

# Projekt: Innehållsbaserade nyhetsrekommendationer

Clas Rydergren, [clas@nic-sys.se](mailto:clas@nic-sys.se). Version 2008-08-19.

*Sammanfattning* - Automatiskt genererade rekommendationer kan användas för att sälla i det informationsflöde som finns på Internet idag. Rekommendationssystem på webben brukar delas in i innehållsbaserade (content-based) eller användarfiltrerade (collaborative filtering) tekniker. Denna text beskriver ett antal tidigare publicerade metoder och system för gruppering och rekommendation av nyheter baserat på artikelinnehåll. Ett tänkt projektupplägg för innehållsbaserade rekommendationer för nyhetsartiklar presenteras.

## Inledning

Automatiskt genererade rekommendationer kan användas för att sälla i det informationsflöde som finns på Internet idag. Rekommendationssystem på Internet brukar klassificeras som innehållsbaserade (content-based) eller användarfiltrerade (collaborative filtering). De innehållsbaserade bygger på textanalys, medan de användarfiltrerade bygger på insamlade surfbeteenden hos nuvarande och tidigare besökare och besök. Dessa surfbeteenden kan vara insamlade från den egna sajten och/eller från andra sajter, som t.ex. socialmedia-sajter.

Den typ av rekommendationer som verkar vara vanligast förekommande på Internet idag rekommenderar produkter att köpa. Givet att du har lagt t.ex. en bok i e-handelssajtens virtuella varukorg så får du förslag på andra böcker som du kanske också skulle vara intresserade av att köpa. Denna typ av rekommendationer kan ha ett kommersiellt värde om de ökar försäljningen, eller ökar besöksintensiteten på webbsajten. För en specifik webbtjänst är det också relativt enkelt att utvärdera effekten av denna typ av rekommendationer, t.ex. genom före-/efter-studier.

Rekommendationen kan endera vara generell eller personligt anpassad. För att göra personliga anpassningar tittar man vanligen på vad besökaren tidigare besökt för sidor på webbsajten, vilka länkar besökaren klickat på, och vad personen köpt eller klickat på vid tidigare besök. För att implementera denna typ av automatiska system för rekommendationer ställs stora krav på informationhantering. Information om hur olika sidor eller produkter hänger ihop behöver skapas, och ska rekommendationen vara personlig så krävs också snabbt tillgänglig besöksstatistik. Både besöksstatistiken och informationen om hur sidorna eller produkterna hänger ihop måste kunna uppdateras snabbt när det tillkommer nya sidor eller produkter.

Att rekommendera ytterligare läsvärda nyhetsartiklar till en läsare av en nyhetsartikel kan även det ha ett kommersiellt intresse. Det är inte, på kort sikt, uppenbart att en nyhetssajt ”gör mer pengar” om du läser en artikel till, jämfört med att t.ex. klicka sig vidare på en annons, men på lång sikt så leder rimligen ett väl fungerande rekommendationssystem till fler sidvisningar och fler nöjda besökare.

Många nyhetssajter har redan idag en lista med länkar till andra nyhetsartiklar på sina start- och artikelsidor. Ofta är det en topplista med artiklar med flest besökare, eller liknande. Dessa länkar är oftast inte anpassade för en aktuell nyhetsartikel, och är heller inte personligt anpassade mot den specifika besökaren. Även om det tidigare skapats försök mer personligt anpassade nyhetsrekommendationer (t.ex. på den numera nedlagda Findory.com, <http://www.findory.com>), så, beroende på informationstillgången, är det generella, opersonliga rekommendationer, som känns aktuellt att implementera på befintliga nyhetssajter idag. Dagens nyhetssajter har relativt få besökare som är inloggade, och nyhetssajterna har oftast inte något system där besökaren kan visa sina

nyhetspreferenser på andra sätt än att klicka runt och läsa nyhetsartiklar. Rekommendationssystem för nyhetsartiklar är därför oftast innehållsbaserade (content-based).

Denna text fortsätter med att beskriva ett antal tidigare publicerade metoder och system för gruppering och rekommendation av nyheter baserat på artikelinnehåll. Sedan följer en sektion som innehåller ett tänkt upplägg för innehållsbaserade rekommendationer för nyhetsartiklar. Upplägget syftar till att ta fram ett system för rekommendation som kan visas på befintliga artikelsidor, rekommendationer som bygger på att besökaren besökt en artikelsida, funnit den intressant, och vill läsa flera relaterade artiklar.

## **Tekniker för att rekommendera nyheter**

De idag mest kända projekten runt utveckling av rekommendationssystem deltar i "the Netflix Prize" (Netflix prize, 2006) som är den tävling som videouthyrarsajten Netflix startade år 2006. Tävlingen pågår fortfarande, och priset på en miljon dollar ges till den som förbättrar Netflix nuvarande (dåvarande) rekommendationssystem med mer än 10%. Resultaten mäts av Netflix med hjälp av ett "benchmarksystem". Idag leder AT&T-labs (AT&T, 2008) med ett system som ger 9.15% förbättring.

Netflix-rekommendationerna bygger på en databas med video-betyg, vilket gör att metoderna som testats i detta sammanhang inte är direkt tillämpbara på nyhetsartiklar, eftersom de sällan betygsätts.

Även utan tillgång till besöksstatistik, besökarspårning eller betyg finns många sätt man skulle kunna påstå att två nyhetsartiklar är relaterade till varandra. Kanske handlar de om samma företag, om samma person, om samma företeelse (konkurs, tsunami, startups, OS, fotboll eller något annat), eller om samma geografiska plats (land, län, stad eller gata). Två nyheter kan också anses relaterade för att de publiceras ungefär samtidigt, eller att de är skrivna av samma person.

Tekniker för att hitta relaterade nyheter kan till viss del hämtas från närliggande problemområden. Dessa kan vara tekniker för klassificering och "klustring" av nyhetsartiklar. Ett grundläggande sätt att skapa grupper av relaterade nyheter är den så kallade  $k$ -Means-algoritmen (Chakrabarti, 2003, Kapitel 4.2.2). Metoden bygger på att varje artikel beskrivs i en ord-vektor, i princip en lista med vilka ord som finns i en artikel (och eventuellt hur många gånger de förekommer). Parametern  $k$  anger antalet kluster som ska bildas. Givet att man skapat dessa kluster (som tyvärr inte automatiskt ges något namn) så kan nyinkommande artiklar placeras i dessa kluster genom t.ex. en  $k$ -Nearest Neighbour-algorithm ( $k$ NN). Algoritmen  $k$ NN bygger på att ord-vektorn för artikeln jämförs med tidigare kategoriserade artiklars ord-vektorer och klassas in artikeln i den grupp (kluster) var ord-vektorer som "ligger närmast". Algoritmen  $k$ NN finns även den beskriven i Chakrabarti (2003, Kapitel 5.4). Ett exempel på ett enklare klassificeringsförsök baserat på dessa tekniker, för spanska nyhetsartiklar, finns i Ullise m.fl. (2004).

Ord-vektorerna kan bestämmas baserat på hela eller delar av innehållet. Normalt ser man dessutom till att reducera antalet olika ord som används i dessa ord-vektorer. Tekniker och vinsterna med denna typ av reduktion (feature selection) beskrivs i Yang och Pedersen (1997). Man kan också tänka sig att man skapar ord-vektorer utifrån inledningen av en artikel, med motivet att nyhetsartiklar är skrivna i en "omvänd informationspyramid", där det viktigaste står först (Borgers och van den Bosch, 2007). Stoppord (d.v.s., informationslösa småord) rensas oftast innan ord-vektorerna bestämmas i denna process. En nackdel med denna stoppordsrensning är att information som kan hjälpa till att finna relationer mellan nyheterna kan försvinna. Detta finns beskrivet i Riloff (1995) – alternativt, att inte rensa bort stoppord och försöka utvinna information därifrån, blir dock ofta väldigt beräknings- och språkinformationskrävande.

Klassificering och klustring kan ses som en variant på "taggning" av artiklar. Automatisk klustring och kategorisering resulterar inte i att någon "tagg" sätts, utan det måste i så fall göras i efterhand.

Tekniker för att göra det beskrivs i t.ex. Thirunarayan m.fl (2007). Att sätta dessa ”taggar” efter det att kategorierna är formade är oftast inte nödvändigt, då det inte påverkar grupperingen. Omvändningen är mer intressant; givet att artiklarna har givits ”taggar” så kan dessa användas för att förbättra grupperingen av artiklarna. Även om ”taggarna” är breda så kan informationen användas för att begränsa antalet artiklar som behöver undersökas i grupperingsalgoritmen. Ett system för detta (kallad hierarkisk kategorisering) finns beskriven i Chiang och Chen (2004). För en nyhet i en specifik klass eller kategori passar det ofta bra att rekommendera en annan nyhets i samma klass eller kategori. En manuell klassificering kan utnyttjas, och kompletteras med enkla tekniker för att identifiera t.ex. namn på företag, personer och platser.

Även om nyhetssajten i sig inte frågar, efter eller lagrar information om, vilka artiklar en specifik besökare läser, eller hur intressant artikeln är, så går det att utnyttja tekniker för ”collaborative filtering”. Informationen hämtas in då från tredje-partssajter, t.ex. socialmedia-sajter som ofta används för att länka och kommendera nyhetsartiklar. Endera används information om länkar till nyhetssajten, eller också samlas data om alla nyhetslänknings. Ett system för där innehållsbaserade och ”collaborative filters” kombineras finns presenterat i Balabanovic och Shoham (1997).

Att kunna utvärdera en kategorisering eller en rekommendation är viktigt. Tintarev och Masthoff (2006) tar upp ett antal exempel på hur mått på närhet mellan nyhetsrubriker kan beräknas och utvärderas. Tyvärr bygger flera vanliga mått på ordklassificeringar och fördefinierade ordhierarkier, vilka finns tillgängliga för det engelska språket (genom bl.a. British National Corpus och WordNet). Att använda samma mått på svenska texter skulle kräva mycket förarbete. Vid automatisk klassificering används normalt måtten ”recall” och ”precision” för att avgöra kvaliteten på en metod för kategorisering. I Borgers och van den Bosch (2007) jämförs några mått för artikelrekommendationer, och där används mått kallade Kandalls tau och MAP för att redovisa hur ”bra” en rekommendation är.

Det går också att göra rekommendationer som inte är baserade på innehåll eller på ”collaborative filters”. Dessa, nästan triviala, sätt att rekommendera nyheter kan vara att skapa en lista med de mest lästa nyheterna eller de nyheter som har flest e-post-rekommendationer (Thorson, 2008) eller som har delats mest på t.ex. Facebook.

## **Förslag på projektmetodik**

Projekt syftar till att ta fram en metod för att kunna ta fram och presentera relaterade nyhetsartiklar inom samma nyhetssajt. Givet en samling med nyhetsartiklar så ska metoden kunna användas för att, för en utvald artikel, rekommendera ett flertal andra artiklar. Metoden ska utgå från analys av innehåll i artiklar, snarare än analys av besöksstatistik. Rekommendationerna ska vara generella, inte individanpassade. Metoden får gärna utnyttja att artikelmaterialen är på svenska, men ska vara underhållsfri, vilket eventuellt kan ställa krav på konstruktion av t.ex. stopppordlista och namnlistor, som eventuellt används i projektet.

Upplägget är tänkt att bygga på enkel kategorisering i flera steg, där det första steget kan liknas vid en ”taggning”. Kategoriseringen görs för att skapa vida grupper baserade på identifiering av namn, plats och tidsindelning. Beroende på hur stor andel av nyhetsartiklarna som kan ”taggas” i denna process, kan mer eller mindre avancerade metoder användas. Vid låg andel som kan ”taggas” kan t.ex. den teknik som beskrivs i BBC Radio labs (2008) provas. Tekniken behöver så förfinas något för att blir snabbare (t.ex. genom att reducera antalet dokument att matcha mot). Denna för-kategorisering är tänkt att fungera som säkerhetsåtgärd för att inte ”helt felaktiga” rekommendationer görs. Denna teknik används i prototypen beskriven i Ha m.fl. (2007). För-kategoriseringen vägs dessutom in genom viktning i de efterföljande ord-vektor-beräkningarna. Dessa beräkningar förväntas följa någon av standardmetoderna beskrivna ovan. Dessa metoder är välbeskrivna och har utprovats i flera sammanhang, bl.a. för att hitta relaterade saker (Oldmoe, 2008) och relaterade blogginslag i form av

Wordpress-plugin (Marsh, 2008). Eventuellt bör beräkningarna även kompletteras med möjligheten att identifiera om den rekommenderade artikeln är nästan identisk – ett problem som är mer aktuellt om det är så att rekommendationerna görs över ett nätverk av nyhetssajter som kan innehålla ”samma nyheter”, men med något olika formuleringar.

För utvärdering av rekommendationerna används ofta kvantitativa mått i litteraturen. Måtten verkar dock relativt abstrakta och svaga. Istället vore det önskvärt att försöka utforma beräkningarna så att det går att spåra varför en artikel anses relaterad. Denna information kan också användas för intrimning och verifiering av rekommendationssystemet. Under projektets gång bör också funktionaliteten för relaterade artiklar hos t.ex. Google News (<http://news.google.se>) och NewsExplorer (<http://press.jrc.it/NewsExplorer>) användas som riktmärke för kvaliteten på rekommendationerna.

Även om projektet fokuserar helt på innehållsbaserade rekommendationer så är det önskvärt att systemet har möjlighet att, vid lämpliga steg i rekommendationsberäkningarna, kunna ta in extern information. Denna information skulle t.ex. kunna vara data från något Twingly-liknande system, eller andra automatiska avläsningar av socialmedia-sajter såsom Del.icio.us, Jaiku, Twitter, FriendFeed, eller det för dagen mest populära blogg- eller mikrobloggssystemen.

Projektet förväntas använda ett Solr/Lucene-index, eventuellt dess funktionalitet för ord-vektorer, och eventuellt genom att bygga vidare på dess ”more-like-this”-funktion. Önskvärt är att funktionen ska vara så snabb att rekommendationerna skulle kunna genereras i samband med en artikelvisning.

## **Referenser**

AT&T Labs, BellKor (2008) <http://korta.nu/99e0>

Balabanovic, M., och Sholam, Y., (1997) Combining content-based and collaborative recommendation, <http://korta.nu/247d>

BBC Radio labs (2008) Wikipedia + Lucene’s More like this = useful bits about the bits, <http://korta.nu/0599>

Bogers, T., och van den Bosch, A., (2007) Comparing and evaluating information retrieval algorithms for news recommendation, <http://korta.nu/5e44>

Chakrabarti, S., (2003) Mining the Web – Discovering Knowledge from Hypertext Data, Morgan-Kaufmann Publishers.

Chiang J-H., och Chen, Y-C, (2004) An intelligent news recommender agent for filtering and categorizing large volumes of text corpus, <http://korta.nu/d0fb> (kräver prenumeration)

Ha, S.H., Joo, S.H., Pae, H.U., (2007) Searching for similar informational articles in the Internet channel, <http://korta.nu/fcba>

Marsh, (2008) Similar posts, blogginlägg/Wordpress-plugin, <http://korta.nu/3a99>

Netflix Prize (2006) <http://www.netflixprize.com/>

Nintarev, N., och Masthoff, J., (2006) Similarity for news recommender systems, <http://korta.nu/1a86>

Oldmoe, (2008) Document matching in Ruby, blogginlägg, <http://korta.nu/113e>

Riloff., E., (1995) Little words can make a big difference for text classification, <http://korta.nu/6a62>

Thirunarayan, K., Immaneni, T., och Shalik, M.V., (2007) Selecting labels for news document clusters, <http://korta.nu/c9ca>

Thorson, E., (2008) Changing pattern of new consumption and participation: news recommendation engines, <http://korta.nu/14ee> (kräver prenumeration)

Ulises., C.B., García Adeva, J.J., Calvo, R.A., och Ceccatto H.A., (2004) Automatic classification of news articles in Spanish, <http://korta.nu/a98e>

Yang, Y., och Pedersen, J.O., (1997) A comparative study on feature selection in text categorization, <http://korta.nu/2a54>