# بررسی مشکلات Collaborative Filtering مبتنی بر شباهت و ارائه راهکارهایی در این زمینه

شیرین ضرغامی[۱] ، مهندس سید پیمان عمادی[۲]

**چکیده** – **با توجه به رشد روز افزون داده و اطلاعات در دنیای امروز نیاز به فیلتر کردن اطلاعات برای استفاده ی مطلوب از آنها بیش از پیش احساس می شود. به این خاطر بر آن شدیم که در این مقاله به بررسی مزایا و معایب انواع مختلف فیلتر کردن اطلاعات پرداخته و دلایل مطرح شدن *Collaborative Filtering* به عنوان یکی از موفق ترین روش های فیلتر کردن اطلاعات و نقش حائز اهمیت آن با توجه به گسترش شبکه های اجتماعی پرداخته شود. در این مقاله علاوه بر مروری کوتاه بر روش مبتنی بر محتوا [۳] و بررسی انواع *CF*, پیشنهادی برای بهینه تر شدن *CF* بر اساس اعتماد [۲] کاربران مطرح شده است.**

**کلید واژه**– **Recomandation System ، Collaborative Filtering , Item-Based CF،Trust**

## ۱– مقدمه

با توجه به گسترش روز افزون شبکه های اینترنتی در زندگی امروز، نیاز به استفاده از روش های مختلف داده کاوی برای ارائه ی خدمات بهتر به کاربران بیش از پیش احساس می شود. در گذشته بیشتر ،سوابق کاربران برای رسیدن به این هدف مورد ارزیابی قرار می گرفت اما به دلیل پراکندگی زیاد اطلاعات، این روش به تنهایی برای گرفتن نتیجه ی ایده آل کافی نیست، اینجا است که اهمیت فیلتر کردن اطلاعات احساس می شود. به طور کلی روش های فیلتر اطلاعات به سه دسته ی کلی زیر تقسیم میشود:[۱]

روش های مبتنی بر محتوا [۲]: این روش بر مبنای محتوای سند است. که می تواند شامل فیلم، آهنگ ، موسیقی، خبر،....باشد.
Collaborative Filtering (CF): در این روش ، مبنای کار گروه بندی افراد مشابه بر اساس نرخ (rate) هایی است که اعلام کرده اند و از تجربه ی افراد برای پیشنهاد دادن به هم گروهی های آنها استفاده می شود.
Hybrid: تلفیقی از دو روش بالا محسوب می شود.

[۱]موسسه آموزش عالی روزبه ، Zarghami.sh@roozbeh.ac.ir

[۲]موسسه آموزش عالی روزبه ، Emadi.p@roozbeh.ac.ir

[۳] Content_ base [۲۰] trust

در روش های مبتنی بر محتوا تنها محتوای شی(item )مورد بررسی قرار می گیرد از جمله الگوریتم های موجود در این زمینه می توان به الگوریتمهای براساس clustering [۳] ، Meta tag [۴] ، پردازش تصویر [۵] و...اشاره نمود. هر چند ،روش های مبتنی بر محتوا ابزاری بسیار ارزشمند در زمینه ی دادن پیشنهاد های مناسب به کاربران محسوب می شود اما با مشکلاتی مواجه است که از آن جمله می توان به موارد زیر اشاره نمود:

- در فهرست کردن منابع چند رسانه ای محدودیت دارد.
- براساس قواعد موضوعی (syntax)است و بعضی مواقع در انتقال معنا (semantic) دچار مشکل می شود.
- بارشناخت دامنه (cognitive load) را به کاربر و سرویس دهنده تحمیل می کند. به عبارت دیگر هنگامی که کاربر در حال جستجو است باید کلمات کلیدی شی(item ) مورد نظر را بداند زیرا در غیر این صورت نمی تواند در زمان کوتاه، جستجوی ایده آلی داشته باشد. از سوی دیگر ارائه کننده ی شی نیز باید کلمات کلیدی مناسبی را برای شی خود انتخاب کند.

می توان گفت روش مبتنی بر CF به دلیل کار کردن بر اساس نظر کاربران توانسته تا حدی پاسخگوی این مشکلات باشد. در حقیقت CF با استفاده ازشبکه های اجتماعی(social network)باعث ارائه پیشنهادات بهتر شده است[۶] در ادامه به بررسی CF برمبنای شباهت(similarity) پرداخته شده و برای رفع مشکلاتی که این روش با آنها مواجه است ، به مبحث CF بر اساس اعتماد به عنوان راه حل، پرداخته می شود.رئوس مطالب ارائه شده عبارت است از:

۲.بررسی انواع CF
۳.مراحل انجام شدن CF سنتی
۴.موارد استفاده از CFو بررسی معایب و مزایای آن
۵.بررسی trust به عنوان راه حلی برای مشکلات CF بر مبنای شباهت
۶.ارائه پیشنهاد، برای بهینه تر شدن CF بر مبنای اعتماد(trust)
۷.کارهای آینده
۸.نتیجه گیری

## ۲- بررسی انواع CF:

دو دسته بندی اصلی در CF عبارت است از:[۷]

**۲-۱ بر اساس مدل( model_ based) :** این روش با ساختن یک مدل از امتیازات کاربران به آنها پیشنهاد می دهد. در این الگوریتم دسته بندی بر اساس احتمال بوده و از الگوریتم های machine learning نظیر چند الگوریتم زیر استفاده می شود:

Bayesian network: مدل احتمال را به صورت فرمول در می آورد تا در CF استفاده کند.این قابلیت را دارد تا به صورت off-line ساعت ها یا روزها بر روی یک مسئله کار کند اما برای محیط هایی که به بروز کردن سریع نیاز دارد مناسب نیست.

Clustering :به صورت یک مسئله از نوع classification کار می کند.به این صورت که احتمال قرار گرفتن کاربری در یک کلاس حساب کرده و بر حسب عدد بدست آمده تصمیم می گیردو کاربران مشابه رادر یک کلاس قرار می دهد.

Rule based: از ارتباطاتی نظیر فروش یک شی به چند کاربر قوانین را استخراج می کند.

### ۲-۲-براساس حافظه(memory_ based):

در این روش با استفاده از پایگاه داده ای که از کاربران و شی ها تشکیل شده و یکسری تکنیک ثابت برای پیدا کردن افراد مشابه با کاربر فعال استفاده شده است. این شباهت میتواند به صورت صریح(explicit) با نظر سنجی و یا از طریق بررسی فعالیت هایی که کاربر انجام می دهد( مانند خرید مشابه) و فرمولی که تعریف شده ،(implicit) استخراج شود.پس از بررسی این اطلاعات به گروهبندی افراد مشابه پرداخته و سپس باالگوریتم هایی نظیر KNN
( K Nearest Neighbors) می توان به کاربران شی پیشنهاد کرد.
در ادامه به بررسی بیشتر این روش و توضیح CF سنتی [۸] پرداخته شده است.

## ۳-مراحل انجام شدن CF سنتی:

### ۳-۱- ارزیابی کاربران(rating):

همان طور که توضیح داده شد به دو صورت explicit , implicit ماتریسی همانند (شکل ۱) از نرخ هایی که کاربران به شی ها داده اند تشکیل می شود.

**۳-۲- تشکیل گروه:** هسته ی اصلی CF محسوب می شود. پس از تشکیل ماتریس مرحله قبل می توان به دو صورت گروه بندی را انجام داد در ادامه به توضیح CF بر مبنای کاربر (user_ based)و CF بر مبنای شی(item-based)پرداخته می شود.

**CF بر مبنای کاربر(USER _ BASED):**کاربران ستون ها و اشیاء سطر ها را تشکیل می دهند و میزان شباهت اشیاء بر اساس کاربرانی که به آن اشیاء رای (rate) داده اند مورد بررسی قرار می گیرد.مشکل این روش در زمانیکه تعداد شی ها زیاد می شود بروز می کند برای مثال در سایتی مانندAmazoon.com که دارای اشیاء زیادی است احتمال اینکه تمام اشیاء حداقل توسط دو کاربر استفاده شده باشد بسیار کم و می توان گفت نا ممکن است .به این دلیل از روش مبتنی بر شی (item -based) استفاده می شود.

**CFبر مبنای شی (ITEM -BASED) :**در این روش همان طور که در (شکل۱) نمایش داده شده است برخلاف روش قبل از روی میزان شباهت itemها به شباهت کاربران می رسند.به عبارت دیگر برای یافتن شباهت دو کاربر به بررسی اشیائی که هر دو ،آن ها را ارزیابی کرده اند پرداخته واز طریق فرمول های نظیر آنچه در ادامه ذکر شده میزان شباهت کاربران محاسبه می شود:[۷]

- شباهت بر اساس کسینوس:

ابتدا شی های i,j را به صورت بردار در نظر گرفته و شباهت به صورت زیر محاسبه می شود:

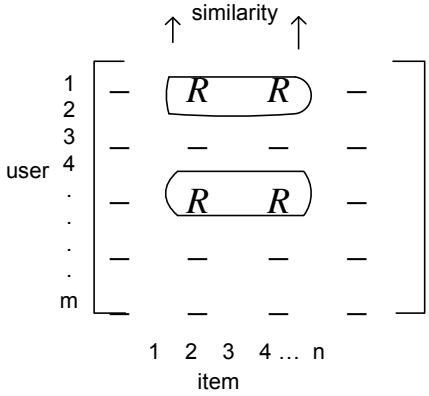
$$Sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2}$$

- روش Pearson :

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\bar{R}_i)(R_{u,j}-\bar{R}_j)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_i)^2}\sqrt{\sum_{u\in U}(R_{u,j}-\bar{R}_j)^2}}.$$

در فرمول بالا U مجموعه کاربرانی است که اشیاء i,j را انتخاب کرده اند.

R u, i معادل ارزیابی هایی است که کاربران عضو u روی شی i تعریف کرده اند.[۸]




(شکل ۱)

## ۳-۳- پیش بینی(prediction ) اشیاء:

از مهمترین قسمت CF، انجام پیش بینی (prediction) اشیاء محسوب می شود. پس از تشکیل همسایگی ها برای یک کاربر، می توان میزان علاقه ی او را برای یک شی مورد نظر پیش بینی کرد.این مقدار بر اساس نرخ (rating) های همسایگان کاربر اندازه گیری می شود و در نتیجه از پیشنهاد دادن شی غیر مرتبط به کاربر پیشگیری می گردد.

[۱۱]

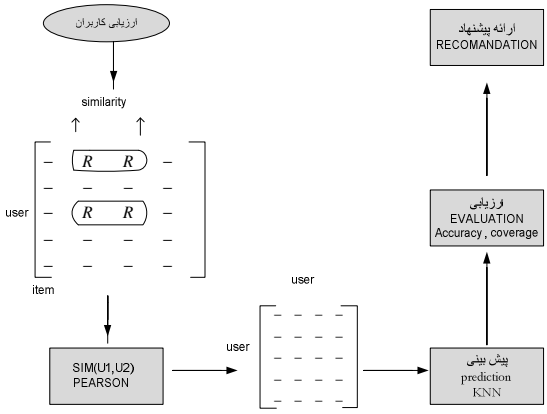
## ۴-۳- پیشنهاد (recommendation) اشیاء:

برای پیشنهاد دادن شی ابتدا باید میزان سازگاری آن را با آنچه که کاربر انتخاب می کند محاسبه کرد. اگر این میزان از آستانه ای بالا تر بود می توان از این روش برای پیشنهاد دادن استفاده نمود. معیار های زیادی برای ارزیابی روش CF وجود دارد اما می توان به دو عامل دقت و میزان پوشش (coverage) به عنوان مهمترین عامل ها نام برد.منظور از دقت معیاری است که نشان دهنده ی میزان تطابق

پیشگویی انجام شده با نظر کاربر و منظو ر از میزان پوشش معیاری است که نشان می دهد چه میزانی از اشیاء مناسب برای پیشنهاد دادن از بین تمام شی های مناسب پیدا شده است.برای بررسی این دو معیار روش های مختلفی موجود است از جمله ی این روش ها می توان به MAE(mean absolute error ) اشاره نمود[۸].در این روش ابتدا مراحل فوق را روی یک data set که می توان از سایت هایی نظیر Movielens تهیه نمود انجام داده و سپس با استفاده از فرمول زیر میزان خطا و یا به عبارت دیگر میزان عدم تطابق بین خواسته ی کاربر و آنچه به او پیشنهاد شده محاسبه می شود:

$$MAE = \frac{\sum_{i=1}^{N} | P_i - q_i |}{N}$$

در حقیقت pi بیانگر پیش بینی که صورت گرفته و qi بیانگر خواسته ی کاربر است و N نیز تعداد اشیاء را مشخص می کند.در انتها اگر MAE از یک آستانه ای پایین تر بود می توان از این روش برای پیشنهاد دادن استفاده نمود مراحل فوق به اختصار در(شکل ۲) نشان داده شده است:




(شکل ۲)

## ۴-موارد استفاده از CFو بررسی معایب و مزایای آن:

از کاربردهای CF می توان به موارد زیر اشاره نمود:

- سیستم های پیشنهاد دهنده:

از CF همانند روش های دیگر فیلتر کردن برای ارائه ی پیشنهاد استفاده می شود.برای مثال اگر کاربر در جستجوی کتابی در زمینه ی داده کاوی باشد می تواند از نظرات افراد مشابه برای یافتن کتاب مناسب استفاده کند.

- سیستم کمک به تصمیم گیری:

در اینجا کاربر شیئی که حدس می زند مناسب است را انتخاب کرده ولی از انتخاب خود مطمئن نیست به همین دلیل با افراد مشابه در مورد آن شی مشورت می کند.برای مثال فرض کنید فیلم جدیدی در سایت Movielens[۴] قرار گرفته که اطلاعات زیادی در مورد آن ندارید .در این صورت با استفاده از نظر افراد مشابه می توان در مورد کیفیت آن مطلع شوید و در صورت مناسب بودن برای آن فیلم وقت صرف کنید .

- فیلتر کردن اسناد مخرب:

با CF می توان اسناد مخربی نظیر spam ها را شناخت و از آسیبی که آنها می رسانند جلوگیری نمود.

- استفاده در وب سرویس:

از سوی دیگر با توجه به رفتن مدل تجاری اینترنت به سمت تکنولوژی وب سرویس ،استفاده از CF به صورت پر رنگ تر جلوه می کند. وب سرویس در حقیقت نرم افزارهایی هستند که از XML و یکسری protocol برای ایجاد تعامل بین نرم افزارهای درون یک شبکه استفاده می کند.[۹]

یکی از امکانات وب سرویس آن است که می توان نرم افزار را روی یک سرویس دهنده قرار داد و کاربر به وسیله ی شبکه از امکانات آن نرم افزار استفاده کند.البته می توان با استفاده از CF کاربر را در یافتن توابع واستفاده از امکانات نرم افزار کمک نمود[۱۰]

از مزایای CF می توان به موارد زیر اشاره نمود:

- مشکل ارزیابی منابع مالتی مدیا را حل کرده است.زیرا این روش نیاز به بررسی محتوای شی ندارد.
- تا حد زیادی مشکل فهم معنایی در وب(semantic) را بر طرف نموده زیرا مانند روش مبتنی بر محتوا براساس قواعد (syntax)نیست و بر اساس نظر کاربر که عامل هوشمندی است پیشنهاد ارائه می شود.
- با توجه به گسترش شبکه های اجتماعی(social network) و تمایل افراد برای تعامل با افراد مشابه می توان نیاز به CF را احساس کرد.

اما CF بر مبنای شباهت دارای معایبی است که در ادامه بررسی می شود:

پراکندگی(sparsity):

به این معنا که اطلاعات به صورت پراکنده است به عبارت دیگر فقدان یا کم بودن اشتراک اشیاء مشابه بین کاربران باعث می شود که نتوان از آنها برای بدست آوردن شباهت استفاده کرد.

مرحله آغازین سیستم (Cold start phase) :

در ابتدای کار و یا زمانی که کاربر جدید هنوز به اشیاء رای نداده است ما اطلاعات زیادی در مورد کاربران نداریم. پس انجام محاسبات شباهت دشوار می شود. زیرا همانطور که مطرح شد تنها شی هایی که هر دو کاربر ارزیابی کرده اند بررسی می شود.

اولین رای دهنده (First rater):

این مشکل زمانی پیش می آید که شی های رای داده شده توسط یک کاربر هنوز توسط سایر کاربران رای داده نشده است. پس شی مشترکی برای محاسبه ی میزان شباهت وجود ندارد. مشکلات ذکر شده برای CF براساس شباهت نیاز به استفاده از پارامتر دیگری مانند اعتماد بین کاربران را محسوس تر می کند و سبب استفاده از اعتماد بین کاربران( trust )به عنوان جایگزینی برای CF بر اساس شباهت شده است . در ادامه به توضیح این مطلب پرداخته می شود.

[۴] http://www.cs.umn.edu/research/GroupLens/data/

## ۵.بررسی trust به عنوان راه حلی برای مشکلات CF بر مبنای شباهت:

مشکلات بیان شده سبب استفاده از اعتماد بین کاربران به عنوان جایگزینی برای CF بر اساس شباهت شده است. به عبارت دیگر در این روش علاوه بر بررسی شی هایی که هر دو کاربر به آن رای داده اند شی های دیگر نیز بررسی می شود و به جای استفاده از فرمول های ارائه شده در بخش ۳ می توان از فرمول ها و روش های زیر استفاده نمود[۱۱]

ابتدا با استفاده از شباهت، میزان اعتماد (trust) کاربر a به b محاسبه می شود . نکته حائز اهمیت آن است که در روش قبل میزان شباهت کاربر a به b با شباهت bبه a متفاوت نبود اما در CF بر اساس اعتماد متفاوت است زیرا در بدست آوردن میزان اعتماد کاربر a به کاربرb تمام شی هایی را که کاربر a به آنها رای داده است بررسی می شود حتی اگر کاربر b بررسی نکرده باشد. در میزان اعتماد کاربر b به a دقیقا" برعکس این موضوع مطرح می شود. نحوه ی محاسبه ی میزان اعتماد کاربر a به b در فرمول زیر مطرح شده است .( رای (rate) ها عدد مثبتی بین صفر تا پنج است) :[۱۱]

$$value(a,b,i) = -\frac{1}{5} | r_{a,i} - r_{b,i} | +1$$

$$trust(a,b,n) = \frac{\sum_{i=0}^{n} value(a,b,i)}{n}$$

در این رابطه میزان اعتماد دو کاربرa,b از رویn شی ارزیابی شده توسط کاربر a محاسبه می شود.

بعد از محاسبه ی میزان اعتماد می توان گرافی از کاربران تشکیل داد که وزن روی یال ها بیانگر میزان اعتماد و گره ها نیز کاربران باشد. سپس نوبت به انتخاب افراد مورد اعتماد کاربر از روی این گراف برای تشکیل گروه می رسد. از مرسوم ترین روش های موجود در این زمینه می توان موارد زیر را نام برد:

- تعین یک آستانه برای اعتماد : در این روش پیدا کردن مقدار آستانه دشوار است و بستگی به نحوه ی توزیع وزن ها در اجتماع کاربران دارد.[۱۲]

- استفاده از الگوریتمKNN : در این روش k همسایه ثابت از افراد مورد اعتماد کاربر تشکیل شده و از نظرات آنها استفاده می شود. مشکلی که پیش می آید آن است که ممکن است افراد هم گروه با کاربر در همه ی زمینه های مورد علاقه ی کاربر اطلاعات لازم را نداشته باشند. در حالی که افرادی خارج از گروه بتوانند نظرات بهتری در آن زمینه ارائه دهند.

- (KNR)K –Nearest Recommenders [۱۱]:

این روش جدید، برای رفع مشکلKNN ایجاد شده است.در این روش همانند روش بالا عمل می کنیم با این تفاوت که مجموعه افراد مورد اعتماد ، ثابت نیستند و برای موضوعات متفاوت همسایگان متفاوتی انتخاب می شود. این روش هرچند پوشش دهی (coverage )بالایی دارد اما به دلیل پردازش کل مجموعه در مورد مباحث مختلف هزینه ی بالایی را تحمیل می کند. در بخش بعد راهکاری برای بهینه تر شدن این روش ارائه شده است.

## ۶.ارائه پیشنهاد برای بهینه تر شدن CF بر مبنای اعتماد(trust):

روش KNR باعث بهینه تر شدن CF از نظر دو معیار ارزشمند دقت و میزان پوشش دهی می شود.اما تنهانکته ی منفی این روش هزینه ی زیاد آن برای بررسی مکرر تمام اجتماع برای یافتن k همسایگی در زمینه های مختلف است. راهکار پیشنهادی ، بر مبنای تلفیقی از روش مبتنی بر محتوا و CF است به این صورت که در ابتدا به صورت خوش بینانه بر مبنای معیار هایی نظیر زمینه ی کاری و یا رشته تحصیلی ،گروه هایی از کاربران تشکیل شده و در ابتدای کار به کاربران در گروهی که قرار میگیرند یک نرخ اعتماد ثابت داده شود. پس از سپری شدن مدتی از کار این سیستم ،با استفاده از نظرافراد اجتماع در مورد پیشنهاد آن کاربر در زمینه ی گروهی که قرار دارد از میزان اعتماد نسبت به او کاسته یا افزوده شود. اگر این میزان اعتماد از حدی بیشتر بود، آن کاربر می تواند به عنوان معیاری برای پیدا کردن افراد دیگری که می توانند عضو آن گروه شوند استفاده شود و اگر میزان اعتماد به کاربر پایین تر از یک آستانه ی تعریف شده باشد از گروه حذف شده و با نمایندگان گروه های دیگر مقایسه شده ،تا گروه مناسب برای آن کاربر تشخیص داده شود. البته این نکته قابل توجه است که گروهبندی ما در این قسمت برای پیدا کردن افراد مورد اعتماد در زمینه های مختلف است که باعث می شود، برای پیدا کردن k همسایه، همه ی افراد اجتماع بررسی نشوند و تنها گروه افرادی که در آن زمینه اطلاعاتی دارند استفاده شود.البته چنین سیستمی با گذشت زمان و جمع آوری نظرات بیشتر در مورد کاربران، نتایج بهتری تولید خواهد کرد.

## ۷ . کارهای آینده:

در ادامه قصد داریم به طراحی و شبیه سازی این روش پرداخته و با استفاده از معیارهای CF تاثیر این روش بر بازده کار را بررسی کنیم.

## ۸.نتیجه گیری:

در این مقاله به بررسی انواع مختلف فیلتر کردن اطلاعات از جمله فیلتر کردن بر اساس محتوا و انواع CF پرداخته و همچنین چند روش از CF بر مبنای اعتماد بین کاربران، به عنوان راه حلی برای رفع مشکلات CF برمبنای شباهت مطرح گردید.در خاتمه ،پیشنهادی برای بهبود CF بر مبنای اعتماد کاربران ارائه شد.

## سپاسگزاری

در آخر از جناب آقای مهندس علیرضا ضرغامی و خانم مهندس سوده فاضلی که ما را در تهیه ی این مقاله یاری نموده اند صمیمانه تشکر می کنیم.

## مراجع

[۱] O.Nouali, S.Kirat, H.Meziani,

A BASIC PLATFORM OF COLLABORATIVE FILTERING

[۲] R. Kosala. and H. Blockeel, Web Mining Research: A Survey, SIGKDD Explorations, ۲(۱):۱-۱۵, ۲۰۰۰.

[۳] Ramiz M. Aliguliyev A Novel Partitioning Based Clustering Method and Generic Document Summarization,۲۰۰۶.

[۴] Danny Sullivan ,*How to use HTML Meta Tags* ,Mar ۲۰۰۷.

[ ۵] A.P. Asirvatham and K.K. Ravi, *Web Page Classification based on Document Structure,* International Institute of Information Technology.

[۶] Sarah K. Tyler and Yi Zhang ,Open Domain Recommendation: Social Networks and Collaborative Filtering
,University of California, Santa Cruz

[۷] Badrul Sarwar, George Karypis Joseph Konstan, and John Riedl , Item-based Collaborative Filtering Recommendation Algorithms
GroupLens Research Group

[۸] Greg Linden ,Brent SMITH., and Jeremy York, Amazon.com Recommendations , Item-to-Item collaborative Filtering
Published by the IEEE computer society , JANUARY * FEBRUARY ۲۰۰۳

[۹] The World Wide Web Consortium (*W۳C*)
http://www.w۳.org/TR

[۱۰] Naoki Ohsugi, Akito Monden S h uuji Morisaki , Collaborative Filtering Approach for Software Function Discovery

[١١] N. Lathia, S. Hailes, and L. Capra, “Trust-based collaborative filtering,” in IFIPTM ٢٠٠٨: Joint iTrust and PST Conferences on Privacy, Trust management and Security, ٢٠٠٨, p. ١۴.

[١٢] ٨. N. Lathia, S. Hailes, and L. Capra. The effect of correlation coefficients on communities of recommenders. In To Appear in ACM SAC TRECK, ٢٠٠٨