 2016).



Figur 8. Förenklad illustration av Scrapys programstruktur.

1) När webbspindeln ska samla in data börjar den med att koppla upp sig mot den URL som den är programmerad att besöka. Från denna URL hämtar den hem webbsidans källkod. Spindeln genomsöker källkoden för att finna specifika elementmarkeringar. Den kan finna elementmarkeringar antingen genom dess unika namn eller genom dess placering i källkoden. Informationen den finner sparas i specifika objekt, i detta fall som ett evenemang. När all eftersökt information är sparad som attribut till varje evenemang skickas den sorterade datan till en så kallad pipeline.

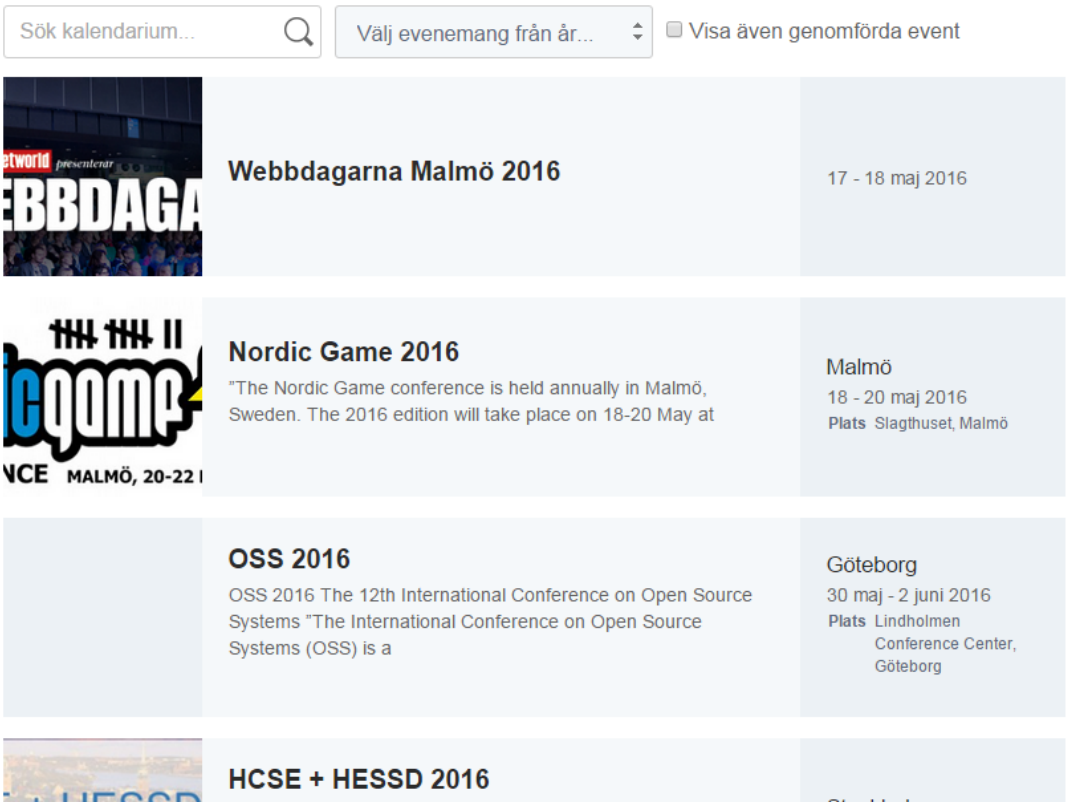
2) Pipelinen tar emot ett objekt med sorterad data. Där bearbetas datan ytterligare för att ge alla attribut generisk form, till exempel samma datumformat. Dessutom genomsöks datan för att tilldela varje evenemang ytterligare attribut som inte varit möjliga att extrahera, till exempel typ av evenemang och ämnesområde.

3) När all data har genomarbetats och är i önskat format skickas objektet för att sparas i en databas. När ett objekt sparats börjar processen om med nästa objekt. Objekt som är sparade i databasen kan senare visas på exempelvis en webbsida.

5.2.1 Exempel på extraheringsprocessen

I följande exempel visas hur extraheringsprocessen ser ut i praktiken. Detta är i förenklad form, där endast för exemplet relevanta komponenter av processen tas i hänsyn.

I figur 9 visas ett exempel på hur en portal kan se ut för en mänsklig användare, i detta fall en portal för evenemang i IT-branschen. I figuren syns de evenemang som är intressanta att extrahera. Värt att notera är att all relevant information inte finns tillgänglig i denna vy, exempelvis en mer ingående beskrivning av evenemangen. Webbspindeln måste således följa en länk till varje evenemang för att få tillgång till ytterligare information.



Figur 9: Exempel på evenemangsportal för IT-evenemang (swedsoft.se).

När programmet startas skickas domännamn och den URL vilken webbspindeln ska utgå ifrån. Domännamnet tvingar webbspindeln att inte följa länkar som leder till en annan domän. Dessutom är webbspindlarna i denna studie programmerade att inte följa länkar mer än ett steg. Det innebär att djup-först-sökning och bredd-först-sökning i detta fall är lika effektiva. Djup-först-sökning skulle vara bättre lämpat om webbspindeln följde länkar i flera steg. Detta på grund av att den följer länkar endast i syfte att få mer information om ett specifikt evenemang. Den extraherar alltså all information som berör ett evenemang innan den går vidare till nästa evenemang. En bredd-först-sökning hade frångått denna logiska ordning.

Nedan visas ett kort utdrag från programmeringskoden för en förenklad webbspindel. Fullständig kod går att finna i bilaga D.

```
11  class MySpider(CrawlSpider):
12      name          = "swedint"
13      allowed_domains = ["swedsoft.se"]
14      start_urls    = ["http://swedsoft.se/kalender/kalendarium/"]
```

Figur 10: Utdrag från en specialanpassad webbspindels kod.

Webbspindeln kopplar sedan upp sig mot webbportalen i fråga och genomsöker dess källkod. Nedan syns ett utdrag från portalens källkod, vilken beskriver det översta

evenemanget i figur 9. Varje enskilt evenemang ligger inom elementmarkeringen `<a class=”article-link”>`. I figuren visas alltså tre evenemang, ett vars innehåll är expanderat, det vill säga “1” i figuren. “2” och “3” motsvarar andra evenemangs kod. Webbspindeln är programmerad att behandla varje sådan elementmarkering den finner i källkoden som ett enskilt evenemang, och varje evenemang har i detta exempel identisk kodstruktur. Webbspindeln söker igenom koden efter på förhand definierade elementmarkeringar, i detta fall evenemangets titel, datum samt länken som den måste följa för att erhålla ytterligare information. Titeln finns i elementmarkeringen `<h1 class=”h2 entry-title”>`, datum i `<p class=”meta medium light”>` och länken i

`<a href=”http://swedsoft.se/event-ovrigt/webbdagarna-malmo-2016”>`.

Figur 11: Utdrag från portalens källkod (swedsoft.se).

Följande kod beskriver, på ett förenklat sätt, hur informationen extraheras. Först definieras en regel för vilka länkar som tillåts följas (rad 16-18 i figur 12). Webbspindeln är programmerad att inte följa länkar den finner direkt, utan vänta en bestämd tid. Detta är gjort för att inte belasta webbsidan i fråga med för många förfrågningar samtidigt.

```
16      rules = (
17          Rule(SgmlLinkExtractor(allow = (), restrict_xpaths=('//*[@id="main"]/div/div/')), callback="parse", follow = True),
18      )
```

Figur 12: Utdrag från webbspindelns kod som beskriver vilka länkar som ska följas.

Den gröna texten till höger i figuren definierar var i webbsidans källkod informationen återfinns med hjälp av en Xpath, det vill säga en viss plats i källkoden (Richards, 2006). På rad 23-27 i figur 13 extraheras tillgänglig information och sparas i attribut, exempelvis i “`item[‘title’]`”.

```
23              for div in response.xpath('//*[@id="main"]/div/div/a'):
24                  item = AfeventItem()
25                  item['title'] = div.xpath('//*[@class ="h2 entry-title"]/text()').extract()[i]
26                  item['location'] = div.xpath('//*[@id]/div[3]/h2/text()').extract()[i]
27                  item['date'] = div.xpath ('//*[@id]/div[3]/p/text()').extract()[i]
```

Figur 13: Utdrag från webbspindelns kod som visar de olika attributens position i källkoden.

Rad 27-30 i figur 14 beordrar spindeln att behålla den hittills extraherade informationen och följa länken för att extrahera ytterligare information.

```
27                  follow_url = div.xpath('//*[@id="main"]/div/div/a/@href').extract()
28                  request = Request(follow_url, callback = self.parse_second)
29                  request.meta['item'] = item
30                  yield request
```

Figur 14: Utdrag från webbspindelns kod som visar hur den behåller information mellan sidor.

Rad 32-35 i figur 15 exekveras när webbspindeln följt länken till en annan sida där resten av informationen finns. Där extraheras de sista attributen, i detta fall "`item['description']`". Slutligen har all relevant information extraherats och all information skickas för att bearbetas med kommandot "`yield item`". Efter det söker webbspindeln upp nästa elementmarkering som motsvarar ett evenemang och upprepar processen.

```
32      def parse_second(self, response):
33          item = response.meta['item']
34          item ['description'] = ''.join(response.xpath('//*[@id]/section/p//text()').extract())
35          yield item
```

Figur 15: Utdrag från webbspindelns kod som visar hur den samlar information från följd länk.

I figur 16 visas dessa moment tillsammans.

```
11  class MySpider(CrawlSpider):
12      name        = "swedint"
13      allowed_domains = ["swedsoft.se"]
14      start_urls  = ["http://swedsoft.se/kalender/kalendarium/"]
15
16      rules = (
17          Rule(SgmlLinkExtractor(allow = (), restrict_xpaths=('//*[@id="main"]/div/div/')), callback="parse", follow = True),
18      )
19
20      def parse(self, response):
21          for div in response.xpath('//*[@id="main"]/div/div/a'):
22              item = AfeventItem()
23              item['title'] = div.xpath('//*[@class ="h2 entry-title"]/text()').extract()
24              item['location'] = div.xpath('//*[@id]/div[3]/h2/text()').extract()
25              item['date'] = div.xpath ('//*[@id]/div[3]/p/text()').extract()
26
27              follow_url = div.xpath('//*[@id="main"]/div/div/a/@href').extract()
28              request = Request(follow_url, callback = self.parse_second)
29              request.meta['item'] = item
30              yield request
31
32      def parse_second(self, response):
33          item = response.meta['item']
34          item ['description'] = ''.join(response.xpath('//*[@id]/section/p//text()').extract())
35          yield item
```

Figur 16: Förenklat utdrag av webbspindelns extraheringsfunktion.

När ett evenemang med tillhörande attribut skapats som objekt skickas det till programmets pipeline. Det första som händer där är att det upprättas en anslutning till en databas, i detta fall MongoDB (rad 18-23 i figur 17).

```
18        connection = pymongo.MongoClient(
19            settings['MONGODB_SERVER'],
20            settings['MONGODB_PORT']
21            )
22        db = connection[settings['MONGODB_DB']]
23        self.collection = db[settings['MONGODB_COLLECTION']]
```

Figur 17: Utdrag från pipelinens kod angående kopplingen till databasen.

På rad 25-27 i figur 18 jämförs det aktuella evenemangets URL med redan lagrade evenemangs. Detta förhindrar att dubbletter lagras i databasen eftersom de inte sparas om de redan återfinns där.

```
25    def process_item(self, item, spider):
26        if self.collection.find_one({'url': item['url']}):
27            raise DropItem('Item already in DB')
```

Figur 18: Utdrag från pipelinens kod angående dubbletter i databasen.

På rad 29-37 i figur 19 jämförs evenemangets titel och beskrivning med en extern lista på cirka 1500 ämnesspecifika nyckelord. Dessa nyckelord tillhör ÅF:s interna kategoriseringar som är en del av deras kompetensnätverk. Återfinns ett visst nyckelord i informationen som hämtas från webbsidan tillskrivs evenemanget en “tagg” i form av ett ämnesområde. Slutligen lagras evenemanget som ett objekt i databasen (rad 38 i figur 19).

```
29    with open("keywords.json") as jsonFile:
30        global jsonData
31        jsonData = json.load(jsonFile)
32
33  for word1 in jsonData:
34      for word2 in jsonData[word1]:
35          if word2 in description or word2 in title:
36              if word1 not in item['tags']:
37                  item['tags'].append(word1)
38                  self.collection.insert(dict(item))
```

Figur 19: Utdrag från pipelinens kod som visar skapande av ämnesområde.

Slutligen ser koden i pipelinen ut som följer.

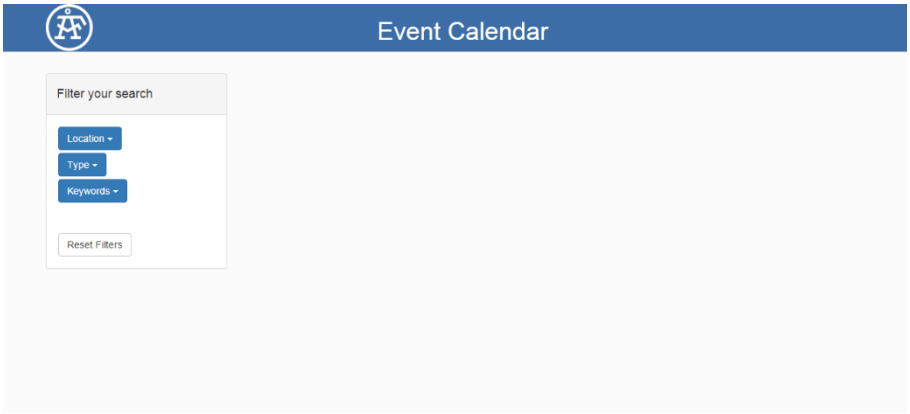
```
15  class AfeventPipeline(object):
16
17      def __init__(self):
18          connection = pymongo.MongoClient(
19              settings['MONGODB_SERVER'],
20              settings['MONGODB_PORT']
21              )
22          db = connection[settings['MONGODB_DB']]
23          self.collection = db[settings['MONGODB_COLLECTION']]
24
25      def process_item(self, item, spider):
26          if self.collection.find_one({'url': item['url']}):
27              raise DropItem('Item already in DB')
28
29      with open("keywords.json") as jsonFile:
30          global jsonData
31          jsonData = json.load(jsonFile)
32
33  for word1 in jsonData:
34      for word2 in jsonData[word1]:
35          if word2 in description or word2 in title:
36              if word1 not in item['tags']:
37                  item['tags'].append(word1)
38                  self.collection.insert(dict(item))
```

Figur 20: Förenklat utdrag från processens pipeline.

5.3 Genomgång av prototypen

Prototypen består av tre huvudkomponenter; en funktion som utvinner data, en funktion som bearbetar data och en webbsida som visualiserar informationen på ett strukturerat sätt. Prototypens uppgift är således att utvinna semistrukturerad information och redovisa den på ett strukturerat sätt. Detta avsnitt beskriver hur informationen som lagrats i den tidigare nämnda extraheringsprocessen presenteras för användare.

Webbsidan är ursprungligen tom och innehåller ingen information från databasen. Dock finns en del grafiska element som satts för att webbsidan ska se ut och fungera som önskat. Nedan kan ses att en filterfunktion återfinns på vänster sida. Detta filter innehåller tre "drop down"-menyer för plats, typ av evenemang och nyckelord, och dessutom en knapp där filtret kan återställas till att innehålla alla evenemang i listan. Överst på sidan har ÅF:s logga och en grafisk board lagts in för att användaren ska känna igen sig.

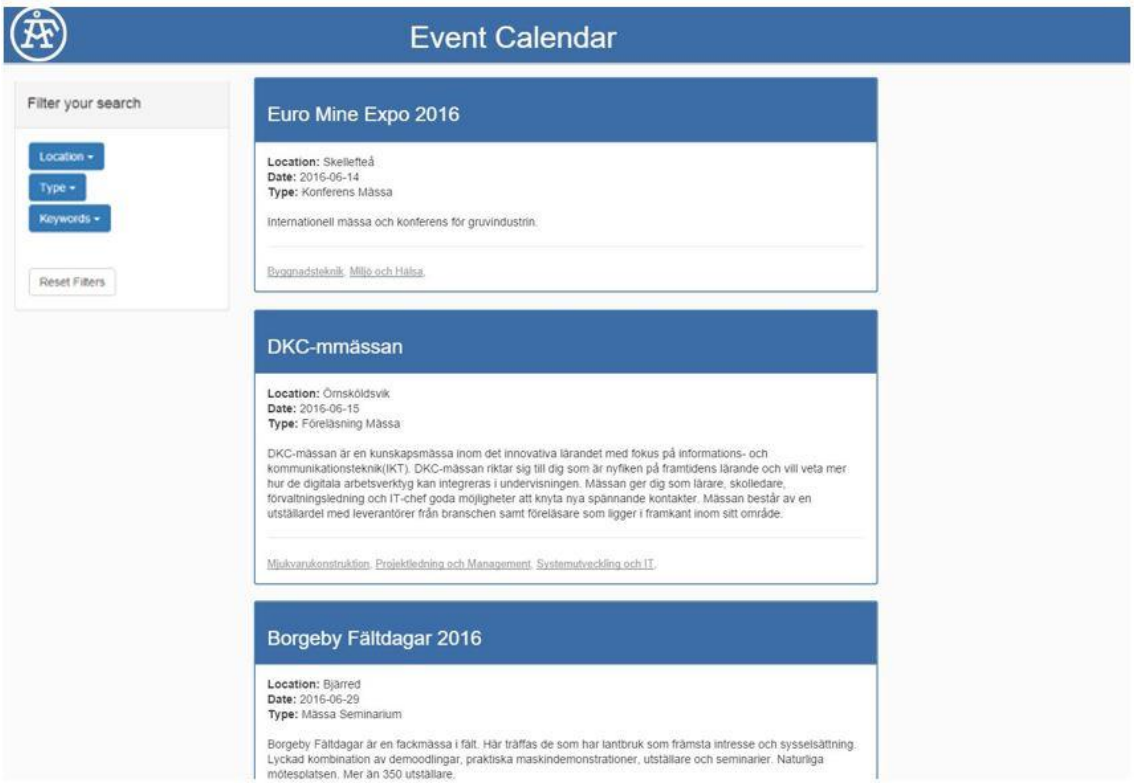


Figur 21: Webbsidan innan information extraherats och förts in i databasen.

Extrahering av information sker en gång per dygn nattetid, då webbspindeln söker igenom på förhand utvalda webbsidor där evenemang presenteras. Programmet söker igenom källkoden på webbsidorna och lokaliserar de evenemang som finns. Varje evenemang tilldelas attribut; exempelvis plats, datum och beskrivning, och skickas sedan vidare för att bearbetas.

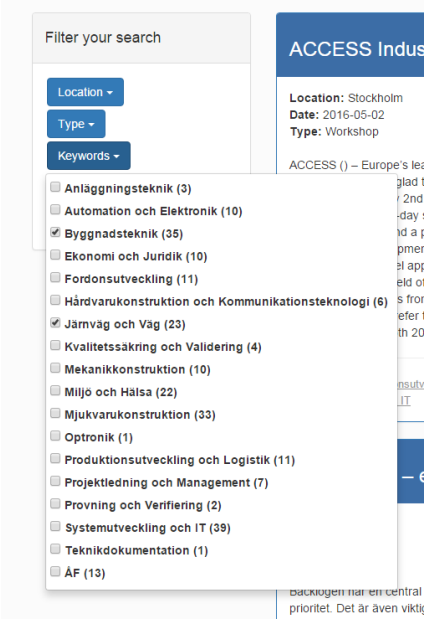
Under bearbetningen sker en rad processer för att strukturera data från alla källor till en generisk form. Detta innebär bland annat att formatera det angivna datumet så att det visas på samma sätt för alla evenemang eller att radera överflödig information från enskilda attribut.

När extraherings- och bearbetningsprocessen gjorts har varje evenemang tilldelats följande attribut: titel, datum, plats, beskrivning, URL-adress, nyckelord och typ av evenemang. Först efter detta lagras evenemanget i en databas för att kunna visas senare. Följande figur visar hur webbsidan ser ut efter att information extraherats och lagrats i databasen.



Figur 22: Exempel på webbsidans struktur efter att information förts in i databasen.

När evenemang lagrats i databasen ges en mer strukturerad framställning av information som är extraherad från flera källor. Detta ger användaren snabbare en överblick av kommande evenemang, och möjligheten att filtrera evenemang baserat på plats, typ av evenemang och ämnesområde. Följs länken kommer användaren direkt till evenemangets hemsida där det går att läsa mer och anmäla sig att delta. Hela processen sker automatiskt utan krav på administratör eller redaktör.



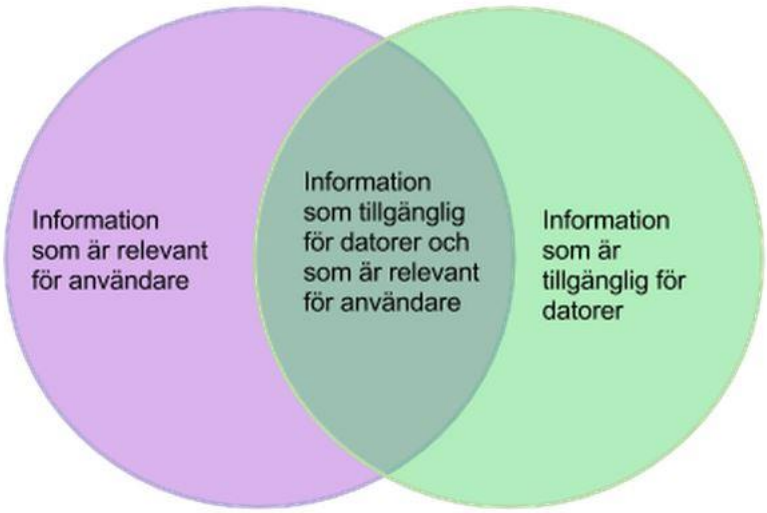
Figur 23: Exempel från filtreringsfunktionen.

6. Analys

I detta kapitel analyseras skillnaden i resultat mellan vad användarna efterfrågar och vad webbspindeln kan extrahera. En egen analysmodell presenteras också innan diskussionskapitlet tar vid.

6.1 Användarkrav i förhållande till webbspindelns förmåga

Webbspindelns processer sker enligt strikta ramverk, vilket innebär att datorn vare sig hämtar mer eller mindre information än vad den blivit tillsagd att hämta. Detta skiljer sig från de kognitiva processer som styr hur människor tar till sig information. För att information ska kunna avläsas effektivt av både datorer och människor krävs, förutom dess relevans, att den presenteras på ett sätt som gör den lättillgänglig för båda. Denna information kan betraktas som ett snitt mellan den information som är relevant för mänskliga användare och information som kan hämtas av datorer. För att kunna utvidga tvärsnittets area krävs både en förståelse för vilken information användare anser vara relevant för ändamålet, samt för hur väl webbspindlar i dagsläget kan hämta denna information.



Figur 24: Illustration av informations tillgänglighet för människor och datorer.

Efter att den för studien utvecklade webbspindeln extraherat och sorterat information från givna källor går det att konstatera att en automatiserad process kan finna relevanta evenemang och presentera dessa på ett strukturerat sätt. Däremot finns stora möjligheter för förbättringar eftersom en rad kompromisser har varit nödvändiga för att få programmet att fungera på ett önskvärt sätt. Dessa gås igenom djupare nedan.

Eftersom varje webbsida krävde att webbspindeln konfigurerades för att fungera på önskvärt sätt visade det sig vara ineffektivt att extrahera information från

förstahandskällor. Förstahandskällor innebär de webbsidor där evenemang först publiceras av arrangören. Därför gjordes valet att extrahera information från existerande portaler där vissa typer av evenemang redan fanns sammanställda. Dessa portaler tillhörde exempelvis branschorganisationer och föreningar. Genom att genomsöka webbsidor som innehöll många evenemang utvann varje spindel jämförelsevis flera evenemang vilket ökade effektiviteten avsevärt.

Från enkätundersökningen framgår det att de viktigaste attributen som anställda på ÅF vill se i en sammanställning är "Ämnesområde" och "Typ av evenemang". Ingen av de extraherade evenemangen innehöll denna information. "Ämnesområde" och "Typ av evenemang" genererades istället i en efterföljande process. Ingenting fungerade automatiskt och att skapa denna process var tidskrävande och krävde olika lösningar. Denna efterföljande process förutsatte att antingen evenemangets titel eller beskrivning innehöll särskilda nyckelord som kunde knytas till ett specifikt ämnesområde eller typ av evenemang. För att kunna knyta ett evenemang till ett ämnesområde krävdes således att titel eller beskrivning innehöll facktermer eller andra branschspecifika ord vilket sällan är fallet; de allra flesta evenemang har en beskrivning i vardagliga ordalag som försvårar en automatisk kategorisering. Detta medförde att de attribut som användarna ansåg vara viktigast var svårast att erhålla från webbsidorna med hjälp av webbspindlarna. Samma problem förekom när det gällde arrangör, men då undersökningen pekade på att detta inte var lika viktigt för användarna ansågs problemet inte vara av samma vikt.

En annan funktion som användarna ansåg viktig var att ha en sammanfattande text om evenemangen i kalendern. Även denna funktion medförde en del problem för webbspindeln. Eftersom alla sidor hade sina egna sätt att presentera information om eventen var även "Sammanfattande text" svår att presentera strukturerat i eventkalendern. Då det inte fanns någon övergripande struktur för hur textbaserad information presenterades på webbsidorna var det svårt att träna spindeln till att extrahera just det användaren finner intressant och relevant i en specifik text. Många texter på hemsidorna innehöll relevant information men ofta var denna information uppblandad med mindre relevant information, exempelvis mailadress till den ansvariga för evenemanget, vägbeskrivning till den plats evenemanget skulle hållas på, information om dagens fika med mera. Då denna information inte anses lämpad för denna typ av eventkalender kunde ingen information om eventet hämtas i dessa fall.

Som nämnt tidigare kan en webbspindel varken hämta mer eller mindre än exakt det den anpassats till att hämta. Då spindeln inte kan tränas till att känna igen relevant information automatiskt, som en människa inte har några som helst bekymmer med att urskilja, går det inte heller att träna spindeln till att bara hämta den typ av relevant information som användaren eftersöker. För att en spindel ska fungera helt automatiserat utan mänskligt ingripande måste den kunna skilja mellan relevant och irrelevant information. Detta kan i bästa fall åstadkommas genom att ge spindeln extremt strikta förhållningsregler men då

alla sidor ser olika ut sinsemellan är det omöjligt att få spindeln att restriktivt finna korrekt information. I och med denna observation kan det konstateras att det här finns en betydande diskrepans mellan vad användare efterfrågar jämfört med vad webbspindeln kan presentera automatiskt.

På fjärde plats i rangordningen gällande vad användarna ansåg som viktiga komponenter i en eventkalender var "Länk till event". Att finna länkarna till evenemangen var den enklaste uppgiften i skapandet av eventkalendern. Eftersom länkar fungerar annorlunda än fritext låg de alltid prydligt sorterade under rätt elementmarkering i webbsidornas källkod. Att extrahera rätt länk till rätt evenemang gick för alla webbsidor som besöktes och här kunde alltså ingen diskrepans urskiljas gentemot vad användarna efterfrågade.

I enkätundersökningen undersöktes även vilka övriga funktioner användarna önskade se i en eventkalender. Som presenterat i enkätresultaten var alternativet "Att kunna få förslag på passande event" i särklass mest populärt då 73% hade med alternativet i sitt svar jämfört med "Att kunna synkronisera med kalender" som var näst populärast på 41%. Även här kan en viss diskrepans urskiljas eftersom detta alternativ är något som en webbspindel inte kan producera. Som nämnt i intervjuresultaten finns en utbredd förväntan hos användare att teknik ska vara "smart" och kunna presentera alternativ för användaren. Idag tycks användare förvänta sig att alla tekniska funktioner ska kunna kopplas sinsemellan och att tekniken ska kunna tänka åt en (Intervju, Respondent 3). Användaren förväntar sig att funktioner ska kopplas samman och fungera ihop, helt automatiskt utan mänsklig närvaro. För att kunna skapa dessa avancerade algoritmer krävs mer tid och större resurser än vad som funnits tillgängligt i denna undersökning och ligger utanför avgränsningarna för denna studie.

7. Diskussion

I diskussionskapitlet diskuteras studiens resultat. Först diskuteras hur skillnaden mellan vad användarna efterfrågar och vad webbspindeln extraherar kan förminskas; antingen genom att användarnas efterfrågan förändras, att webbspindlar förbättras eller att webbsidors struktur förändras. Därefter diskuteras etiken kring webbspindlar och hur dessa kan användas på ett medvetet sätt. Slutligen diskuteras alternativ till webbspindlar.

7.1 Hur diskrepansen kan förminskas

För att utveckla en webbspindel som förbättrar kategorisering av den extraherade datan och som dessutom tillhandahåller mer information till slutanvändaren finns vissa åtgärder som kommer att diskuteras nedan. För att en automatiserad process ska kräva så lite underhåll som möjligt krävs en flexibel lösning som även är skalbar, vilket öppnar möjligheten för att anpassa den för olika användningsområden och förändrade behov.

7.1.1 Användarnas förändrade vanor och förväntningar

Enkätundersökningen och intervjuerna som gjorts i samband med denna studie visar att det råder en gemensam uppfattning bland anställda på ÅF att det i dagsläget tar för mycket tid att aktivt leta efter evenemang som eventuellt är intressanta. På grund av detta väljer användare att inte leta aktivt, utan får istället bara reda på evenemang som skickas till dem via e-post. Dessa får de endast om de tidigare deltagit i evenemang och registrerat sin e-postadress till en arrangör. Detta innebär att det är svårt för användare som inte är frekventa deltagare i evenemang att överhuvudtaget få information om vad som pågår för evenemang inom deras respektive yrkesområde.

Vidare pekar enkätundersökningen på att de flesta användare önskar se ytterligare funktionalitet i en eventkalender i form av att kunna få skräddarsydda tips om relevanta evenemang. Att kunna se deltagande kollegor och tipsa varandra om evenemang var också önskvärt. Detta tycks stämma överens med de slutsatser som Bachrach och McKean (2014) drar, nämligen att användare hellre skapar sig en uppfattning baserat på egna och andras bedömning av en tjänst eller vara, snarare än genom uppmaning från leverantören.

Användare av webbaserade tjänster höjer sina förväntningar i takt med den standard som andra, allt mer avancerade tjänster som de använder, håller (Intervju, Respondent 3). Därför kan det tyckas orimligt att användare plötsligt ska sänka sina förväntningar gällande vissa typer av tjänster, trots att tekniken tillåter en förbättrad upplevelse. Om tekniska lösningar kan möta användares högre krav bör det därför vara en punkt att utgå ifrån.

7.1.2 Förbättring av webbspindlar

I teorin finns ingen gräns för hur mycket data som kan extraheras av en webbspindel. Det finns heller ingen begränsning för hur mycket extraherad data kan bearbetas för att möta användarens behov. Detta innebär att en spindel skulle kunna hämta all tillgänglig information från hundratusentals webbsidor och bearbeta data i syfte att kategorisera den. Det finns dock en rad problem med detta där det största problemet är frånvaron av semantisk struktur, vilket diskuteras av bland andra Alvarez et al. (2008). Stora företag, exempelvis Google, arbetar med detta för att göra information på internet sökbar men det gör det inte lättare att kategorisera data efter vad det är för typ av information den innehåller. Det krävs med andra ord mycket resurser för att hantera så stora mängder data.

Andra problem med att hämta stora mängder data från webbsidor är att det innebär en belastning på trafiken till och från sidan. Denna belastning påverkar driftkostnad och hastighet, vilket medfört att många större webbsidor bygger in spärrar som gör det svårare för webbspindlar att utvinna information. Detta kan komma att försvåra användandet av webbspindlar i framtiden och utvecklare kan komma att tvingas anpassa spindlarna ytterligare för att komma runt denna problematik.

7.1.3 Förbättring av webbstruktur

Ett alternativ för att minska skillnaden mellan vad användarna efterfrågar och vad som faktiskt finns att tillgå på internet via datainsamling med hjälp av webbspindlar är att i större utsträckning standardisera strukturen av hur information lagras i HTML-kod på internet. Liksom inom de flesta discipliner finns inom programmering vissa konventioner som följs av professionella utövare. Det tycks dock råda en generell frånvaro av gemensam struktur när det gäller presentation av viss information på internet. Elementmarkeringar som dykt upp på de sidor som denna studie berört har haft godtyckliga namn som sällan varit intuitiva eller logiska. Det medför att varje enskild sida på internet kräver en webbspindel som är särskilt konfigurerad för webbsidan i fråga för att fungera.

Skulle det existera konventioner för hur yrkesverksamma webbutvecklare benämner evenemang i webbsidans källkod skulle effektiviteten för webbspindlar öka avsevärt. Nedan följer en jämförelse för hur källkoden i figur 25, som innehåller källkoden från exemplet i avsnitt 5.2.1, skulle kunna skrivas mer generisk.

Figur 25: Källkod från exempel i avsnitt 5.2.1 (swedsoft.se).

Figur 26: Förslag till hur källkoden i figur 25 kan göras mer generisk.

I originalkoden i figur 25 benämns den del av koden som innehåller information om ett evenemang `<a class=”article-link”>`. Titeln på evenemanget finns i elementmarkeringen `<h1 class=”h2 entry-title”>` och datum i `<p class=”meta medium light”>`. I figur 26 benämns motsvarande element `<div class=”event”>`, `<div class=”event-title”>` samt `<div class=”event-date”>`. I figur 26 är med andra ord källkoden mer logisk och enklare att följa. Figur 26 är endast ett förslag på hur en standardiserad benämning på elementmarkeringar skulle kunna se ut i praktiken.

Fördelarna med att ha en konvention bland webbutvecklare som ger generiska namn på särskilda elementmarkeringar är tydliga. En webbspindel skulle inte behöva anpassas efter enskilda webbsidor utan istället skulle den programmeras att söka igenom ett godtyckligt antal domäner som är av relevans för användaren. Vidare skulle utvecklare kunna ge publicerade evenemang nyckelord i källkoden vilket skulle innebära att webbspindeln direkt kunde avgöra evenemangets relevans utan att först extrahera informationen.

Generiska benämningar på evenemang som samverkar med webbspindlar av det slag som utvecklats i denna studie skulle öppna möjligheten att marknadsföra evenemang snabbare

och mer effektivt än vad som är möjligt idag. En arrangör skulle ges möjlighet att informera alla potentiella intressenter om sitt evenemang inom loppet av några minuter, förutsatt att evenemanget publiceras under en domän som en webbspindel genomsöker. Metoden har vidare potential att sträcka sig förbi branschspecifika evenemang. Med hjälp av generiska elementbenämningar och nyckelord i källkoden skulle alla typer av evenemang ges möjligheten till automatisk marknadsföring.

Standarder som RDF-elementmarkeringar i källkoden skulle kunna bidra till att effektivisera hur information presenteras på internet. Denna befintliga teknik används dock inte på några av de webbsidor som behandlats i denna studie, vilket innebär ett hinder för webbspindlar som då måste konfigureras efter varje enskilt fall.

7.2 Medveten användning av webbspindeln

Webbspindlar är effektiva när det gäller att extrahera information från webbsidor. Detta är anledningen till att de används av både stora och små aktörer på internet. Det finns dock vissa faktorer som talar emot användningen av webbspindlar bland annat överbelastning, användarintegritet och upphovsrättskyddat material (Thelwall, 2006).

Webbspindeln som är utvecklad för denna studie extraherar en förhållandevis liten mängd data och innebär därför inga problem för ägarna av webbsidorna. Varje besökt sida behandlas under ett fåtal sekunder trots att webbspindeln är programmerad att inte följa länkar så snabbt som möjligt, allt för att undvika intensiv belastning på webbsidan. Det innebär att spindeln inte medför nämnvärt större belastning än en mänsklig användare. Vidare besöks sidorna nattetid av den utvecklade webbspindeln eftersom belastningen på sidorna då betraktas som låg. På samma sätt innebär det att kostnaden för webbsidans ägare inte påverkas av att webbspindeln besöker deras sida. Eftersom stor del av kritiken som riktas mot webbspindlar handlar om att de ökar trafiken på webbsidorna och gör dem långsammare för andra användare, var det av stor vikt att detta undveks i utvecklingen av denna eventkalender.

All information som extraheras av denna webbspindel är publik. Det är med andra ord inte information som bör betraktas som känslig på något sätt. Informationen som visas på prototypens webbsida är densamma som den publicerade informationen. Den information som publiceras på prototypens webbsida är tänkt att informera fler personer om specifika evenemang. All information som publiceras står i anslutning till en länk som leder till originalkällan. Detta innebär att det inte sprids något upphovsrättsskyddat material av webbspindeln utan denne istället hänvisar användaren till källan. Det innebär också att den utvecklade eventkalendern inte bara gynnar användare som söker efter evenemang utan även arrangörerna som bör uppleva större intresse för sina evenemang utifrån, samt för de som skapat webbportalerna för dessa evenemang. På grund av dessa faktorer fungerar eventkalendern som en förlängning av evenemangarrangörernas och webbportalernas intressen att sprida och skapa publicitet kring evenemangen.

7.3 Alternativa lösningar

För att åstadkomma samma resultat som eventkalendern i nuläget presenterar utan att använda automatiserade webbspindlar krävs mycket tid av användaren. Ett tänkbart scenario för att skapa en liknande eventkalender utan webbspindel skulle vara att användaren letar upp alla evenemang för att sedan manuellt föra in dem i databasen eller direkt in på eventkalenderns hemsida. Fördelen med att göra på detta sätt skulle vara att rätt information i större utsträckning skulle hamna på rätt plats och på rätt format. Dessutom skulle användaren som skapar sidan kunna avlägsna irrelevant information från exempelvis "Sammanfattande text" eftersom denne i högre utsträckning än webbspindeln kan urskilja viktig information från oviktig information så som om det erbjuds fika eller hur tillgången på parkeringsplatser ser ut. Det skulle dock krävas orimligt mycket tid och en outtömlig energidepå för denne användare vilket gör detta till ett nästintill befängt alternativ.

Ytterligare ett alternativ till att skapa en motsvarande eventkalender utan att använda webbspindlar finns. Detta alternativ skulle innebära att en formulärbaserad webbsida utvecklades där evenemangsarrangörerna själva kunde gå in och ladda upp information om sina evenemang; under rätt ämneskategorier, med egen sammanfattande text med mera. Detta skulle likna ovanstående alternativ men med skillnaden att det då går att undvika att samma person måste stå för väldigt mycket producerande av information. Dessutom kunde evenemangsarrangörerna själva bestämma vad som skulle publiceras i eventkalendern. För att detta alternativ skulle fungera krävs det att marknadsföringen av eventkalendern, som idag är obefintlig, skulle behöva öka markant. För att få en fungerande eventkalender där arrangörerna själva fyller i formulär och lägger upp sina evenemang skulle det vara viktigt att nästintill alla arrangörer kände till eventkalendern och faktiskt tog sig tiden att lägga upp informationen kring sina evenemang. Även detta alternativ anses mycket sämre än alternativet att använda sig av webbspindlar.

8. Slutsatser

Studiens frågeställning har varit *I vilken utsträckning skiljer sig informationen som kan hämtas via en webbspindel från den information som ett företags användare efterfrågar, och hur kan denna skillnad förminskas?*, med syftet att ta reda på hur informationsspridning kan effektiviseras och anpassas efter användare. Denna frågeställning, med syftet i åtanke, kommer att besvaras i följande kapitel.

Studien har visat att en webbspindel endast kan finna och presentera relevanta evenemang, och motsvarande information gällande dessa evenemang, om webbspindeln anpassas efter de aktuella webbsidorna i fråga. Studien har visat att det finns en betydande diskrepans mellan vad användare efterfrågar och vad en webbspindel av detta slag kan presentera automatiskt i den grad att de för användarna viktigaste attributen även är de svåraste att automatiskt extrahera för webbspindeln.

Denna diskrepans kan förminskas bland annat genom att användarna förändrar vad de efterfrågar eller att tekniken kring webbspindlarna förbättras. Slagkraftigaste förändringen skulle dock vara om strukturen av hur information lagras i HTML-kod på internet kan standardiseras. Det har visats att det råder en generell frånvaro av gemensam struktur när det gäller presentation av viss information på internet. Elementmarkeringarna som dykt upp i denna studie har haft godtyckliga namn som sällan varit logiska. Detta medför att varje enskild sida på internet kräver en webbspindel som är särskilt konfigurerad för webbsidan i fråga för att fungera optimalt. Detta skulle lösas om standarden istället blir att webbsidor utvecklas inom samma övergripande struktur med konventioner för hur yrkesverksamma webbutvecklare benämner evenemang i webbsidans källkod. Det skulle medföra möjligheten att göra enkla webbspindlar mer kraftfulla och öppna upp för att göra tekniken mer skalbar. För företag skulle detta innebära en enkel sammanställning av relevanta evenemang, och för arrangörer en mycket effektiv marknadsföringsmetod.

Denna studie har visat att webbspindlars effektivitet är svår att förkasta och att tekniken är enkel att utveckla och använda. Resultaten av denna studie tillsammans med diskussionen förd i analys- och diskussionskapitlet resulterar i slutsatsen att det inte finns några likvärdiga alternativ till att skapa en eventkalender för detta ändamål utan webbspindlar.

8.1 Vidare forskning

I en mer omfattande studie hade det varit intressant att belysa om och i så fall hur en eventkalender likt den som är utvecklad i samband med denna studie skulle förändra webbtrafiken på de sidor eventkalendern hänvisar till. Det borde vara av intresse för de webbportaler som lägger upp evenemang i dagsläget att se så mycket trafik som möjligt

på sina sidor eftersom deras syfte är att marknadsföra evenemang till så många som möjligt. Det vore också intressant att undersöka om och i så fall hur eventkalendern som utvecklats i denna studie skulle förändra antalet människor som deltar i evenemangen som den hänvisar till. Om resultatet av en sådan studie skulle visa att eventkalendern bidrar till att fler människor går på evenemangen skulle det vara ett väldigt värdefullt marknadsföringsverktyg för de som skapar evenemangen och vill se många deltagare närvara.

Denna studie är genomförd med ÅF som fallföretag och ÅF:s anställda som användare. Det skulle vara relevant att undersöka om studien skulle ge liknande resultat på andra företag med andra användare eller om det är stor skillnad mellan olika företag, användare och geografisk tillhörighet för att bara nämna några. Denna studie är utförd ur användarnas perspektiv och det skulle därför även vara givande att forska mer utifrån evenemangsägarnas synvinkel. Resultaten från en undersökning gällande ytterligare webbstruktur och förenkling för webbspindlar skulle eventuellt kunna gynna de som skapar evenemangen och evenemangsportalerna i framtiden.

9. Referenser

Ahuja, M., Jatinder, S. 2014. “Web Crawler: Extracting the Web Data”. International Journal of Computer Trends and Technology (IJCTT). s. 132-137.

Álvarez, M., Alberto, P., Raposo, J., Bellas, F & Cacheda, F. 2008. “Extracting Lists of Data Records from Semi-structured Web Pages”. Data & Knowledge Engineering. Vol. 64, s. 491-509.

Bachrach, D., McKean, J. 2014. “Customer’s New Voice: Extreme Relevancy and Experience through Volunteered Customer Information”. Wiley. s. 26-27, 124-126.

Batsakis, S., Petrakis, E. & Milios, E. 2009. “Improving the Performance of Focused Web Crawlers”. Data & Knowledge Engineering. Vol. 68:11, s. 1001-1013.

Beamer, S., Asanović, K. & Patterson, D. 2013. “Direction-Optimizing Breadth-First Search”. Scientific Programming. Vol. 21, s. 137-148.

Bootstrap. “About”. http://getbootstrap.com/about/. Hämtad: 2016-05-12.

Brace, I. 2008. “Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research”. 2nd ed. London, GBR: Kogan Page Ltd.

Bryman, A. 2011. “Samhällsvetenskapliga metoder”. 2. [rev.] uppl. Malmö: Liber.

Carter, J. 2015. “Google Got it Wrong: The Internet Won't be Global by 2020”. http://www.techradar.com/news/internet/google-got-it-wrong-the-internet-won-t-be-global-by-2020-1306432. Hämtad: 2016-03-23.

Chooralil, V. & Gopinathan, E. 2015. “A Semantic Web Query Optimization Using Resource Description Framework”. Procedia Computer Science. Vol. 70, s. 723-732.

Computer Sweden. “IT-ord”. http://it-ord.idg.se/ord/dynamisk-webbsida/. Hämtad: 2016-05-02.

Domenech, J., Gil, Sahuquillo, J., Pont, A. 2010. “Using Current Web Page Structure to Improve Prefetching Performance”. Computer Networks. Vol. 54(9), s. 1404-1417.

Duval, E., Hodgins, W., Sutton, S. & Weibel, S. 2002. “Metadata Principles and Practicalities”. D-Lib Magazine. Vol. 8(4).

Guo, Y., Wang, J. 2012. "Scrapy-Based Crawling and User-Behavior Characteristics Analysis on Taobao". IEEE. S. 44.

Gyorödi, C., Gyorödi, R. & Sotoc, R. 2015. “A Comparative Study of Relational and Non-Relational Database Models in a Web- Based Application”. International Journal of Advanced Computer Science and Applications. Vol. 6.

Internet Live Stats. “Internet Users in the World”. http://www.internetlivestats.com/internet-users/. Hämtad: 2016-02-11.

Internet Live Stats. “Total Number of Websites”. http://www.internetlivestats.com/total-number-of-websites/. Hämtad: 2016-02-11.

Kim, K., Kim, K., Lee, K., Kim, T. & Cho, W. 2012. “Design and Implementation of Web Crawler Based on Dynamic Web Collection Cycle”. CoreEngineering. s. 643-649.

Kongsved SM., Basnov M., Holm-Christensen K. & Hjollund NH. 2007. “Response Rate and Completeness of Questionnaires: A Randomized Study of Internet Versus Paper-and-pencil Versions”. Journal of Medical Internet Research. Vol. 9(3).

Mehlhorn, K. & Sanders, P. 2008. “Algorithms and Data Structures: The Basic Toolbox”. http://people.mpi-inf.mpg.de/~mehlhorn/ftp/Toolbox/GraphTraversal.pdf. Hämtad: 2016-02-10.

Powers, D. 2010. “PHP Solutions: Dynamic Web Design Made Easy”. 2nd ed. New York: Friendsof.

Richards, R. 2006. “Pro PHP XML and Web Services”. 1st ed. Apress. S. 124-125

Rouse, M. 2014. “Semi-structured data”. http://whatis.techtarget.com/definition/semi-structured-data. Hämtad: 2016-05-11.

Sauers, M. 2010. “Blogging and RSS: a Librarian’s Guide”. 2nd edition. Information Today, Inc.

Saunders, M., Lewis, P. & Thornhill, A. 2015. “Research Methods for Business Students”. 7th edition. Harlow: Pearson Education.

Schultz, D. & Cook, C. 2007. “Beginning HTML with CSS and XHTML: Modern Guide and Reference”. Apress.

Scrapy. “Architecture Overview”. http://doc.scrapy.org/en/latest/topics/architecture.html. Hämtad: 2016-02-12.

Serrano, Á., Maguitman, A., Boguñá, M., Fortunato, S. & Vespignani, A. 2007. “Decoding the Structure of the WWW: A Comparative Analysis of Web Crawls”. ACM Transactions on the Web. Vol. 1(2).

Sidiropoulos, A., Pallis, G., Katsaros, D., Stamos, K., Vakali, A. & Manolopoulos, Y. 2008. “Prefetching in Content Distribution Networks via Web Communities Identification and Outsourcing”. World Wide Web. Vol. 11(1), s. 39-70.

Sirsap, S. 2014. “Extraction of Core Contents from Web Pages”. International Journal of Engineering Trends and Technology. Vol. 8(9), s. 484-489.

Stiftelsen ÅForsk. “Stiftelsen ÅForsk”. http://www.aforsk.se/. Hämtad: 2016-02-08.

Tafesse, W. 2016. “Conceptualization of Brand Experience in an Event Marketing Context”. Journal of Promotion Management. Vol. 22(1), s. 34-48.

Thelwall, M. 2006. “Web Crawling Ethics Revisited: Cost, Privacy, and Denial of Service”. Journal of the American Society for Information Science and Technology. Vol. 57(13), s. 1771-1779.

Tracy, P. & Carkin, D. 2014. “Adjusting for Design Effects in Disproportionate Stratified Sampling Designs Through Weighting”. Crime & Delinquency. Vol. 60(2), s. 306-325.

Zulqurnan, A. 2016. “Semantic Web Mining in E-Commerce Websites”. International Journal of Computer Applications. Vol. 137(2), s. 1-4.

ÅF. 2015. “Årsredovisning 2014”. Årsta: Ineko.

ÅF. “I korthet”. http://www.afconsult.com/sv/om-af/i-korthet/. Hämtad: 2016-02-08.

ÅF. “Utveckling sedan 1895”. http://www.afconsult.com/sv/om-af/historia/. Hämtad: 2016-02-08.

Bilaga A: Intervjufrågor

- Vad är din roll på ÅF och dina arbetsuppgifter?
- Hur länge har du jobbat på ÅF?
- Hur ser du framför dig att en sådan här kalender skulle fungera?
- Skulle det vara bra med en kalender där du kan finna evenemang?
- Varför vill anställda gå på evenemang?
- Vad kan anställda få ut från dem?
- Hur skulle det underlätta för dig med en kalender?
- Hur hittar du relevanta evenemang idag?
- Hur måna är chefer att konsulter deltar i evenemang?
- Kan du urskilja någon trend i extern utbildning?
- Hur vanligt är det att folk i din position går på evenemang?
- Hur går denna kunskapsöverföring till när anställda väl har varit iväg på något evenemang?
- Hur vanligt är det att folk på din avdelning vill gå/går på sådana evenemang?
- Finns det en strävan efter att bli mer frekventa deltagare på evenemang?
- Skulle folk gå på fler evenemang om de var lättare att söka upp?
- Vilka attribut är viktigast för att snabbt hitta relevanta evenemang?
- Vilka områden/ämneskategorier hade du varit mest intresserad av?
- Vilka arrangörer är ni och skulle kunna vara särskilt intresserade av?
- Vilka tycker du att tjänsten bör rikta sig mot?
- Hur vill du att eventkalendern ska se ut och innehålla?
- Hur mycket jobbar anställda ihop med andra divisioner för att marknadsföra ÅF?
- Hur mycket jobbar anställda ihop med andra divisioner för att dela kompetenser mellan divisionerna?

Bilaga B: Enkätfrågor

Hej!

Vi heter Åsa Bengtson och Karl Söderlund och vi gör vårt examensarbete på ÅF i Solna. Vår uppgift är att utveckla och presentera en webbaserad eventkalender där ÅF:s medarbetare kan söka upp och finna relevanta evenemang. Dessa evenemang kan vara exempelvis mässor eller föreläsningar och kan vara arrangerade antingen internt eller externt. Vi håller i nuläget på att undersöka användarbehoven och skulle uppskatta om du kan avvara några minuter för att hjälpa oss. Enkäten är anonym och de anonyma resultaten kommer presenteras i en rapport samt ligga till grund för den utformade eventkalendern.

Tack på förhand!

- Åsa och Karl.

---

1. Hur gammal är du?

- -30
- 31-40
- 41-50
- 51-60
- 60-

2. Vilken position har du på företaget?

- Konsult
- Chef
- Projektledare
- Övrigt: _____

3. Hur stor andel tid tillbringar du i snitt på ÅF:s kontor?

- Mindre än 50%
- Cirka 50%
- Cirka 75%
- 100%

4. Har du någon form av personalansvar?

- Ja
- Nej

5. Välj **TVÅ** typer av evenemang du är mest intresserad av att gå på:

- Konferenser
- Föreläsningar

- Mässor
- Work shops
- Övrigt:

  _____

6. Vilka evenemang skulle du vilja se i en eventkalender?

- Konferenser
- Föreläsningar
- Mässor
- Work shops
- Övrigt:

  _____

7. Välj ut de **FYRA** viktigaste attributen du skulle vilja få presenterade i en eventkalender:

- 1. Pris
- 2. Arrangör
- 3. Typ av evenemang (exempelvis mässa, föreläsning etc.)
- 4. Ämnesområde
- 5. Bransch
- 6. Sammanfattande text om event
- 7. Länk till event
- 8. Övrigt:

  _____

7a. Rangordna de, från ovanstående fråga, valda attribut sinsemellan:

Mest viktig_____
_____
_____
Minst viktig_____

8. Vilka **TVÅ** funktioner finner du viktigast?

- Att kunna se vilka kollegor som deltar i vilka event
- Att kunna synkronisera med annan kalender
- Att kunna betygsätta evenemang och arrangör
- Att kunna skriva utlåtanden om eventet
- Att kunna tipsa en kollega om ett event
- Att kunna få förslag på passande evenemang för just dig
- Övrigt:

  _____

9. Övriga förslag eller önskemål gällande en eventkalender?

_____
_____
_____

_____
_____
_____
_____

10. Har du några tips på bra webbsidor där du idag söker upp relevanta event?

_____
_____
_____

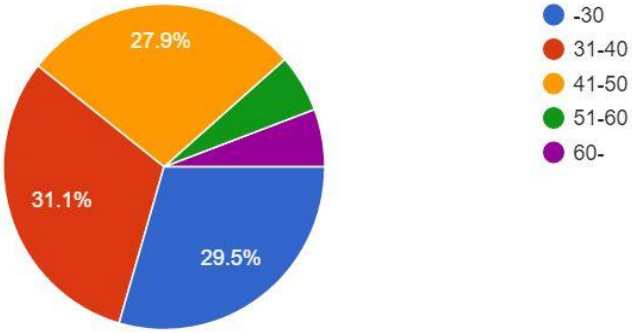
11. Är du intresserad av att ställa upp på en ev. demotestning av eventkalendern senare i vår?

Mailadress:
_____

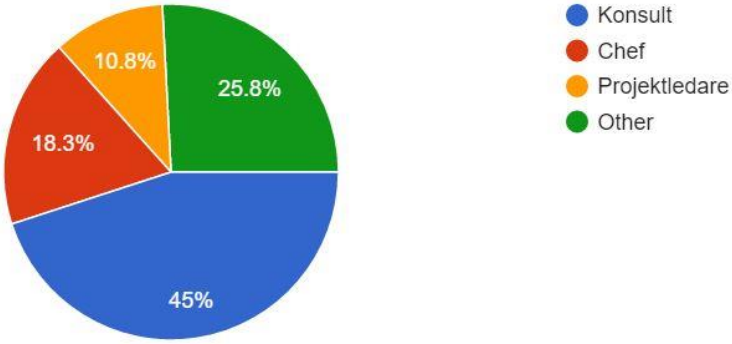
Tack för din medverkan!
- Åsa och Karl

Bilaga C: Svar från enkätundersökningen

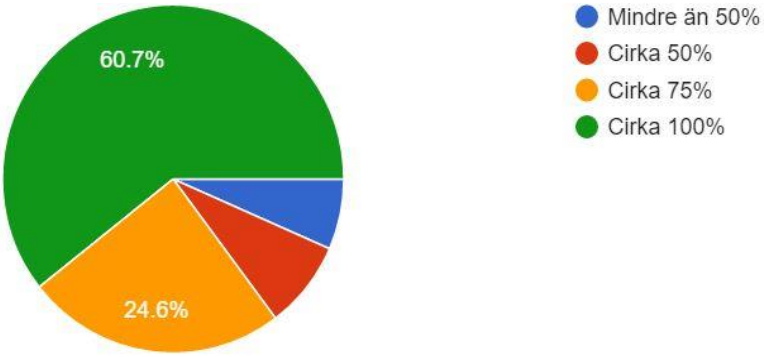
1. Hur gammal är du?



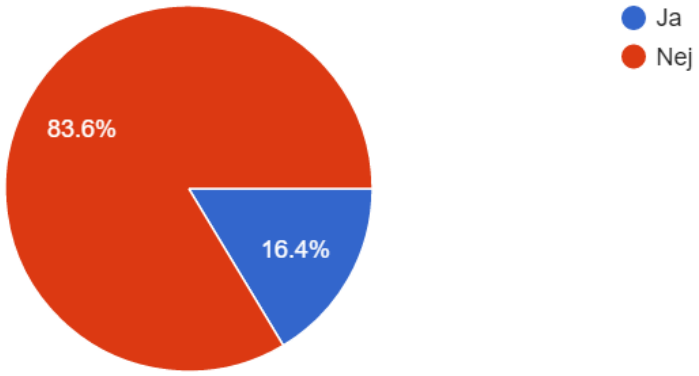
2. Vilken position har du på företaget?



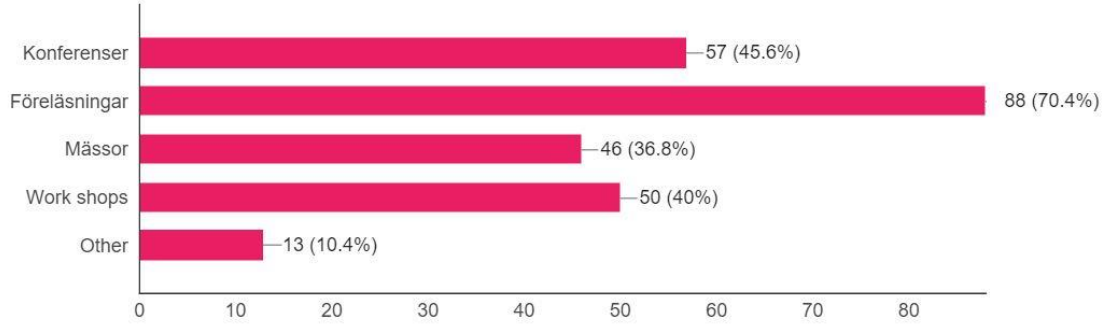
3. Hur stor andel tid tillbringar du i snitt på ÅF:s kontor?



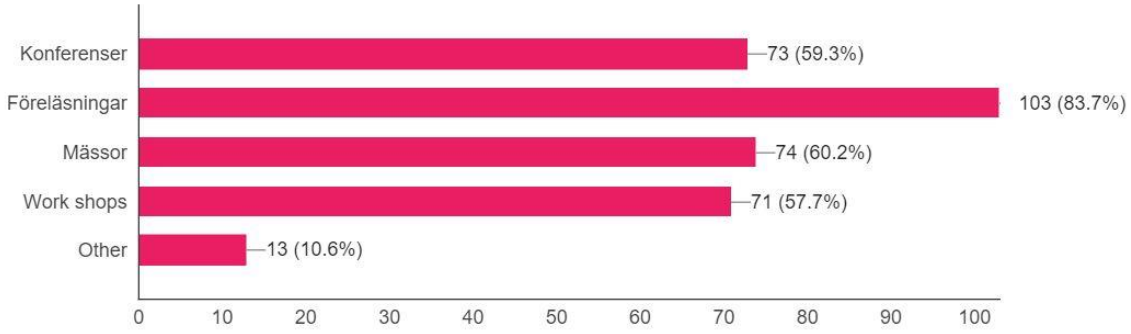
4. Har du någon form av personalansvar?



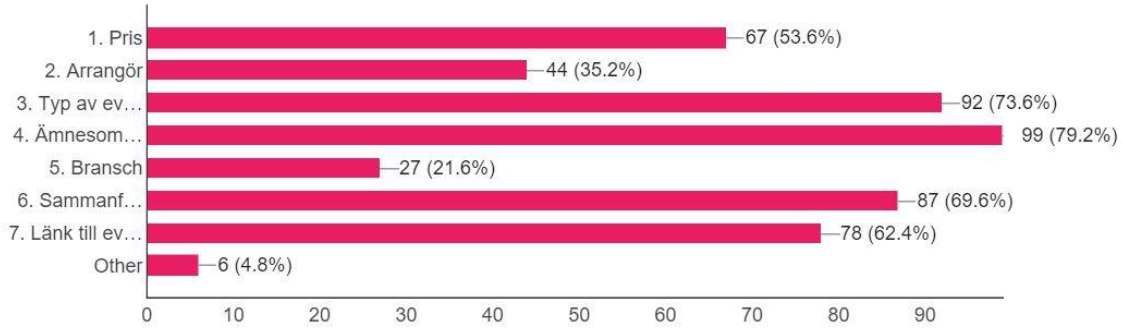
5. Välj **TVÅ** typer av evenemang du är mest intresserad av att gå på:



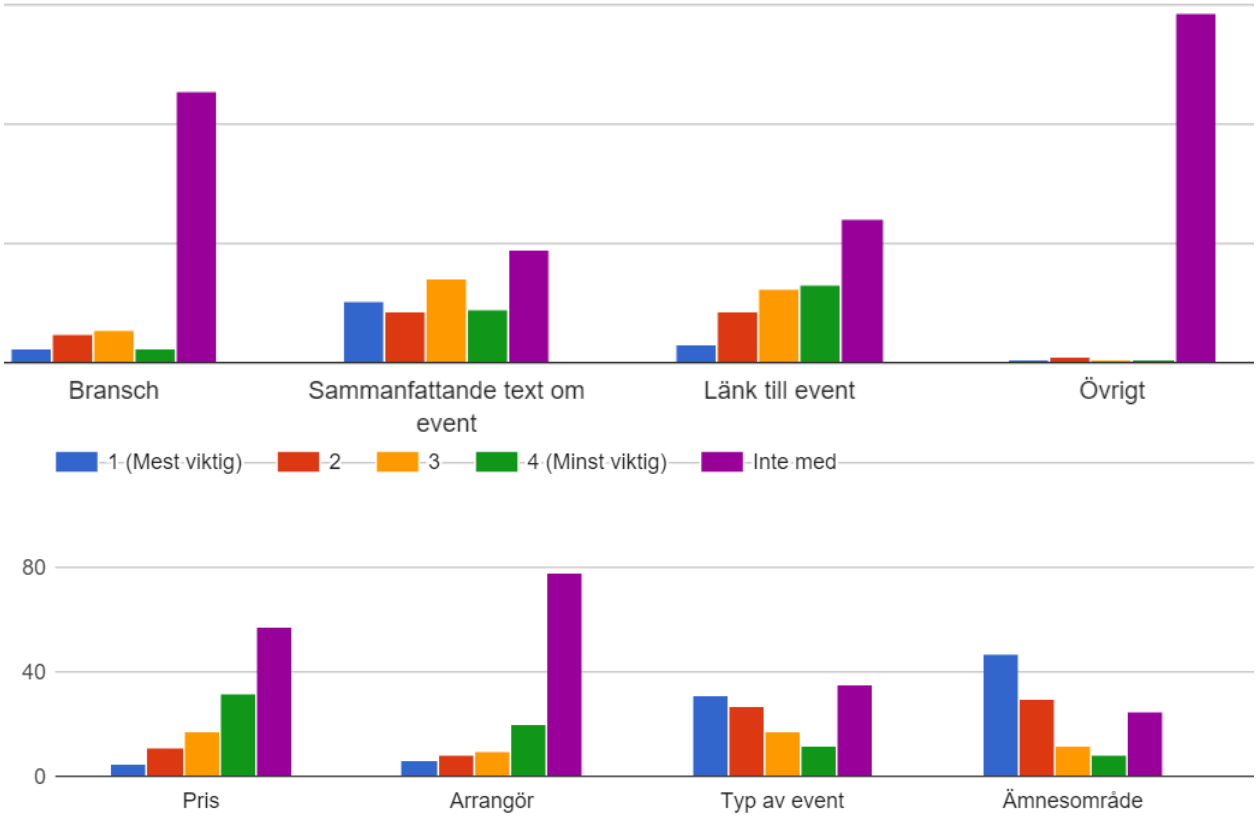
6. Vilka evenemang skulle du vilja se i en eventkalender?



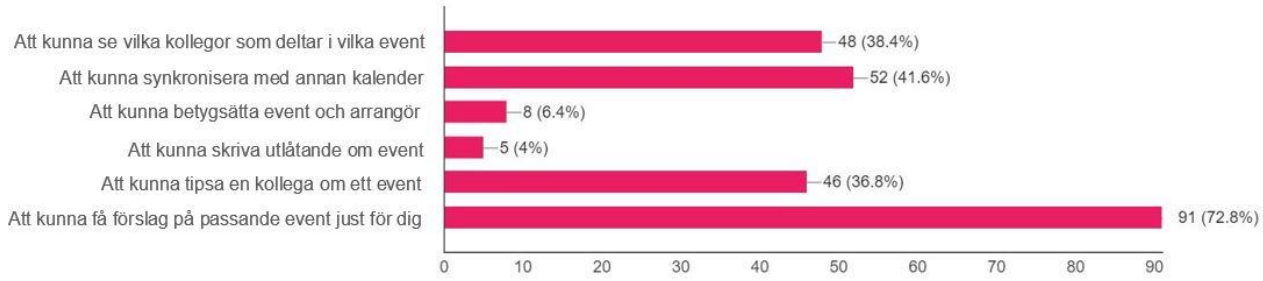
7. Välj ut de **FYRA** viktigaste attributen du skulle vilja få presenterade i en eventkalender:



7a. Rangordna de, från ovanstående fråga, valda attribut sinsemellan:



8. Vilka **TVÅ** funktioner finner du viktigast?



Bilaga D: Viss kod från webbspindlar och pipeline

Kod från webbspindeln som hämtar information från swedsoft.se:

```
import scrapy
from scrapy.spiders             import CrawlSpider, Rule
from scrapy.selector            import HtmlXPathSelector
from scrapy.linkextractors.sgml import SgmlLinkExtractor
from afevent.items              import AfeventItem
from scrapy.http                import Request
from scrapy.crawler             import CrawlerProcess
from urlparse                   import urljoin


class MySpider(CrawlSpider):
    name        = "swedint"
    allowed_domains = ["swedsoft.se"]
    start_urls  = ["http://swedsoft.se/kalender/kalendarium/"]

    rules = (
        Rule(SgmlLinkExtractor(allow = (), restrict_xpaths=('//*[@id="main"]/div/div/')),
         callback="parse", follow = True),
    )

    def parse(self, response):
        i = 0

        for div in response.xpath('//*[@id="main"]/div/div/a'):
            item = AfeventItem()

        #Store data into lists
            item['title'] = div.xpath('//*[@class ="h2 entry-title"]/text()').extract()[i]
            item['location'] = div.xpath('//*[@id]/div[3]/h2/text()').extract()[i]
            item['venue'] = div.xpath('//*[@id]/div[3]/table/tr[1]/td[2]/text()[1]').extract()[i]
            #the following code changes the format of the date
            origDate = div.xpath ('//*[@id]/div[3]/p/text()').extract()[i]
            #split up the text in the date
            newDate = origDate.split()

            #handles if date is between two dates, e.g. "10 - 11 maj 2016"
            if len(newDate) > 3:
                rightDate = []
                rightDate.extend((newDate[0], newDate[3], newDate[4]))
                newDate = rightDate

            #Assign values to month names
            month = ["", "januari", "februari", "mars", "april", "maj", "juni", "juli",
             "augusti", "september", "oktober", "november", "december"].index(newDate[1])

```

```
            #Assign a "0" in the beginning if month number is < 10
            if month < 10:
                zeroMonth = [0, month]
                zeroMonth = ''.join(map(str, zeroMonth))
            else:
                zeroMonth = month

            #same thing as above with day
            if int(newDate[0]) < 10:
                zeroDate = [0, newDate[0]]
                zeroDate = ''.join(map(str, zeroDate))
            else:
                zeroDate = newDate[0]

            #Puts everything together and stores into item['date']
            finalDate = [newDate[2], zeroMonth, zeroDate]
            item['date'] = '-'.join(finalDate)

            item['host'] = "Swedsoft"
            item['description'] = div.xpath('//*[@id]/div[2]/section/p/text()').extract()[i]
            item['url'] = div.xpath('//*[@id="main"]/div/div/a/@href').extract()[i]
            follow_url = div.xpath('//*[@id="main"]/div/div/a/@href').extract()[i]
            request = Request(follow_url, callback = self.parse_second)
            request.meta['item'] = item

            if i < len(response.xpath('//*[@id="main"]/div/div/a')):
                i = i + 1
            yield request


    def parse_second(self, response):
        item = response.meta['item']
        item ['description'] = ''.join(response.xpath('//*[@id]/section/p//text()').extract())
        yield item
```

Kod från pipeline:

```
# -*- coding: utf-8 -*-

import pymongo
import json

from scrapy.conf        import settings
from scrapy.exceptions  import DropItem
from pymongo            import MongoClient


class AfeventPipeline(object):

    def __init__(self):
#Connect to Mongodb
        connection = pymongo.MongoClient(
            settings['MONGODB_SERVER'],
            settings['MONGODB_PORT']
        )
        db = connection[settings['MONGODB_DB']]
        self.collection = db[settings['MONGODB_COLLECTION']]

#Using the json file consisting of keywords
        with open("keywords_final.json") as json_file:
            global json_data
            json_data = json.load(json_file)

#Using the json file consisting of types of events
        with open("type.json") as json_file:
            global json_data2
            json_data2 = json.load(json_file)

    def process_item(self, item, spider):
#List in order to correct misspelled locations
        locationDict = {'Gothenburg' : 'Göteborg', 'Göteborg ': 'Göteborg',
                        'Malmo': 'Malmö', 'ÖstersundSenior': 'Östersund', 'Stockholm ': 'Stockholm'}

#Drop item if it is already in the database
        if self.collection.find_one({'url': item['url']}):
            raise DropItem('Item already in DB')
        else:
            description = item['description'].lower()
            title = item['title'].lower()
            item['tags'] = []
            item['type'] = []


#Go through list of words in the json files in order to assign each event with a type and with tags
            for type1_word in json_data2:
                for type2_word in json_data2[type1_word]:
                    if type2_word.lower() in description or type2_word.lower() in title:
                        if type1_word not in item['type']:
                            item['type'].append(type1_word)

            for level1_word in json_data:
                for level2_word in json_data[level1_word]:
                    if level2_word.lower() in description or level2_word.lower() in title:
                        if level1_word not in item['tags']:
                            item['tags'].append(level1_word)

            for key, value in locationDict.iteritems():
                if key == item['location']:
                    item['location'] = value
                    item['location'] = item['location'].strip()
            self.collection.insert(dict(item))

        for data in item:
          if not data:
              valid = False
              raise DropItem("Missing {0}!".format(data))
        return item
```