Jämförelse av kollaborativa, innehållsbaserade och sociala rekommendationssystem

Fredrik Fagerholm
37164
fredrik.fagerholm@abo.fi
Kandidatavhandling i datavetenskap
Handledare: Mats Neovius
Fakulteten för naturvetenskaper och teknik
Åbo Akademi
4 april 2016

Referat

Rekommendationssystem omfattar sådana verktyg och tekniker som används för att hitta produkter en användare kan uppfatta som intressanta. Rekommendationerna som systemet ger är menade att fungera som stöd vid beslut som användaren måste göra angående produkter. Den här avhandlingen behandlar några av de vanligaste typerna av rekommendationssystem samt en mindre vanlig metod, som är sociala rekommendationssystem. De typer av system som diskuteras är kollaborativa, innehållsbaserade och sociala rekommendationssystem. I avhandlingen redogörs för hur dessa system fungerar samt vilka fördelar och brister som finns hos vardera system. Vidare undersöks om sociala rekommendationssystem kan prestera bättre än de två andra typerna av system.

Nyckelord: Rekommendationssystem, Kollaborativ filtrering, Innehållsbaserad filtrering, Social filtrering

Innehållsförteckning

1 Inledning

Med rekommendationssystem avses verktyg och tekniker som genom olika metoder försöker göra förutsägelser om användares preferenser för produkter inom någon domän. Termen produkter syftar här inte endast på fysiska produkter eller tjänster, utan den har en bredare betydelse som förklaras mer ingående i stycke 3.2. Systemen använder förutsägelserna när de rekommenderar en produkt för en användare. För att göra förutsägelserna används data om produkterna och användarna. Systemet försöker hitta samband och mönster i data och utnyttja dessa för att göra förutsägelserna. Den här avhandlingen ger en överblick över de metoder som används i rekommendationssystem och hur data om relationer mellan användarna kan användas inom dessa system.

De två huvudsakliga kategorierna av rekommendationssystem är innehållsbaserade (eng. content-based) och kollaborativa (eng. collaborative) [1, 2]. En tredje kategori utgörs av system som skapar rekommendationer för en viss användare genom att använda sig av preferenser hos andra användare som har någon form av relation till den användare rekommendationen är riktad till. Dessa kallas sociala eller tillitsbaserade (eng. trust-enhanced) rekommendationssystem. Det finns indikationer på att sociala rekommendationssystem kan ge bättre resultat än de andra typerna av system under vissa förutsättningar. De är följaktligen av intresse och undersöks i denna avhandling.

Syftet med avhandlingen är att förklara hur system i dessa tre kategorier fungerar och skriva om deras fördelar och brister. Sociala rekommendationssystem kommer att behandlas i större detalj och det undersöks hur dessa kan användas för att överkomma brister hos innehållsbaserade och kollaborativa system.

Metoder som används för att skapa förutsägelser är analys av historiska data, till exempel genom statistisk analys och maskininlärning. Båda metoder används för att hitta samband. Man förlitar sig på att dessa samband, åtminstone i någon grad, kommer att hålla också i framtiden. Trots att vissa samband möjligtvis håller så är användarnas beteende ofta kaotiskt i den mening att framtida beteendemönster är svåra att förutsäga utgående från information om det förflutna.

Rekommendationssystem används bland annat inom näthandel, tjänster för strömning av film och musik, sökmotorer och marknadsföring. I en nätbutik kan rekommendationer ges till exempel utgående från vad användaren köpt tidigare, vilka produkter liknande användare har köpt, de mest köpta produkterna överlag och så vidare. För de andra tjänsterna används motsvarande data, så som tidigare spelade låtar eller de mest populära låtarna. Vilken typ av data som används och hur den används varierar beroende på domän. Något som fungerar inom ett visst domän behöver nödvändigtvis inte fungera inom ett annat, till exempel är det mer troligt att en användare lyssnar på en låt upprepade gånger medan det däremot är mindre troligt att någon köper flera stycken bilar. I det exemplet skulle det intuitivt sett vara rimligt att anta att upprepade spelningar av en låt tyder på att användaren gillar låten, och således kan ett system för låtrekommendationer använda sig av detta antagande.

Rekommendationssystem utvecklas och används på grund av den potentiellt stora ekonomiska nytta de skapar genom att förbättra användarupplevelsen av en tjänst, och effektivera marknadsföringen. En tjänst som ger korrekta rekommendationer åt sina användare uppfattas troligen som överlägsen en tjänst som saknar denna funktionalitet. På grund av det enorma utbudet av innehåll hos strömningstjänster för media eller antalet produkter i en nätbutik, så underlättar rekommendationer då användaren skall göra ett val. Det är mer sannolikt att en användare väljer en produkt den är nöjd med om en liten välvald mängd av produkter presenteras för användaren, i motsats till om hela sortimentet visas.

2 Motivering

Rekommendationssystem är verktyg och tekniker som ger användare förslag om produkter. Förslagen är relaterade till beslutsfattande som användaren gör angående produkter. Typen av beslut och produkt varierar beroende på domän. Tillämpningsområdena är många, det kan exempelvis handla om ett köpbeslut i en nätbutik, val av låt i en musikströmningstjänst eller ett val bland resultat som ges av en sökmotor.

Rekommendationssystem kan spela en mängd olika roller. Systemen kan vara till nytta både för parten som erbjuder en tjänst och slutanvändaren. Det främsta målet för den som erbjuder tjänsten är i allmänhet att öka omsättningen för företaget eller organisationen, medan slutanvändaren strävar efter att hitta en så bra och lämplig produkt som möjligt.

Människor tenderar att använda sig av, och ofta lita på, rekommendationer de får av andra. De använder sig av rekommendationerna när de skall fatta beslut [3]. Till exempel är det troligt att man väljer en bok som en vän rekommenderat över en annan bok man inte vet något om. Vidare väljer man troligen hellre filmer, musik och restauranger som klarat sig bra i recensioner eller fått höga betyg. I förlängningen antas detta gälla även för rekommendationer som är gjorda av ett system. Detta är ett grundantagande som görs för att motivera användningen av rekommendationssystem. Det vill säga att användarna kommer att ta i beaktande rekommendationer som är gjorda av ett system då de gör ett beslut, givet att systemet ger pålitliga rekommendationer.

3 Termer

Systemen kommer till stora delar att behandlas på ett allmänt plan, inte för någon specifik tillämpning. Detta kräver att man definierar de grundläggande begrepp som är nödvändiga för att behandla dem generellt. I detta stycke förklaras de allmänna begreppen som används i avhandlingen.

3.1 Användare

Med användare avses de agenter som rekommendationerna görs för. Systemen använder data om användarna för att göra dessa rekommendationer. Dessa data utgör en modell för användaren och data om en användare representerar den användaren i systemet. Vilken slags data som används och strukturen hos den varierar beroende på domän och vilken typ av rekommendationssystem det är frågan om. I stycke 4 beskrivs några typiska slag av data och strukturen hos dem.

Det antas att majoriteten av användarna är rationella och välvilliga. Därför kan man lita på att deras interaktioner med systemet speglar vad de verkligen tycker. Därmed kan man använda data som samlas in om dessa interaktioner för att skapa en verklighetstrogen modell av användarna. Det kan dock förekomma användare som avsiktligt interagerar med systemet på ett sätt som inte motsvarar deras verkliga preferenser. Man kan ta åtgärder för att motverka de negativa effekter det har på systemet och på så sätt skapa ett mer robust system.

3.2 Produkter

Med produkter avses i den här avhandlingen de objekt som rekommenderas åt användaren. Det behöver nödvändigtvis inte vara fysiska objekt, utan det kan handla om tjänster, artiklar, hemsidor, tips för investering och försäkringar. Ett system är ofta anpassat för ett specifikt domän. Teknikerna systemet använder är då också anpassade för egenskaper hos den typ av produkt som systemet skapar rekommendationer för.

3.3 Betygsättning

Med betyg avses de preferenser användarna visat för någon produkt. Betygen kan vara explicita, till exempel om användaren gett ett konkret betyg på en skala eller “gillat” en produkt. Betygen kan också vara implicita. Då tolkar man användarens handlingar och försöker avgöra om de tyder på preferens för en produkt. Till exempel om användaren öppnat sidan för en produkt i en nätbutik kan det tolkas som att användaren visat intresse för produkten.

4 Tekniker

Genom att använda data om användarna och produkterna vill man räkna ut en rankning av varje rekommenderbar produkt. Man söker efter en funktion som givet en viss användare och produkt ger ett värde som representerar produktens relevans för användaren. Värdena skall gå att ordna så att man kan välja produkterna av högsta relevans för användaren. Då kan produkterna ordnas så att man får en rankning där de som är högst upp är de produkter som är mest relevanta och som såvida är bra kandidater för rekommendation. Eftersom systemen ”sållar fram” de mest relevanta produkterna i hänsyn till någon användare så kallas metoderna de använder ofta filtrering. Mer specifikt kollaborativ, innehållsbaserad samt social filtrering.

Det är vanligt att lagra data om produkter och användare i matrisform. I matrisen motsvarar kolumnerna produkter och raderna användare. För *n* stycken användare och *m* stycken produkter fås en $n\times m\text{-}matris$. Ett element i rad *i* och kolumn *j* representerar då användare *i*:s betyg för produkt *j*, alternativt är elementet det uträknade preferensvärdet. Om varken betyg eller preferensvärdet finns att tillgå har elementet något specifikt värde som indikerar att betyget är okänt. I figur 1 visas en sådan matris. Matrisen har 4 användare och 5 produkter och är således en $4\times5\text{-}matris$. Betygen är heltalsvärden mellan 1 och 10, till exempel har *användare 2* gett betyget 6 för *produkt 4*. Tomma positioner indikerar att användaren inte ännu gett ett betyg för den produkten.

	$produkt\ 1$	$produkt\ 2$	$produkt\ 3$	$produkt\ 4$	$produkt\ 5$
$användare\ 1$	5		7		8
$användare\ 2$		9		6	
$användare\ 3$	4		7		
$användare\ 4$	9		1	9	4

Figur 1. En användar–produkt matris.

Varje användare kan ses som en m-dimensionell vektor där elementen i vektorn är de betyg de har gett åt produkterna. På motsvarande sätt kan en produkt ses som en n-dimensionell vektor där elementen är de betyg produkten fått.

5 Kollaborativa system

Grundidén bakom kollaborativa system är att använda sig av data om hela eller delar av användarbasen för att producera rekommendationer för en enskild användare. Om man vill göra en rekommendation för en viss användare så söker man först efter användare som liknar målanvändaren och skapar en grupp av sådana användare. Sedan ser man på betygen som användare i gruppen gett åt sådana produkter som målanvändaren inte betygsatt. Utgående från dessa betyg försöker man räkna ut ett värde som förhoppningsvis motsvarar ett betyg målanvändaren skulle ge åt produkterna. Därefter kan man rekommendera sådana produkter vars uträknade betyg överstiger något godtyckligt tröskelvärde.

Man antar att det är möjligt att hitta en grupp av användare vars tidigare preferenser sammanfaller med en viss målanvändare. Förutsatt det så antar man också att deras preferenser kommer att likna varandra i framtiden. Under dessa antaganden så kan man använda gruppens preferenser för att förutsäga målanvändarens preferens [3].

5.1 Mått för likhet mellan användare

För att räkna ut ett värde för en användare och produkt måste man först bestämma hur man mäter likheten mellan användare [2]. Om man utgår från att varje användare representeras av en vektor bestående av de betyg de gett åt produkterna, så finns det ett antal mått som vanligen används inom system av denna typ. Måtten baserar sig på distans eller vinkel mellan vektorerna, alternativt mäter man korrelationen mellan värdena i vektorerna [4]. Här presenteras tre stycken ofta använda mått på likhet mellan användare.

5.1.1 Euklidiskt avstånd

Ett av de enklaste och mest intuitiva måtten på distans är det euklidiska avståndet. För två punkter är det längden av en linje mellan dem. För två m-dimensionella vektorer x och y fås det euklidiska avståndet d av x och y genom formeln i ekvation (1). Värdet

nummer k i vektorerna betecknas x_k respektive y_k. Varje värde i y subtraheras från varje värde i x, kvadreras och summeras. Sedan tas kvadratroten av summan [4].

$$d(x,y) = \sqrt{\sum_{k=1}^{m}(x_k - y_k)^2} \tag{1}$$

Euklidiskt avstånd

5.1.2 Vinkel mellan vektorer

Cosinus av vinkeln mellan två vektorer kan användas som ett mått för likhet mellan användarna de representerar. En liten vinkel mellan vektorerna motsvarar en stor likhet medan en stor vinkel motsvarar en liten likhet. Eftersom man beräknar cosinus av vinkeln så är värdena mellan -1 och 1. Värdet 1 tolkas som total likhet och -1 som total olikhet mellan användarna. Cosinus av vinkeln α mellan vektorerna x och y beräknas enligt formeln i ekvation (2). I figuren betecknar $x \cdot y$ skalärprodukten mellan vektorerna x och y och $\|x\|\|y\|$ betecknar produkten av normerna av vektorerna [4]. Funktionen som mäter likheten mellan två användare kallas här sim från engelskans ”similarity”.

$$sim(x,y) = cos(\alpha) = \frac{x \cdot y}{\|x\|\|y\|} \tag{2}$$

Cosinus av vinkeln mellan vektorer

5.1.3 Korrelation

Korrelationen mellan två vektorer anger hur deras värden förhåller sig till varandra. Det mest använda måttet för korrelation är Pearsons korrelationskoefficient [2] och betecknas med ett r. Pearsons korrelationskoefficient mäter det linjära beroendet mellan datapunkter. När man beräknar r för två vektorer får man ett värde mellan -1 och 1. Värdet -1 innebär fullständig negativ korrelation, värdet 0 avsaknad av korrelation och 1 fullständig positiv korrelation. I likhet med cosinus av vinkeln mellan vektorerna så

tolkar man värdet 1 som total likhet och -1 som total olikhet mellan användarna. Korrelationskoefficient r för vektorerna x och y beräknas enligt formeln i ekvation (3). I figuren representerar $\bar{x}$ och $\bar{y}$ det aritmetiska medelvärdet av värdena i respektive vektor.

$$r(x,y) = \frac{\sum_{k=1}^{m}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_k - \bar{x})^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}} \quad (3)$$

Pearsons korrelationskoefficient

5.1.4 Krav hos måtten

I de presenterade måtten för likhet är det vissa detaljer som måste tas i beaktande vid beräkningarna. Som tidigare påpekats är det möjligt att varje användare inte betygsatt varje produkt. När likheten mellan två användare beräknas kan den ena användaren alltså ha gett ett betyg för en produkt som den andra inte betygsatt, och detta måste hanteras av de algoritmer som utför beräkningarna. Till exempel i matrisen i figur 1 har $användare\ 3$ och $användare\ 4$ båda gett betyg åt $produkt\ 1$ och $produkt\ 3$. Däremot har $användare\ 4$ betygsatt $produkt\ 4$ vilket $användare\ 3$ inte har gjort. Om likheten för dessa två användare skall beräknas så är man tvungen att välja hur man gör med de värden som saknas. Ett alternativ är att utelämna de par där någotdera värde saknas. I beräkningarna för $användare\ 3$ och $användare\ 4$ skulle man enligt denna metod endast använda betygen de gett för $produkt\ 1$ och $produkt\ 3$. Likhetsvärdet mellan användare som endast har en gemensam betygsatt produkt saknar betydelse [2], värdet är missvisande och bör därför inte användas. Till exempel $användare\ 2$ och $användare\ 4$ har endast $produkt\ 4$ som båda har betygsatt. Formeln i ekvation (2) ger värdet 1 för dessa användare vilket motsvarar fullständig likhet mellan användarna, trots att de gett olika betyg åt produkten.

5.2 Utförande

Om man vill räkna ut preferensvärdet för ett användar–produkt-par som från tidigare saknar ett värde så försöker man hitta en mängd användare som är så lika som

målanvändaren som möjligt. Två användare anses vara lika om de har gett liknande betyg åt samma produkter. Mängden av likande användare skall också ha gett ett betyg för den produkt som man beräknar preferensvärdet för. Man utgår från de betygen då man beräknar värdet som exempelvis kan vara ett medeltal av mängdens betyg för produkten. Olika modifikationer kan göras för att få ett bättre värde. Till exempel kan man ge högre vikt för betygen hos användare som är mer lika målanvändaren. Kollaborativa system gör inga antaganden om sociala relationer mellan användarna utan utgår endast ifrån hur användarna har interagerat med produkterna.

Den klassiska formeln för kollaborativ filtrering ges av ekvation (4) [2]. Funktionen f ger en förutsägelse för betyget användaren a skulle kunna ge åt produkt p. $\bar{x}_a$ är det aritmetiska medeltalet av användare a:s betyg. L är mängden av alla användare som liknar a, kriteriet för att en användare b skall tas med i L kan vara att likhetsvärdet mellan a och b, som kan beräknas med hjälp av någon av formlerna i ekvation (1), (2) eller (3), överstiger något förutbestämt tröskelvärde. $v_{a,b}$ är vikter som exempelvis kan vara de likhetsvärden som beräknats mellan a och b. $x_{b,p}$ är det betyg användare b gett åt produkt p. I formeln kan kvoten ses som ett viktat medeltal av betygen användare som liknar a har gett åt produkten p.

$$f(a,p) = \bar{x}_a + \frac{\sum_{b \in L} v_{a,b} \cdot (x_{b,p} - \bar{x}_a)}{\sum_{b \in L} v_{a,b}} \tag{4}$$

Förutsägelse av betyg.

Om man beräknar ett värde för $anv\ddot{a}ndare\ 3$ och $produkt\ 5$ genom att använda likhetsmåttet i ekvation (2) och utelämnar värden som fattas från beräkningarna, så går det till på följande sätt. Man ser på $anv\ddot{a}ndare\ 3$ och $anv\ddot{a}ndare\ 1$, tar de betygen för produkterna som båda har betygsatt alltså för $produkt\ 1$ och $produkt\ 3$, beräknar cosinus av vinkeln mellan dessa vektorer enligt formeln i ekvation (2) och får värdet $0{,}995$ (avrundat till tre decimaler). Motsvarande beräkningar för $anv\ddot{a}ndare\ 3$ och $anv\ddot{a}ndare\ 4$ ger $0{,}589$, $anv\ddot{a}ndare\ 2$ lämnas bort från beräkningarna eftersom den inte har betygsatt $produkt\ 5$. Dessa två tal är likhetsvärdena för paren av användare.

Om tröskelvärdet för likhetsvärdena satts till 0,50 kan båda användarna tas med i mängden av liknande användare. Formeln i ekvation (4) används för att beräkna ett möjligt värde för betyget. Det värdet är 6,539 och används som ett betyg $användare\ 3$ skulle kunna ge åt $produkt\ 5$. Eftersom betyget är relativt högt på skalan så kan $produkt\ 5$ anses vara en bra kandidat för rekommendation till $användare\ 3$. Det uträknade betygets värde är närmare 8 än 4 på grund av att $användare\ 1$ är mer lik målanvändaren än $användare\ 4$ enligt måttet som valts, betyget $användare\ 1$ gett ges mer vikt vid uträkningen av det slutliga värdet. I figur 5 visas uträkningarna som görs för att få det slutliga värdet.

$$sim(a3, a1) = \frac{a3 \cdot a1}{\|a3\|\|a1\|} = \frac{(4 \quad 7)\begin{pmatrix}5\\7\end{pmatrix}}{\sqrt{4^2+7^2}\sqrt{5^2+7^2}} = \frac{4*5+7*7}{\sqrt{65}\sqrt{74}} = \frac{69}{\sqrt{4810}} = 0{,}9948\ldots \approx 0{,}995$$

$$sim(a3, a4) = 0{,}5889\ldots \approx 0{,}589$$

$$f(användare\ 3, produkt\ 5) = 5{,}5 + \frac{0{,}995 * (8 - 5.5) + 0{,}589 * (4 - 5.5)}{0{,}995 + 0{,}589} = 6{,}5388\ldots \approx 6{,}539$$

Figur 2 Beräkning av förutsägelse.

5.3 Fördelar med kollaborativa system

En av fördelarna med kollaborativa metoder är att de inte kräver information om produkterna som rekommenderas [2] till skillnad från innehållsbaserade system som diskuteras i stycke 6. Algoritmerna använder sig endast av de betyg som användarna gett åt produkterna för att avgöra hur lika varandra användarna är och hur mycket en specifik användare skulle kunna gilla en viss produkt. Vidare så fördelas arbetsmängden för utvärderingen av produkterna mellan användarna, vilket utförs centralt i innehållsbaserade system [2].

6 Innehållsbaserade system

Innehållsbaserade system utnyttjar antagandet att användare som visat preferenser för en viss typ av produkt kommer att gilla liknande produkter. Det utförs genom att jämföra data om egenskaper hos produkter med data om vad användarna tycker om olika egenskaper [5]. För att detta ska vara möjligt måste det finnas data om egenskaper hos produkterna, dessa data används för att beräkna likheten mellan produkter. Det krävs dessutom att man samlar in data om vilka produkter användarna interagerar med för att skapa en modell eller profil för användarna [6]. En bra rekommendation för en viss användare utgörs då av en produkt vars egenskaper stämmer väl överens med den användarens profil. Det är önskvärt att profilen för användaren överensstämmer med användarens verkliga preferenser för att rekommendationerna ska kunna göras så bra som möjligt.

Ett enkelt exempel är om en användare av en tjänst för strömning av musik lyssnat på flera låtar som klassats som ”rock” av systemet, så kan systemet rekommendera andra låtar som också klassats som ”rock”. I det här fallet är egenskapen hos produkten ”rock” och ”rock” ingår i profilen för användaren i fråga.

6.2 Uppbyggnad av innehållsbaserade system

En allmän struktur för innehållsbaserade system består av tre delar [5]. En del analyserar information om produkterna för att skapa en representation av dem i systemet. En annan del skapar profiler för användarna, och slutligen använder en tredje del resultaten från de två tidigare för att skapa rekommendationer.

Den första delen tar in ostrukturerade data om produkterna för att sedan extrahera relevant information och lagra den i strukturerad form. Denna del kallas Innehållsanalyseraren (eng. Content Analyser). Innehållsanalyseraren omvandlar informationen till en form som sedan kan användas av de andra delarna i systemet. Från produktbeskrivningar kan till exempel relevanta nyckelord plockas ut och användas för att representera produkterna i systemet [5].

Den andra komponenten i systemet kallas Profilbyggaren (eng. Profile Learner). Profilbyggaren samlar in data om användarnas beteende och preferenser [5]. Dessa data är i form av respons om produkterna, exempelvis att användarna betygsatt eller på något annat sätt visat intresse för någon produkt. Responsen kan tolkas antingen som positiv eller negativ. Alternativt kan användarna explicit fastställa vad de gillar och ogillar genom att fylla ett frågeformulär eller genom någon annan motsvarande metod.

Responsen om produkterna kallas explicit om användaren är tvungen att aktivt bedöma och rapportera sin åsikt åt systemet. Sådan respons kan erhållas genom att låta användarna betygsätta produkterna: binärt: gilla eller ogilla, på en skala eller skriva kommentarer. Respons som samlas in genom att följa och analysera användarnas beteende men inte kräver aktivt deltagande av användarna kallas implicit respons. Den formen av respons kan samlas in genom att följa användarnas beteende och ge specifik betydelse åt vissa handlingar. Att en användare öppnar beskrivningen för en produkt kan tolkas som att användaren har intresse för produkten och tiden en användare låter en video spela innan hen stänger av den kan antyda hur intresserad hen är av videon. Fördelen med implicit respons är att den inte kräver någon aktiv insats av användarna, utan kan samlas in kontinuerligt under normal användning av tjänsten. Det finns ändå en viss osäkerhet angående riktigheten hos den formen av respons, eftersom den beror på hur användarnas handlingar tolkas. I exemplet med videon kan det hända att användaren inte aktivt ser på videon utan glömt videon på och trots det tolkas handlingen som att användaren visar intresse för videon. Det finns även risk för införandet av partiskhet i systemet eftersom valet för hur handlingarna skall tolkas är subjektivt.

Profilbyggaren använder en mängd av representationer för produkter tillsammans med den respons som getts för motsvarande produkter för att skapa en profil för användarna. Vanligtvis används maskininlärning för att skapa en modell som gör förutsägelser om en användares preferenser, modellen utgör profilen för användaren [5]. Modellen tränas med representationerna och responsen för att sedan kunna förutsäga vad en användare skulle kunna ge för respons givet vissa attribut hos produkten. Till exempel kan det vara

en funktion som tar en representation av en produkt och ger som resultat ett värde som anger hur mycket användaren med den profilen skulle gilla den produkten.

Den tredje delen som kallas Filtreringskomponenten (eng. Filtering Component) skapar rekommendationerna för användarna. Den använder sig av representationerna för produkterna och profilerna för att hitta produkter som skulle kunna intressera användarna [5]. För en viss profil testas olika produkter och rangordnas enligt hur relevanta de är enligt profilen, de mest relevanta produkterna kan sedan rekommenderas åt användaren. Namnet för systemkomponenten kommer från att dess uppgift är att filtrera ut intressanta produkter från mängden av alla produkter.

6.2 Fördelar hos innehållsbaserade system

Innehållsbaserade system kräver inte betyg av andra användare än målanvändaren då den skapar rekommendationer så som kollaborativa system gör [5]. Därför är skapandet av rekommendationer för en viss användare oberoende av andra användares beteende. Om andra användare gett ett väldigt lågt antal betyg åt produkter så lider inte en aktiv användares rekommendationer, vilket skulle vara fallet i ett kollaborativt system.

Eftersom representationen för produkterna skapas av systemet så lider inte innehållsbaserade system heller av problemet med att nya produkter är svåra att rekommendera [5]. För kollaborativa system gör avsaknaden av betyg för en ny produkt som införs i systemet det svårt att skapa rekommendationer för den produkten. Detta problem förklaras mer ingående i stycke 8.1.

Dessa system kan ge explicita motiveringar för rekommendationerna genom att hänvisa till produkter användaren tidigare visat tycke för [5]. Kollaborativa system kan endast motivera rekommendationerna med att liknande användare gillade produkten som rekommenderas. Hur rekommendationerna motiveras kan vara avgörande för mycket användaren kommer att lita på dem och ta dem i beaktande vid ett beslut [5].

7 Sociala system

Sociala rekommendationssystem använder sig av information om användarna och relationer mellan användarna för att skapa rekommendationer [2]. Relationen mellan användarna kallas oftast helt enkelt för en tillitsrelation och därför går systemen också under namnet tillitsbaserade system eller också tillitsstärkta system (eng. Trust-enhanced recommender systems). Den senare benämningen används eftersom de ofta baserar sig på tekniker som även används i andra system, såsom kollaborativa system [7, 8].

Människor har en tendens att hellre använda sig av rekommendationer som kommer från eller på något sätt är kopplade till personer de litar på till skillnad från rekommendationer som är kopplade till anonyma användare [7]. Denna observation kan utnyttjas i sociala rekommendationssystem. Ett exempel skulle vara att meddelandet ”A gillade den här produkten” visas i samband med en rekommendation, där A är en person målanvändaren känner. Alternativt skulle meddelandet ”personer som liknar dig gillar denna produkt” kunna visas, men det första alternativet antas fungera bättre.

Användningen av sociala rekommendationssystem kräver att systemet har tillgång till data om relationerna mellan användarna. Vilken sorts relation eller relationer det är frågan om och hur de representeras varierar. Informationen om relationerna kan anges direkt av användarna, exempelvis genom förfrågningar där användaren explicit anger hur mycket den litar på en annan användare. Alternativt kan data om tillitsrelationerna hämtas från yttre källor. Till följd av att populariteten hos sociala medier och likande tjänster har ökat så är mängden sådana data riklig [7].

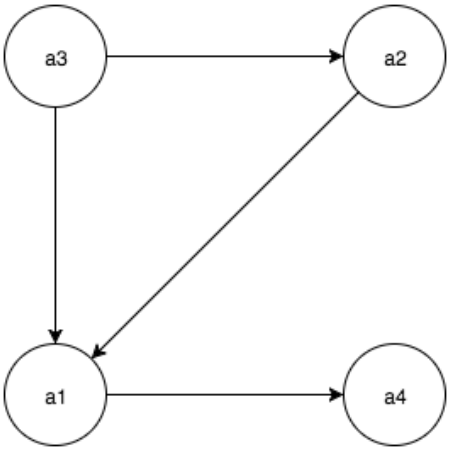
7.1 Tillit

Som det påpekades i det inledande stycket benämns ofta relationen mellan användarna med tillitsrelation. Begreppet tillit kan i det här sammanhanget täcka en mängd olika interpersonliga relationer, så som likhet i smak och uppfattningen av den andras rykte eller kompetens [7].

Mängden av tillit som en användare har till en annan användare kan representeras på olika sätt. En probabilistisk modell beräknar en sannolikhet för varje tillitsrelation, där ett högt värde motsvara en högre sannolikhet att användaren går att lita på. I en gradvis modell har tillitsrelation ett värde som representerar mängden tillit, till exempel kan värdena vara på en skala mellan 0 och 1 där 0 motsvarar total avsaknad av tillit och 1 fullständig tillit.

7.1.1 Tillitsnätverk

En grundläggande struktur i sociala rekommendationssystem är ett tillitsnätverk (eng. web of trust) [2, 7]. Tillitsnätverket för en viss användare byggs upp av de tillitsrelationer hen har direkt till andra användare. Det fullständiga nätverket av alla användare kan representeras som en graf där noderna är användarna och kanterna är relationerna mellan användarna [8]. Kanterna kan vara viktade så att vikten anger styrkan hos relationen, exempelvis graden av tillit, för oviktade grafer antas vikten vara 1.



Figur 3 Ett nätverk.

Figur 5 illustrerar ett enkelt nätverk. I detta nätverk representeras användarna av cirklar och tillitsrelationerna av pilar. Cirklarna med etiketterna a1 till a4 motsvarar samma användare som i matrisen i figur 1, a1 motsvarar $användare\ 1$ och så vidare. Pilarna, och därmed också relationerna de representerar, är enkelriktade i detta nätverk, till

exempel har $användare\ 1$ tillit till $användare\ 4$ medan det omvända inte gäller. Tillitsnätverket för $användare\ 3$ utgörs av relationerna till $användare\ 1$ och $användare\ 2$.

7.1.2 Tillitsmått

Vanligtvis har de flesta par av användare i systemet ingen tillitsrelation mellan sig. Förutsatt att två användare är indirekt relaterade kan man med hjälp av räknesätt för spridning och sammansättning av tillit beräkna ett värde för mängden tillit mellan dem, trots att de inte är direkt relaterade. Detta gör det möjligt att mäta tillit mellan godtyckliga användare, det vill säga man får mått för tillit (eng. Trust metrics).

För att beräkningarna skall kunna utföras krävs det att man inför räknesätt för tillitsrelationerna. Ett av dem används för att beräkna hur tillit sprider sig i nätverket (eng. trust propagation) och ett annan för hur tillit kombineras (eng. trust aggregation) [7]. Motivering för införandet av spridningsoperatorn är att tillit kan antas ha transitiva egenskaper. Om person a har tillit till person b, och b har tillit till person c, så kan man anta att a har tillit till c i någon grad. Detta kräver att typen av tillit som mäts innefattar hur mycket en användare litar på en annan användares förmåga att hänvisa en till andra pålitliga användare.

För att flera tillitsvärden skall kunna kombineras införs en operator för att göra det. Om till exempel tillitsvärdet för användare 3 till användare 1 i figur 5 beräknas så finns det två stigar i grafen som går från noden a3 till a1, en som går direkt och en via a2. Man kan räkna ut värden för de två stigarna genom spridning. För att det ska vara möjligt att kombinera värdena för dessa stigar behövs sammansättning.

Räknesätten kan implementeras på olika sätt beroende på vad man vill att tillitsvärde representerar och vad man vill åstadkomma. Ett vanligt sätt att beräkna spridningen på är genom multiplikation [7]. Det fungerar för tillitsnätverk där värdena är mellan 0 och 1, eftersom mängden av tillits då avtar för varje steg i nätverket. Det är vanligt att definiera räknesättet på ett sätt där tillitsvärdena avtar för varje steg från målanvändaren,

så att användare längre bort från målanvändaren får ett lägre tillitsvärde, även om varje relation i kedjan har full styrka. Man kan även definiera ett maximalt avstånd för spridningen och att använda så att endast tillitsvärdet för användare inom ett visst avstånd från målanvändaren tas i beaktande. Ett sådan mått visas i ekvation (5) [2]. Måttet kan användas i ett nätverk där alla tillitsvärden har vikt 1. Det vill säga att det finns en kant i grafen med vikt 1 om en användare är relaterad till en annan användare, annars finns ingen kant. Det maximala avståndet är d och kan sättas till något godtyckligt värde beroende på hur långt man vill att tillit skall spridas i nätverket, n är antalet steg från användare a till användare b.

$$t(a,b) = \frac{d-n+1}{d} \tag{5}$$

Mått för tillit.

Sammansättningen av flera tillitsvärden kan exempelvis beräknas som ett medelvärde av de beräknade värdena för varje stig till målanvändaren.

7.2 Utförande

Sociala rekommendationssystem kan implementeras med metoder som liknar de som används i kollaborativa system [2, 8]. Till skillnad från de kollaborativa systemen kan man använda sig av tillitsrelationerna, utöver eller i stället för likheter i betygsättningen, då man söker efter liknande användare. Vid beräkning av en förutsägelse för betygen kan en formel som liknar den i ekvation (4) användas, och då kan vikterna $v_{a,b}$ sättas som värdet som representerar hur mycket tillit användare a har till användare b. Graden av tillit kan också användas som kriteriet för vilka användare som tas med i mängden L. Valet att göra dessa förändringar kan motiveras av idén att man får en bra förutsägelse eftersom användare a litar på användarna i mängden L, och i förlängningen på betygen de gett. Man har även, i vissa data, funnit korrelation mellan grupper av relaterade användare och deras preferenser [8], det vill säga att relaterade användare har liknande preferenser. Det skulle tyda på att användningen av tillitrelationerna i sådana fall kan ge bättre förutsägelser.

Om man vill beräkna en förutsägelse för betyget $användare\ 2$ skulle kunna ge åt $produkt\ 3$ så skulle den vanliga kollaborativa metoden inte fungera. Eftersom $användare\ 1$ och $användare\ 3$ inte har betygsatt samma produkter som $användare\ 2$ kan man inte beräkna likhetsvärdena för dessa par av användare. Som det argumenterades för i stycke 5.1.4 kan man inte heller använda likhetsvärdet för $användare\ 4$ eftersom de endast har en produkt gemensamt. Således kan man inte heller skapa den grupp av liknande användare som krävs för att den kollaborativa metoden skall fungera. Men om man har tillgång till nätverket i figur 3 kan man i stället använda den tillitsbaserade metoden.

För den tillitsbaserade metoden skapar man först en grupp av användare målanvändaren litar på genom att använda sig av tillitsvärdena och spridningsoperatorn. Målanvändaren $användare\ 2$ litar på $användare\ 1$ med värdet 1. Om man använder formeln i ekvation (5) för spridning av tillit med det maximala avståndet $d = 3$, så kan man beräkna ett tillitsvärde för $användare\ 2$ till $användare\ 4$. Det beräknade värdet är $t(användare\ 2\ , användare\ 4\ ) = \frac{3-2+1}{3} = \frac{2}{3}$, n är i det här fallet 2 eftersom $användare\ 4$ är två steg från $användare\ 2$ i nätverket. Tillitsvärdet för $användare\ 2$ till $användare\ 3$ är 0 eftersom det inte går någon stig från a2 till a3.

Man kan använda de uträknade tillitsvärdena och ekvation (4) för att beräkna ett förutsägelse för betyget $användare\ 2$ skulle kunna ge åt $produkt\ 3$. Mängden L i ekvationen innehåller de användare för vilka tillitsvärdet är olika 0, alltså $användare\ 1$ och $användare\ 4$. Tillitsvärdena används som vikter. Det beräknade värdet är $f(användare\ 2, produkt\ 3) = 4{,}6$.

8 Problem

Det finns en del faktorer som kan orsaka problem i rekommendationssystem och leda till att rekommendationerna systemet genererar inte är optimala. I detta stycke presenteras två vanligt förekommande problem.

8.1 Otillräckliga data

Med otillräckliga data eller glesa data (eng. Data Sparsity) avses problemet att det inte finns tillräckligt med information för att ett rekommendationssystem skall kunna generera bra rekommendationer. Det är vanligt att endast en liten del av paren av produkter och användare har ett värde, det beror på att varje användare endast betygsätter ett fåtal av den enorma mängden produkter [2, 6]. Exemplen som används i den här avhandlingen är väldigt små jämfört med mängden data systemen måste arbeta med i verkliga tillämpningar. Det kan finnas miljontals användare och produkter varav varje enskild användare endast har betygsatt några få produkter.

Om data representeras i matrisform så är det endast en liten del av positionerna i användare-produkt matrisen som har något värde, det vill säga det finns glest med värden i matrisen. Detta utgör ett problem speciellt för kollaborativa system där glesheten leder till att det är svårt att hitta användare som liknar varandra. När en användare har gett betyg åt några få produkter så är sannolikheten att man hittar andra användare som har betygsatt samma produkter låg. Därför är de också svårt att skapa bra rekommendationer för denna användare [2].

8.2 Kallstart

Ett problem som liknar problemet med gles data till sin karaktär, kan förekomma då nya användare eller produkter införs i systemet. I de här situationerna benämns det kallstart problemet (eng. cold start problem) [6, 8]. När en ny användare införs i systemet har den inte betygsatt någon produkt, för kollaborativa system leder det till att man inte kan hitta andra liknande användare och på det sättet generera rekommendationer. För innehållsbaserade system innebär det att man inte kan skapa en verklighetstrogen profil för användaren och med hjälp av den göra rekommendationer [5]. Att man inte kan generera rekommendationer bidrar dessutom till att den nya användaren har det svårare att hitta produkter den skulle vara intresserad av och samtidigt är det mindre troligt att den kommer i kontakt med produkter som den kan betygsätta vilket förvärrar problemet [2]. Motsvarande gäller för produkter, när en ny produkt införs i systemet så saknar den

betyg och kommer därmed troligen inte heller att bli rekommenderad för någon. Det minskar på sannolikheten att den kommer att betygsättas vilket gör problemet värre.

8.3 Möjliga lösningar

De problem som presenterades i styckena 8.1 och 8.2 är vanliga och påverkar systemens prestanda negativt, därför är det av intresse att försöka överkomma dessa problem. De sociala eller tillitsbaserade systemen har egenskaper som kan hjälpa att lösa dessa problem.

Med hjälp av tillitsvärdena och spridning av tillit kan man öka antalet användare som kan användas i beräkningen av betyg [2]. Som det förklarades i stycke 5.1.4 så krävs det att användarna har betygsatt ett antal samma produkter för att ett likhetsvärde skall kunna beräknas. Men det finns ofta många användare som endast har betygsatt ett fåtal produkter och för sådana användare är det möjligt att antalet betygen de gett inte räcker till för att beräkna likhet mellan användare med få betyg. Således kan man inte skapa en bra grupp av likande användare som kan användas vid beräkningen av förutsägelserna för betygen. I sociala system kan man däremot utnyttja tillitsrelationerna och spridningen av tillit för att hitta en grupp av användare som kan användas vid beräkningen av förutsägelserna för betygen. Detta gör det möjligt att göra rekommendationer även för användare som gett väldigt få betyg. Därför gör det systemet bättre på att hantera kallstart situationer. I exemplet i stycke 7.2 var det inte möjligt att beräkna en förutsägelse genom att enbart använda den kollaborativa metoden. Men användningen av tillitsvärden och spridning av tillit möjliggjorde beräkningen av ett värde.

Det bör dock påpekas att likt glesheten i data om betygen så är även data om tillitsrelationerna ofta gles [2] på grund av att en användare vanligtvis endast interagerar med ett fåtal andra användare. Och användare som gett få betyg är också ofta samma användare som gett lite data om tillitsrelationer. Det vill säga de användare som är kallstart-användare med avseende på betyg är ofta också kallstart-användare med

avseende på tillitsrelationerna. Men det har visats att det är mer gynnsamt att få några tillitsvärden jämfört med motsvarande mängd betyg in i systemet [2]. Data om tillitsrelationerna kan användas mer effektivt för att öka antalet användare som förutsägelser kan göras för och minska mängden fel i förutsägelserna [2].

9 Diskussion och sammanfattning

De typer av rekommendationssystem som presenterats i avhandlingen har var för sig egna fördelar och brister. De förlitar sig på olika tillvägagångssätt för att generera rekommendationer, och med de olika utgångspunkterna och metoderna följer egna problem. Kollaborativa system kräver ingen information om produkterna de rekommenderar förutom betygen de fått medan detta är essentiellt i innehållsbaserade system. I kollaborativa system fördelas arbetet och ansvaret att evaluera produkter mellan användarna vilket leder till mindre arbete för systemet. I innehållsbaserade system sköts evalueringen av eller skapandet av representationer för produkterna centralt vilket innebär mer arbete för systemet. Inom vissa domän, exempelvis där produkterna är dokument, kan detta göras automatiskt medan för andra måste skapandet av representationer göras manuellt vilket blir för arbetsamt då mängden produkter är stor [2]. De innehållsbaserade systemen kräver till skillnad från de kollaborativa systemen inte information om andra användares preferenser för att skapa rekommendationer för en specifik användare. I likhet med de sociala systemen kan de motivera sina rekommendationer genom att hänvisa till produkter målanvändaren visat preferenser för, liksom de sociala systemen kan hänvisa till användare som är relaterade till målanvändaren.

Bristfällig eller gles data leder till problem i samtliga system, men sociala system kan hjälpa minska mängden problem det förorsakar genom att använda sig av tillitsrelationer och räknesätt för dessa. Som det argumenterades för i stycke 8.3 så kan de sociala systemen fungera bättre än de andra typerna av system då data om betygen är gles och i kallstart situationer.

Källförteckning

[1] D. H. Park et al., ”A literature review and classification of recommender systems research,” *Expert Systems with Applications,* vol. 39, nr 11, p. 10059–10072, September 2012.

[2] P. Massa och P. Avesani, ”Trust-Aware Collaborative Filtering for Recommender Systems,” i *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, Springer Berlin Heidelberg, 2004, pp. 492-508.

[3] F. Ricci et al., ”Introduction to Recommender Systems Handbook,” i *Recommender Systems Handbook*, F. Ricci et al., Red., Springer US, 2011, pp. 1-35.

[4] X. Amatriain et al., ”Data Mining Methods for Recommender Systems,” i *Recommender Systems Handbook*, F. Ricci et al., Red., Springer US, 2011, pp. 39-71.

[5] P. Lops et al., ”Content-based Recommender Systems: State of the Art and Trends,” i *Recommender Systems Handbook*, F. Ricci et al., Red., Springer US, 2011, pp. 73-105.

[6] R. Burke, ”Hybrid web recommender systems,” i *The adaptive web*, Chicago, Illinois: Springer Berlin Heidelberg, 2007, pp. 377-408.

[7] P. Victor et al., ”Trust and Recommendations,” i *Recommender Systems Handbook*, F. Ricci et al., Springer US, 2011, pp. 645-675.

[8] G. Groh och C. Ehmig, ”Recommendations in taste related domains: collaborative filtering vs. social filtering,” i *GROUP '07 Proceedings of the 2007 international ACM conference on Supporting group work*, New York, 2007.