BENEFITS OF APPLYING COLLABORATIVE FILTERING ON STEAM PLATFORM
– AN EXPLORATIVE STUDY

Examensarbete Systemarkitekturutbildningen

Martin Bergqvist
Jim Glansk

HT 2016:KSAIXX

UNIVERSITY OF BORÅS
SCHOOL OF BUSINESS AND IT

Systemarkitekturutbildningen är en kandidatutbildning med fokus på programutveckling. Utbildningen ger studenterna god bredd inom traditionell program- och systemutveckling, samt en spets mot modern utveckling för webben, mobila enheter och spel. Systemarkitekten blir en tekniskt skicklig och mycket bred programutvecklare. Typiska roller är därför programmerare och lösningsarkitekt. Styrkan hos utbildningen är främst bredden på de mjukvaruprojekt den färdige studenten är förberedd för. Efter examen skall systemarkitekter fungera dels som självständiga programutvecklare och dels som medarbetare i en större utvecklingsgrupp, vilket innebär förtrogenhet med olika arbetssätt inom programutveckling.

I utbildningen läggs stor vikt vid användning av de senaste teknikerna, miljöerna, verktygen och metoderna. Tillsammans med ovanstående teoretiska grund innebär detta att systemarkitekter skall vara anställningsbara som programutvecklare direkt efter examen. Det är lika naturligt för en nyutexaminerad systemarkitekt att arbeta som programutvecklare på ett stort företags IT-avdelning, som en konsultfirma. Systemarkitekten är också lämpad att arbeta inom teknik- och idédrivna verksamheter, vilka till exempel kan vara spelutveckling, webbapplikationer eller mobila tjänster.

Syftet med examensarbetet på systemarkitekturutbildningen är att studenten skall visa förmåga att delta i forsknings- eller utvecklingsarbete och därigenom bidra till kunskapsutvecklingen inom ämnet och avrapportera detta på ett vetenskapligt sätt. Således måste de projekt som utförs ha tillräcklig vetenskaplig och/eller innovativ höjd för att generera ny och generellt intressant kunskap.

Examensarbetet genomförs vanligen i samarbete med en extern uppdragsgivare eller forskningsgrupp. Det huvudsakliga resultatet utgörs av en skriftlig rapport på engelska eller svenska, samt eventuell produkt (t.ex. programvara eller rapport) levererad till extern uppdragsgivare. I examinationen ingår även presentation av arbetet, samt muntlig och skriftlig opposition på ett annat examensarbete vid ett examinationsseminarium. Examensarbetet bedöms och betygssätts baserat på delarna ovan, specifikt tas även hänsyn till kvaliteten på eventuell framtagen mjukvara. Examinator rådfrågar handledare och eventuell extern kontaktperson vid betygssättning.

HÖGSKOLAN I BORÅS
INSTITUTIONEN HANDELS- OCH IT-HÖGSKOLAN

BESÖKSADRESS: JÄRNVÄGSGATAN 5 · POSTADRESS: ALLÉGATAN 1, 501 90 BORÅS
TFN: 033-435 40 00 · E-POST: INST.HIT@HB.SE · WEBB: WWW.HB.SE/HIT

Svensk titel: Fördelar med att applicera Collaborative Filtering på Steam – En utforskande studie

Engelsk titel: Benefits of Applying Collaborative Filtering on Steam – An explorative study

Utgivningsår: 2016

Författare: Martin Bergqvist, Jim Glansk

Handledare: Henrik Linusson

Abstract
(på engelska)

CSL@BS

Keywords: Collaborative filtering, Recommender system, Steam, Computer Science, Machine learning.

Sammanfattning
(på svenska)

Nyckelord: Collaborative filtering, Rekommendationssystem, Steam, Datorvetenskap, Maskininlärning.

Innehållsförteckning

1 Inledning

Datorspel har blivit en djupt rotad företeelse i vår kultur vilket tydligt visas genom att spelindustrin ständigt växer sig större. Spelindustrin omsatte 91.8 miljarder USD 2015 och växer mellan 6-9 procent varje år (Newzoo 2016). En av anledningarna till framgången är digital distribution och ökad bandbreddskapacitet, vilket gör att spel på ett enkelt sätt kan köpas och laddas ner över nätet (Toivonen & Sotamaa 2010).

En av de största aktörerna på marknaden idag är Steam, som har över 125 miljoner aktiva användarkonton (Kotaku 2015), och över 10 000 speltitlar (Steam 2016). Steam hade 2011 50-70% av marknaden för digital spelförsäljning till PC (Forbes 2011). Under 2015 hade spel för 3,5 miljarder USD sålts via Steam. Ca. 25% av spelen som såldes har aldrig spelats (Medium 2016), och ytterligare 19% av spelen har spelarna ägnat inte mer än en timmas speltid åt (Ars Technica, 2014).

E-handel är en distributionsform som antagit webben med storm där interaktion mellan kunder och produkter ständigt växer större informationsbaser. För att tillgodose kunder att skapa sammanhang i all information så vänder sig fler företag mot att personifiera kundbesöken genom rekommendationer.

Netflix annonserade 2006 en tävling, där man lovade 1 miljon USD i prispengar till den som lyckades finna ett sätt att optimera deras rekommendationer med 10 procent. När man avslutade tävlingen 2009 hade man nått en optimeringsgrad på 10.6 procent och engagerat över 41305 teams från 186 länder (Netflix Price 2016).

Rekommendationssystem är till för att hitta relevanta biprodukter för den specifika användaren, eller, om man så vill, erbjuda direktriktad reklam/information. I publikationen om Tapestry, ett av de tidigaste rekommendationssystemen, kan man läsa om hur det explosivt ökande e-mailanvändandet i början på 90-talet har skapat ett behov av att filtrera informationsflödet, för att överhuvudtaget kunna ha nytta all tillgänglig information. De resonerar att istället för att registrera sig till en maillista, för att få alla mail relevanta för den listan, ska man ha ett filter som söker alla maillistor efter de dokument som är personligt relevanta. Detta ska då åstadkommas genom att användare rekommenderar dokument de finner relevanta, och låter systemet hitta användare med liknande intressen. Detta kallar de *collaborative filtering (CF)* (Goldberg, Nichols, Oki & Terry 1992).

Rekommendationssystem kan titta på flera sorters input. Där det finns, är explicit feedback (typiskt betyg angivet av användare, recensioner och liknande aspekter) det mest användbara; Om användaren värderar produkten ger det mycket pålitlig data. I de fall där explicit feedback saknas, kan istället implicit feedback (t.ex. användningsfrekvens, sökmönster och annat som helt enkelt *antyder* att användaren har värderat någonting) nyttjas, som indirekt påvisar korrelation genom observation av användarens beteendemönster (Oard 1998).

Rekommendationssystem kan vara byggda enligt CF, men andra principer förekommer. Här menas vanligtvis antingen *Content-based filtering (CBF)*; Att filtrera efter liknande produkter, eller ha fler aspekter än vad som ryms i CF (Hybrid mellan CF och CBF).

CBF går enkelt uttryckt ut på att klassificera alla produkter så noggrant som möjligt och väga dessa taggar mot användarens angivna och/eller historiska preferenser, för att rekommendera de produkter som liknar de användaren gillat tidigare. Steam använder sig av detta system i grunden för sina rekommendationer, där användarens köpta spel, kombinerat med användarens vänners spel, utgör basen för rekommendationerna (Steam 2014). Fördelarna med det här systemet är att resultaten är enkla för användaren att förstå, det är billigt att använda gällande beräkningsmängder, och det är möjligt med personliga resultat även utan en användarprofil. Nackdelarna å sin sida blir att resultaten blir ytliga och uppenbara (Shi 2008).

CF arbetar utifrån principen att liknande användare gillar liknande saker, och analyserar beteendemönster för att sedan hitta de användare som ligger närmast i beteende, och använda deras preferenser som rekommendationer. LinkedIn är ett bra exempel, där du rekommenderas nya företag att följa baserat på vad liknande profiler följer (Wu, Shah, Choi, Tiwari & Posse 2014). Fördelarna med CF är att det fungerar mot vilken domän som helst & "lyckoträffar", när CF hittar till synes orelaterat material av intresse. Nackdelarna rör främst skalbarheten, då en liten databas kommer att ge svaga kopplingar; allmänt kallat *kallstart*, och en omfattande databas kommer ge starka kopplingar, men använda mer beräkningsresurser (Shi 2008). Det anses allmänt vara så, att CF är effektivare än CBF på befintliga datamängder, där man kan bortse från kallstarts-aspekten.

Hybrid Filter är ett samlingsnamn för de system som på något sätt kombinerar både CBF & CF. Detta går att göra på många olika sätt, vilket Netflix är ett utmärkt exempel på, då de 2007 använde en lösning baserad på 107 algoritmer för att rekommendera film (Bell, Koren & Volinsky 2007). Fördelen med att kombinera olika lösningar blir att man kombinerar styrkorna från varje dellösning, och kompenserar i viss mån svagheterna (Shi 2008).

1.1 Problemdiskussion

På större system kan man med fördel använda just Hybrid filtering, men för ett optimalt resultat krävs, som i exemplet med Hybridlösningen tillämpad av Netflix, en mycket avancerad kombination av algoritmer, varför vi avgränsar vårt arbete till att se om vi kan ge bättre prediktioner för lämpliga spel på Steam med hjälp av CF, som är en mer mångsidig och exakt metod för att utvärdera mer implicit data mellan olika användare.

1.2 Problemformulering & Syfte

Då det är ett faktum att en stor andel av spelen som säljs via Steam faktiskt inte spelas, eller spelas mycket lite, konstateras att det finns ett allmän-intresse i att undersöka potentialen i ett system baserat på det tillvägagångssätt som konstaterats vara lämpligare på omfattande databaser.

Syftet med denna rapport är således att svara på frågan "Kan man med CF överträffa CBF på en databas som Steams?"

2 Teori

[Placeholder]

3 Metod

Det här kapitlet ämnar erbjuda en översikt av forskningsstrategi och metodologi relevant för studien. Hur empirin har insamlats, bearbetats och analyserats kommer förklaras, och därefter kommer reflektioner över eventuella etiska aspekter och andra möjliga tillvägagångssätt presenteras. För att uppfylla studiens forskningsmål har data insamlats automatiskt, med hjälp av en egenutvecklad artefakt, varför design science med en kvantitativ analysmetod har ansetts lämpligt.

3.1 Forskningsstrategi

För att besvara vår frågeställning följde vi en experimentell approach där avsikten är att experimentera med olika features för att hitta en optimal CF, vilket faller väl inom ramarna för definitionen av den experimentella strategin (Robson 2011).

3.2 Datainsamling

Genom Valves Steam's öppna API har det varit fri access till den primärdata som legat till grund för studien, där vi fått skapa en enkel implementation för att samla in de SteamID:n som krävts för de användarattribut (ägda spel, speltid, troféer) som var aktuellt. För insamling av nycklar användes snöbollseffekten via en användares vänlista, vilket på ett smidigt sätt genererade en uppsättning om ca 1000 SteamID:n. Eftersom studien inte kräver en demografisk representation för att uppnå svar, har den aspekten åsidosatts.

SteamID:n som samlats in användes sedan för att göra förfrågningar mot Steam API:t i syfte att utvinna primärdata. Utöver den primärdata som samlades in krävdes även insamling av sekundärdata, i form av användarrecensioner, vilket användes som referensdata mot våra preciseringar. För insamling av referensdata använde vi oss av enkäter där användare fick värdera spel på en skala mellan 1-5 och uppge sitt SteamID.

3.3 Dataanalys

Via byggd artefakt har data processerats, där känd data dolts i ML-syfte enligt CF-matrix-factorization, och jämförts mot dold data. För att kunna analysera den dataoutput som vår hearustik predicerar för en specifik användare krävdes att jämförelser gjordes mellan denna dataoutput och den faktiska betygsättning användaren gjort för ett specifikt spel vilket framgår i sekundärdatan. Vår dataoutput är då implicit data omvandlat till predicerad explicit data som jämförs med den faktiska explicita datan betygsatt av användaren. För att åstadkomma det här så krävdes att vi plocka bort de spel som explicit har betygsatts av användaren genom enkätinsamlingen.

Eventuell annorlunda representation av s.k. dolda profiler får utebli från studien, då datan är otillgänglig, och det inte finns någon anledning att bedöma köpmönstren som signifikant avvikande.

3.4 Etiska aspekter

Under enkäten ombads individer uppge sitt SteamID. Detta kan tänkas kännas som ett intrång i den privata integriteten, men då inga personliga data omfattas av heuristiken, och nytta endast kan dras av sådana användare som har en publik profil, där all sådan information redan ligger öppen, bedöms intrånget som acceptabelt. Åtgärder har dessutom tagits för att försäkra att enkät-deltagarna är insatta i innebörden av enkät-resultaten, och därefter haft möjlighet att avstå helt.

4 Problem and Requirements

This part provides an elaborated description and analysis of the practical problem addressed, possibly including a root cause analysis. It also defines the requirements on the artefact, which are justified based on the problem analysis. The part can also describe the processes of problem analysis and requirements elicitation, in particular, the application of the selected research strategies and methods and the use of the knowledge base.

5 Artefact

This part describes the artefact. It explains the structure, behaviour, and function of the artefact, preferably with examples. This part can also describe the development process, including the design alternatives and design rationale as well as the knowledge base used. This part is often the main part of a design science paper.

6 Evaluation

This part describes how the artefact has been evaluated. It describes the evaluation strategy and the evaluation process, especially how the selected research strategies and methods were applied. This part is often a large part of a design science paper but can be smaller if a highly innovative artefact has been developed. Sometimes, this part only describes a demonstration of the artefact.

7 Discussion

This part reflects on the research carried out and its contributions. It identifies limitations in the study, discusses the novelty and value of the artefact compared to existing ones, outlines the practical and theoretical significance of the contributions, and discusses ethical aspects of the use of the artefact. It can also suggest areas for future research. Optionally, this part can reflect on design science and the application of the method framework.

8 Conclusion

This part ties the paper together and shows how the research questions have been answered (or how the research goals have been achieved).

Referenser

Bell, R. M., Koren, Y., & Volinsky, C. (2007). *The BellKor solution to the Netflix prize*.

Forbes (2011). *The Master of Online Mayhem*. http://www.forbes.com/forbes/2011/0228/technology-gabe-newell-videogames-valve-online-mayhem.html [2016-11-24]

Goldberg, D., Nichols, D., Oki, B. & Terry, D. (1992). Using collaborative filtering to weave an information tapestry, *ACM*, NEW YORK.

Netflix (2009). *Netflix Prize*. http://www.netflixprize.com/leaderboard.html [2016-11-25]

Newzoo (2016). *The Global Games Market Reaches $99.6 Billion in 2016*. https://newzoo.com/insights/articles/global-games-market-reaches-99-6-billion-2016-mobile-generating-37 [2016-11-25]

Oard, D.W.,Kim, J. (1998). Implicit Feedback for Recommender Systems, *AAAI Technical Report WS-98-08*.

Robson, C. (2011). *Real world research: a resource for users of social research methods in applied settings.* Chichester: Wiley.

Shi, C. (2008). Exploring Movie Recommendation System Using Cultural Metadata, *IEEE*, ss.431.

Steam (2014). *Recommendation Feed*. http://store.steampowered.com/about/newstore?l=swedish [2016-11-24]

Toivonen, S. & Sotamaa, O. (2010). Digital distribution of games: the players' perspective. *Proceedings of the International Academic Conference on the Future of Game Design and Technology, ACM*, New York, ss. 199-206.

Wu, L., Shah, S., Choi, S., Tiwari, M. & Posse, C. (2014). *The browsemaps: Collaborative filtering at LinkedIn*.

Högskolan i Borås är en modern högskola mitt i city. Vi bedriver utbildningar inom ekonomi och informatik, biblioteks- och informationsvetenskap, mode och textil, beteendevetenskap och lärarutbildning, teknik samt vårdvetenskap.

På **institutionen Handels- och IT-högskolan (HIT)** har vi tagit fasta på studenternas framtida behov. Därför har vi skapat utbildningar där anställningsbarhet är ett nyckelord. Ämnesintegration, helhet och sammanhang är andra viktiga begrepp. På institutionen råder en närhet, såväl mellan studenter och lärare som mellan företag och utbildning.

Våra **ekonomiutbildningar** ger studenterna möjlighet att lära sig mer om olika företag och förvaltningar och hur styrning och organisering av dessa verksamheter sker. De får även lära sig om samhällsutveckling och om organisationers anpassning till omvärlden. De får möjlighet att förbättra sin förmåga att analysera, utveckla och styra verksamheter, oavsett om de vill ägna sig åt revision, administration eller marknadsföring. Bland våra **IT-utbildningar** finns alltid något för dem som vill designa framtidens IT-baserade kommunikationslösningar, som vill analysera behov av och krav på organisationers information för att designa deras innehållsstrukturer, bedriva integrerad IT- och affärsutveckling, utveckla sin förmåga att analysera och designa verksamheter eller inrikta sig mot programmering och utveckling för god IT-användning i företag och organisationer.

Forskningsverksamheten vid institutionen är såväl professions- som design- och utvecklingsinriktad. Den övergripande forskningsprofilen för institutionen är handels- och tjänsteutveckling i vilken kunskaper och kompetenser inom såväl informatik som företagsekonomi utgör viktiga grundstenar. Forskningen är välrenommerad och fokuserar på inriktningarna affärsdesign och Co-design. Forskningen är också professionsorienterad, vilket bland annat tar sig uttryck i att forskningen i många fall bedrivs på aktionsforskningsbaserade grunder med företag och offentliga organisationer på lokal, nationell och internationell arena. Forskningens design och professionsinriktning manifesteras också i InnovationLab, som är institutionens och Högskolans enhet för forskningsstödjande systemutveckling.

HÖGSKOLAN I BORÅS
VETENSKAP FÖR PROFESSION

BESÖKSADRESS: JÄRNVÄGSGATAN 5 · POSTADRESS: ALLÉGATAN 1, 501 90 BORÅS
TFN: 033-435 40 00 · E-POST: INST.HIT@HB.SE · WEBB: WWW.HB.SE/HIT