Die Arbeit „*Multi-Domain Collaborative Filtering*“[1] wurde von Yu Zhang, Bin Cao und Dit-Yan Yeung (alle von der Universität für Wissenschaft und Technologie in Hong Kong) erstellt.

Im ersten Abschnitt (*Introduction*) erklären die Autoren kurz was kollaboratives Filtern im Kontext von Empfehlungssystemen (*Recommender Systems*) bedeutet und wo diese Methoden Anwendung finden. Sie stellen fest, dass kollaborative Filtersysteme in den letzten Jahren immer mehr an Verbreitung und Popularität im WWW finden. Basierend auf einer Umfrage unterteilen sie kollaborative Filtersysteme in drei weitere Unterkategorien: erinnerungsbasierte (*memory based*), modelbasierte (*model based*) und Hybridmethoden (*hybrid*).

Erinnerungsbasierte Methoden berechnen Empfehlungen basierend auf Objektbewertungen von anderen Benutzern mit ähnlichen Vorlieben (ähnlich der Nächste-Nachbarn-Klassifikation). Ein großer Nachteil dieser Methode ist, dass die darunterliegenden Datensätze (also die Bewertungen) sehr dicht sein müssen (also wenige Bewertungslücken aufweisen sollten). Eine Anforderung die nur von wenigen Datensätzen erfüllt werden kann.

Modelbasierte Methoden wurden entwickelt um diesem Nachteil entgegenzuwirken. Anhand vom Benutzerverhalten wird ein Model erstellt und dieses Model dient als Grundlage für die Berechnung der Empfehlungen. Es existieren verschiedene bekannte Modelle, die Autoren nennen hierbei das Bayes'sche Netz, grafische Modelle und Abhängigkeitsnetze. Die Autoren stellen weiters fest, dass fast alle modelbasierten Methoden auf Matrizenfaktorisierung zurückgreifen. Diese Methoden nehmen an, dass Benutzer- und Objekteigenschaften auf einer niedrigdimensional, verborgenen Ebene liegen und generieren basierend darauf Empfehlungen für die Benutzer.

Hybridmethoden sind ein Versuch die Relevanz der Empfehlungen weiter zu steigern. Durch Kombination von erinnerungs- und modelbasierten Methoden aber auch durch Hinzunahme von weiteren Aspekten.

Auch wenn Empfehlungssysteme mit den bekannten Methoden bereits sehr gute Dienste leisten, besitzen sie laut den Autoren ein Problem wenn wenige Daten bzw. große Bewertungslücken vorliegen. Genau dieses Problem wollen die Autoren in ihrer Arbeit angehen. Dabei wollen sie sich auf „*multi-domain collaborative filtering*“-Probleme spezialisieren. Dieses Problem liegt in der Regel bei großen e-Commerce Projekten oder Sozialen Netzwerken vor, also bei Applikationen die Empfehlungen für Objekte aus vielen verschiedenen Domänen anbieten wollen. Als Multidomänenbeispiel nennen die Autoren Bücher und Elektronik. Die Autoren behaupten, dass sie durch Nutzung der Korrelation zwischen Bewertungen aus verschiedenen Domänen mit anschließendem übertragen des gesammelten Wissens auf ähnliche Domänen, das Problem der Bewertungslücken kompensieren können.

Im zweiten Abschnitt (*Multi-Domain Collaborative Filtering*) präsentieren und erläutern die Autoren ihr System. Der gesamte Abschnitt ist eine Formelsammlung, die dem Leser helfen soll die damit final erstellte mathematische Formel besser zu verstehen. Die final erarbeitete Formel sieht dann aus wie folgt:

$$
\begin{aligned}
J(\{\mathrm{U}^i\},\{\mathrm{V}^i\},\sigma,\lambda,\eta,\Omega) &= \sum_{i=1}^{K}\frac{1}{2\sigma_i^2}\sum_{j=1}^{m}\sum_{k=1}^{n_i} I_{jk}^i\left(X_{jk}^i-\left(\mathrm{U}_j^i\right)^T\mathrm{V}_k^i\right)^2 \\
&+\frac{1}{2}\sum_{i=1}^{K}\left(\ln\sigma_i^2\sum_{j=1}^{m}\sum_{k=1}^{n_i} I_{jk}^i\right) \\
&+\frac{md}{2}\sum_{i=1}^{K}\ln\left(\sum_{j=1}^{m}\left(\mathrm{U}_j^i\right)^T\mathrm{U}_j^i\right) \\
&+\sum_{i=1}^{K}\frac{dn_i}{2}\ln\left(\sum_{k=1}^{n_i}\left(\mathrm{V}_k^i\right)^T\mathrm{V}_k^i\right) \\
&+\frac{1}{2(md)^{K-1}}\ln|\mathrm{U}^T\mathrm{U}|
\end{aligned}
$$

In den darauf folgenden Unterabschnitten präsentieren die Autoren Verfeinerungen ihrer mathematischen Formeln, die das Endergebnis verbessern sollen.

Im dritten Abschnitt (*Incorporation of Link Function*) erklären die Autoren, dass ihre Definition der konditionalen Distribution der Bewertungen (im Artikel ist das die Definition mit der Beschriftung (1)) ein Problem aufweist. Die Anwendung der gauß'schen Klammern zur Umwand-

[1] http://arxiv.org/abs/1203.3535

lung in Binärwerte bei der Produktbildung führt dazu, dass die abzählbaren Ganzzahlen der Bewertungen verloren gehen. Um diesem Problem entgegenzuwirken wird, die zuvor bereits erwähnte, Linkfunktion von den Autoren eingeführt.

Im weiteren Verlauf des dritten Abschnitts wird die Linkfunktion erläutert. Wird die Linkfunktion nun in die finale Formel eingefügt ergibt sich folgende neue Formel:

$$
\begin{aligned}
J_1(\{\mathrm{U}^i\},\{\mathrm{V}^i\},\sigma,\lambda,\eta,\Omega,\theta) &= \sum_{i=1}^{K}\frac{1}{2\sigma_i^2}\sum_{j=1}^{m}\sum_{k=1}^{n_i} I_{jk}^i\left(g\left(X_{jk}^i\right)\right. \\
&\quad \left.-\left(\mathrm{U}_j^i\right)^T\mathrm{V}_k^i\right)^2+\sum_{i=1}^{K}\frac{1}{2\lambda_i^2}\sum_{j=1}^{m}\left(\mathrm{U}_j^i\right)^T\mathrm{U}_j^i \\
&\quad +\sum_{i=1}^{K}\frac{1}{2\eta_i^2}\sum_{k=1}^{n_i}\left(\mathrm{V}_k^i\right)^T\mathrm{V}_k^i \\
&\quad +\frac{1}{2}\sum_{i=1}^{K}\left(\ln\sigma_i^2\sum_{j=1}^{m}\sum_{k=1}^{n_i} I_{jk}^i\right) \\
&\quad +\frac{md}{2}\sum_{i=1}^{K}\ln\lambda_i^2+\sum_{i=1}^{K}\frac{dn_i}{2}\ln\eta_i^2 \\
&\quad +\frac{1}{2}\mathrm{tr}(\mathrm{U}\Omega^{-1}\mathrm{U}^T)+\frac{md}{2}\ln|\Omega| \\
&\quad -\sum_{i=1}^{K}\sum_{j=1}^{m}\sum_{k=1}^{n_i} I_{jk}^i\ln g'\left(X_{jk}^i\right) \\
&\quad +\text{Konstante}
\end{aligned}
$$

$tr()$ in der Formel steht für Englisch *trace* und ist im deutschen die Spurfunktion oder Spurabbildung aus der linearen Algebra.

Da es keine analytische Aktualisierungslösung für θ gibt, verwenden die Autoren eine verlaufsbasierte Methode namens „*scaled conjugate gradient method*“ (zu Deutsch „Verfahren der konjugierten Gradienten“). Der Rest des Abschnittes befasst sich noch genauer mit der Formel und deren Einzelteilen.

Der fünfte Abschnitt (*Experiments*) samt Unterabschnitten zeigt wie die Methode auf realen Datensätzen angewendet werden kann. Für das Experiment verwenden die Autoren zwei öffentlich zugängliche Datensätze. Zum einen den von MovieLens, der sich mit Filmen befasst, und zum anderen den Book-Crossing-Datensatz, der sich mit Büchern befasst. Es ist für die Methode essentiell, dass die Datensätze mindestens zwei unterschiedliche Domänen abdecken, deshalb werden die Datensätze in die verschiedenen Genres bzw. Kategorien unterteilt. Beim MovieLens-Datensatz kommen die fünf populärsten Genres zu Tragen und bei Book-Crossing die fünf generellsten Bücherkategorien. 80% der so eingeschränkten Datensätze dienen dem Training und die restlichen 20% dem eigentlichen Test. Jede Konfiguration wurde zehnmal durchiteriert.

Im letzten Unterabschnitt (*Results*) stellen die Autoren die verschiedenen Algorithmen den Ergebnissen ihrer Verfahren gegenüber. Verglichen werden folgende Methoden:

- Probabilistische Matrixfaktorisierung (PMF)
- Kollektive Matrixfaktorisierung (CMF)
- Multidomänen kollaboratives Filtern (MCF)
- MCF plus Linkfunktion (MCF-LF)

Wie zu erwarten liefert die multidomänen-Methode mit Linkfunktion der Autoren die besten Ergebnisse zwischen den verschiedenen Domänen.

Im letzten Abschnitt (*Conclusion*) schreiben die Autoren noch davon, dass sich gezeigt hat, dass aktive Lernmethoden die Ergebnisse signifikant verbessern können und sie deshalb ihre Methode um eben dieses Verfahren erweitern wollen.