

Analyse von Folksonomy basierten Netzwerken

als komplementärer Ansatz für

Autorenempfehlungen in der Wissenschaft

Tamara Heck

Heinrich-Heine-University

Dept. of Information Science

D-40225 Düsseldorf, Germany

Tamara.Heck@hhu.de

Abstract

Wissenschaftliche Kooperationen sind für Forscher in akademischen Einrichtungen und in forschungsintensiven Unternehmen unerlässlich. Wichtig für einen Forscher ist daher, dass er seine Forschungsgemeinschaft und deren Themenfelder kennt. Ein Recommender System kann hierbei Unterstützung bieten, indem es Wissenschaftlern relevante Ressourcen wie Literatur, Tags und auch mögliche Kooperationspartner vorschlägt. Ein solches Empfehlungssystem basiert auf der Ähnlichkeit zwischen zwei Wissenschaftlern. In der Scientometrie bekannte Methoden sind die Co-Zitationsanalyse und das Bibliographic Coupling, mit denen Relationen zwischen Autoren bzw. Forschern gemessen werden können. In diesem Ansatz soll die Methode des Collaborative Filtering hinzugenommen werden. Wir nutzen Daten vom Folksonomy basierten Social Bookmarking-Dienst CiteULike. Zum einen soll die Frage beantwortet werden, ob relevante Autoren für einen bestimmten Forscher auch mit Collaborative Filtering empfohlen werden können. Zum anderen soll analysiert werden, ob diese Methode zu anderen Ergebnissen führt als Co-Zitationsanalyse und Bibliographic Coupling. Im Folgenden werden erste Ergebnisse der Analysen und der Evaluation vorgestellt.

Keywords

Recommender System, Autoren Ähnlichkeiten, Co-Zitation, Bibliographic Coupling, Folksonomy, Tagging

1. Einführung

Als Forscher in der Wissenschaft und in der Wirtschaft ist es wichtig, Personen mit ähnlichen Forschungsinteressen und Arbeitsfeldern zu kennen. Diese Experten, die entweder gleiche Erfahrungen mitbringen oder die eigenen zweckvoll ergänzen, kommen als potentielle Kooperationspartner in Frage. Gebraucht werden sie z.B. in folgenden Situationen:

- Gründung einer Arbeitsgruppe an einem Institut,
- Zusammenführung von Forschern um einen (geförderten) Projektantrag einzureichen (institutsintern oder –extern),
- Bildung einer Community of Practice, unabhängig von der Forschung am Institut mit Bedacht auf individuelle Interessen,
- Anfrage an Kollegen für einen Workshop- oder Konferenzbeitrag oder Beitrag in einem Handbuch oder Journal,
- Suche von passenden Co-Autoren.

Um die Reputation eines Experten zu prüfen, damit er als möglicher Kandidat für Kooperationen in Frage kommt, können bei einem Forscher seine bisherigen Publikationen in wissenschaftlichen Zeitschriften und deren Zitationen herangezogen werden. In diesem Ansatz nutzen wir dafür die Informationsdienste Web of Science (WoS) und Scopus (Meho & Sugimoto, 2009). Zusätzlich soll untersucht werden, wie die Reputation eines Wissenschaftlers in den Social Media vertreten ist und analysiert werden kann. Hierfür werden Daten aus dem Social Bookmarking-Dienst CiteULike ermittelt. Somit wird nicht nur die Perspektive der Wissenschaftler betrachtet, indem deren Publikationen, Referenzen und Zitationen in WoS und Scopus untersucht werden. Sondern es werden auch die Perspektiven der Nutzer in CiteULike miteinbezogen, indem ihre Bookmarks und Tags zu den Publikationen der Forscher analysiert werden. Zwei Forschungsfragen sollen beantwortet werden: 1. Können wir einem Zielautor ein Netzwerk mit relevanten und neuen unbekanntenen Autoren anhand der Social Media Daten und Collaborative Filtering Methoden vorschlagen? 2. Unterschieden sich diese Ergebnisse von denen, die mit Co-Zitationsanalyse und Bibliographic Coupling gewonnen werden?

Das Ziel von Recommender Systemen (RS) ist es, einem Nutzer bestimmte Ressourcen, z.B. Produkte, Filme, Musik, Artikel, die er für interessant und relevant halten könnte, vorzuschlagen und ihm somit die Suche danach sowie die Entscheidung für ein Objekt zu erleichtern. Ein RS sollte möglichst individuelle Empfehlungen, die auf einen Nutzer zugeschnitten sind, aussprechen (Berkovsky et al., 2007). Dabei sollen die „besten“ Empfehlungen für einen Nutzer a in einer Ranking Liste ganz oben stehen (Desrosiers & Karypis, 2011). In der Regel werden zwei Hauptansätze unterschieden: die content-basierte Methode und Collaborative Filtering (CF). Erstere versucht, Ähnlichkeiten zwischen den Inhalten von denjenigen Ressourcen zu finden, die von Nutzer a positiv bewertet worden sind. CF betrachtet nicht nur die Nutzerbewertungen von a , sondern auch diejenigen der anderen Nutzer eines Dienstes [a.o. Parra & Brusilovsky, 2009; Resnick et al., 1994]. Der Vorteil ist, dass Empfehlungen nicht nur auf Grundlage der teils ungeeigneten Inhalte der Ressourcen ausgesprochen werden, sondern auch Bewertungen anderer Nutzer berücksichtigt werden.

In Social Bookmarking Services (SBS) werden zum einen die Nutzer-Ressource Relationen als Bewertung genutzt (unary user-item response), d.h. setzt ein Nutzer einen wissenschaftlichen Artikel als Bookmark in seine Online Literaturliste des genutzten Dienstes, wird dies als positive Resonanz auf den Artikel gesehen (Desrosiers & Karypis, 2011). Somit können diese Systeme als Basis für Empfehlungen dienen. Zusätzlich gibt es die den Ressourcen (in unserem Fall Bookmarks zu wiss. Artikeln) zugeordneten Tags, d.h. wir haben eine Folksonomy-Struktur mit Nutzer-Bookmark-Tag Beziehungen. Unserem Zielautor wollen wir als potentielle Kooperationspartner nun Autoren vorschlagen, die von den Nutzern als relevant empfunden werden, d.h. sie haben deren Artikel als Bookmark in ihre Literaturliste gesetzt und diesen optional Tags hinzugefügt.

2. Recommender Modelle

Damit RS einem Nutzer die bestmöglichen Empfehlungen aussprechen können, sind einige Faktoren zu beachten, z.B. die Wahl des passenden Algorithmus (Shepitsen et al., 2008), die Definition der Nutzer Interaktionen (Ramezani et al., 2008), die Effektivität, Stabilität und Genauigkeit des Systems (Desrosiers & Karypis, 2011) und die Fähigkeit des RS zu Lernen und die Rankings zu verbessern (Rendle et al., 2009). Für kooperative Nutzerdienste wie SBS wurden verschiedene Algorithmen und kombinierte Methoden

getestet, die sich u.a. auf diverse Nutzer-Ressource-Tag Beziehungen stützen und Relationsgewichtungen nutzen. Cacheda et al. (2011) testen verschiedene Methoden gegen ihren eigenen Algorithmus, der positive und negative Nutzerbewertungen berücksichtigt. Bogers & van den Bosch (2008) vergleichen drei CF-Algorithmen, deren Empfehlungen entweder auf gleichen Ressourcen oder gleichen Nutzern basieren. Ein Problem dabei ist der Kaltstart, d.h. neue Ressourcen können nicht vorgeschlagen werden, da sie noch keine Relationen haben (Ahn et al., 2003). Der Kaltstart bezogen auf die Testautoren wird bei diesem Ansatz teilweise umgangen, da die Zielforscher das Bookmarking-System selbst nicht nutzen müssen, sondern es darauf ankommt, welche Nutzer deren Artikel in ihrer Bookmarking-Liste haben. Zudem steht hier zunächst der Vergleich der verschiedenen Ähnlichkeitsberechnungen im Vordergrund. Daher sollen keine Algorithmen untereinander verglichen werden. Alle Berechnungen werden mit dem Cosinus-Maß gemacht, da dieser zu guten Ergebnissen führt (Heck, 2011; Rorvig, 1999). Die Gegenüberstellung verschiedener Algorithmen soll in weiteren Projekten folgen.

Ein dem PageRank ähnliches Modell, das alle Beziehungen zwischen Nutzern, Ressourcen und Tags berücksichtigt, ist der FolkRank (Hotho et al., 2006). Hierbei werden die Gewichtungen der einzelnen Relationen im Graphen weitergegeben und es können je nach Bedarf Nutzer, Ressourcen oder Tags vorgeschlagen werden. Der folgende Ansatz nimmt keine Gewichtungen der Relationen vor, sondern misst die Ähnlichkeit unter den Autoren in CiteULike entweder basierend auf Nutzern oder basierend auf Tags. Eine gewichtete Methode wäre als weitere Methode denkbar. Allerdings lässt sich die FolkRank Variante nicht ohne Modifikationen auf die Co-Zitationsanalyse und das Bibliographic Coupling anwenden.

Neben Expertenempfehlungen für Wirtschaftsunternehmen (u.a. Cai et al., 2011; Reichling & Wulf, 2009) werden RS für Webnutzer und Wissenschaftler konstruiert. Au Yeung et al. (2009) analysieren dafür Relationen im Bookmarking System Del.icio.us, wobei sie einen Experten dadurch definieren, dass dieser viele wichtige Ressourcen (solche, die viele wichtige Nutzer haben) als erster in seiner Bookmark Liste hat. Verglichen mit diesem Ansatz sind die wichtigen Ressourcen hier die Artikel des Zielautors, dem relevante Wissenschaftler vorgeschlagen werden sollen. Haben viele Nutzer Artikel von Zielautor a und einem zweiten Autor b gemeinsam in ihrer Liste, erhöht sich die Ähnlichkeit zwischen den beiden Autoren. Weitere Ansätze für Experten- und Ressourcenempfehlungen sind u.a. Blazek (2007), der sich auf Literaturvorschläge für junge Wissenschaftler konzentriert, Cabanac (2010), der Nutzernetzwerke für die Empfehlung relevanter Artikel analysiert, Heck & Peters (2010), die Expertenempfehlungen auf Grundlage der Daten in drei verschiedenen Bookmarking-Systemen testen sowie Zanardi & Capra (2008), die Tagähnlichkeiten vergleichen und für Empfehlungen Anfrageterme der Nutzer berücksichtigen.

Ein Aspekt bei RS, der nicht vernachlässigt werden sollte, ist die Evaluation der Systeme. Neben Tests zur Effektivität und Genauigkeit der RS ist vor allem die Bewertung durch die Nutzer wichtig, da diese die Empfehlungen als hilfreich und relevant empfinden müssen (Herlocker et al., 2004). Neben den Modellbasierten Evaluationen sind daher die direkten Nutzerevaluationen wichtig. McNee et al. (2006) haben aufgezeigt, welche Aspekte bei RS beachtet werden müssen, damit sie auch von den Nutzern als Wissensmanagement-Tool angenommen werden. Da der hier vorgestellte Ansatz individuelle Empfehlungen von relevanten und neuen Autoren für einen Wissenschaftler als Ziel hat, lassen wir die Ergebnisse von den Forschern bewerten. Wichtig ist, dass der Zielautor die Empfehlungen als relevant für seine aktuelle wissenschaftliche Arbeit sieht.

3. Autoren Empfehlungen

3.1 Collaborative Filtering mit CiteULike

Mit SBS wie BibSonomy, Connotea und CiteULike können Forscher ihre Literatur online verwalten, sie anderen zugänglich machen und selbst auf die Literatur aller Nutzer zugreifen (Linde & Stock, 2011). Solche Systeme weisen eine Folksonomy basierte Struktur auf. Eine Folksonomy (Marinho et al., 2011; Peters, 2009) kann als Tupel $F = (U, T, R, Y)$ definiert werden, wobei die Mengen U , T und R die Elemente Nutzernamen, Ressourcen und Tags enthalten und Y deren Relationen zueinander. In dieser tri-partiten Struktur können mit CF Ähnlichkeiten zwischen Nutzern, Tags und Ressourcen (Bookmarks) mit Hilfe der Relationen gemessen werden. Unterteilt man die Folksonomy, erhält man drei Untermengen: die Docsonomy $DF := (T, R, Z)$ mit $Z \subseteq T \times R$, die Personomy $PUT := (U, T, X)$ mit $X \subseteq U \times T$ und die Individuelle Bookmark Liste (PBL) $PBLUR := (U, R, W)$ mit $W \subseteq U \times R$.

Da unsere Zielforscher Autorenempfehlungen erhalten sollen, sind primär nicht die Nutzer selbst, sondern deren Tags und Bookmarks von Bedeutung, die einen Artikel von Zielforscher a markieren. Wir erweitern daher unsere Folksonomy um die Menge A mit den Elementen Autoren: $F_E := (U, T, R, A, Y)$ mit der Relation $Y \subseteq U \times T \times R \times A$.

Im Folgenden enthält die Menge U_a (bzw. U_b) alle Nutzer, die mindestens einen Artikel von Zielautor a (bzw. Autor b) in ihrer Bookmark-Liste haben, die Menge R_a alle Ressourcen (hier: wissenschaftliche Artikel), die mindestens zwei Tags gemeinsam mit einem Artikel von Zielautor a haben, und die Menge T_a (bzw. T_b) alle Tags, die mindestens einem gebookmarkten Artikel des Zielautors a (bzw. b) zugeordnet sind. Es gibt zwei Möglichkeiten, die Datenbasis für die Empfehlungen zu ermitteln:

1. Wir suchen alle Nutzer $u \in U_a$, die mindestens einen Artikel von a in ihrer Bookmark Liste haben: $U_a = \{u \in U \mid \exists r \in R, a \in A, (u, r, a) \in Y\}$.

2. Wir suchen nach allen Ressourcen, denen ein Nutzer den gleichen Tag zugefügt hat wie einem der Artikel von a : $R_a := \{r \in R \mid t \in T_a, (r, t) \in Y\}$.

Der Nachteil der ersten Methode in unserem Fall war, dass es nur wenige Nutzer mit Artikeln von a gab und die Ähnlichkeitsberechnungen auf dieser Basis schwierig sein können (Lee & Brusilovsky, 2010a). Daher wird die Methode 2 zur Ermittlung der Datenbasis genommen, d.h. Bookmarks sind sich ähnlich, wenn ihnen gleiche Tags hinzugefügt wurden. Unsere Annahme ist nun, dass sich auch die Autoren dieser Bookmarks, sprich der gebookmarkten wissenschaftlichen Artikel, ähnlich sind, da die Nutzer in CiteULike ihre Artikel mit den gleichen Tags beschrieben haben. Tags stellen die thematische Beziehung zwischen sich und dem Artikel bzw. dem Autor her. Diese liefert Aussagen über das Forschungsgebiet eines Autors, was für die Empfehlung eines Kooperationspartners wichtig ist. Unsere Datenbasis für CF in CiteULike ergibt sich somit wie folgt:

$$R_a := \{r \in R \mid t \in T_a, (r, t) \in Y \text{ with } |T_a| \geq 2\}. \quad (1)$$

wobei die Ressourcen mindestens zwei gleiche Tags $t \in T_a$ aufweisen müssen. Somit ist ausgeschlossen, dass die Ähnlichkeit nur auf einem Tag beruht, was dann kritisch werden kann, wenn es sich um einen sehr allgemeinen Tag handelt, wie z.B. in einem unserer Fälle die Tags „nanotube“ oder „spectroscopy“.

Mit dem Cosinus-Maß werden Autorenähnlichkeiten nun auf zwei Wegen ermittelt: A) Zum einen basierend auf gleichen Tags (CULT = CF in CiteULike basierend auf gleichen Tags) und B) Zum anderen basierend auf gleichen Nutzern (CULU = CF in CiteULike basierend auf gleichen Nutzern):

$$A) \text{sim}(a, b) := \frac{|T_a \cap T_b|}{\sqrt{|T_a| * |T_b|}} \quad B) \text{sim}(a, b) := \frac{|U_a \cap U_b|}{\sqrt{|U_a| * |U_b|}} \quad (2)$$

Dabei ist zu beachten, dass die auf Nutzer basierte Variante (2b) nicht der Methode 1 bei der Datenerhebung entspricht. Es kann bei Methode 2 durchaus sein, dass Nutzer außer Acht gelassen werden, die einen Artikel von a in ihrer Bookmark-Liste haben. Methode 2 berücksichtigt aber alle Nutzer, die ähnliche Artikel (basierend auf Tags) wie die Artikel von a haben. Beide Varianten der Ähnlichkeitsberechnung (Formel 2) werden auf Grundlage von Methode 2 (Formel 1) berechnet. Haben bei Artikel des Zielforschers a jedoch keine Tags, wurde nach den Nutzern gesucht, die Artikel von a gebookmarkt haben (siehe Abschn. 4).

3.2 Autoren Co-Zitation and Bibliographic Coupling bei Autoren

Betrachtet man die Beziehungen zwischen zwei Autoren, können vier Relationstypen ausgemacht werden: Co-Autoren, direkte Zitationen, Bibliographic Coupling und Autoren Co-Zitationen. Die beiden ersten Relationen werden in diesem Ansatz nicht berücksichtigt, da einem Zielforscher hauptsächlich Wissenschaftler vorgeschlagen werden sollen, die er noch nicht kennt, d.h. Co-Autoren und zitierte Autoren sind ihm bereits bekannt. Daher betrachten wir Bibliographic Coupling (BC) und Co-Zitationen (CC). BC und CC sind ungerichtete, gewichtete Verbindungen zwischen zwei wissenschaftlichen Artikeln, die entweder an gemeinsamen Referenzen (BC) oder Co-Zitationen durch andere Artikel (CC) gemessen werden. Wir übertragen diese Verbindungen auf die Relationen zwischen zwei Autoren.

BC bei Autoren bedeutet, dass eine Beziehung zwischen zwei Autoren a und b besteht, wenn sie die gleichen Referenzen in ihren Artikeln nutzen. Diese Relationen ermitteln wir in der Datenbank WoS mit der Funktion „related records“. Diese ermöglicht die Auflistung gleicher Dokumente zu einem Artikel von Zielforscher a basierend auf der Anzahl gleicher Referenzen (Stock, 1999). Wir haben die Mengen D , Ref und A mit den Elementen Dokumente (wiss. Artikel), Referenzen (wiss. Artikel) und Autoren. Zwei Autoren a und b sind sich ähnlich, wenn ihre Artikel eine hohe Anzahl gleicher Referenzen aufweisen. Wir nehmen weiter an: Sind diese gemeinsamen Referenzen auf wenige Artikel verteilt, so erhöht dies die Ähnlichkeit, d.h. die Anzahl der gemeinsamen Referenzen pro Artikel ist entscheidend. Zum Beispiel: Autor a hat sowohl mit Autor b als auch mit Autor c sechs gemeinsame Referenzen. Bei a und b verteilen sich diese Referenzen auf jeweils zwei Artikel, bei c jedoch auf sechs. Für uns ist dann Autor a ähnlicher zu b als zu c .

Daher dient folgende Datenbasis als Grundlage: Wir nehmen alle Autoren mit Dokumenten, die über die gemeinsamen Referenzen den Dokumenten des Zielforschers a ähnlich sind, mit einer Höchstzahl n an gemeinsamen Referenzen:

$$D_{BC} := \{d \in D \mid |Ref_{dj} \cap Ref_{da}| \geq n, n \in \mathbb{N}\} \quad (3)$$

wobei Ref_{dj} die Anzahl der Referenzen eines Dokuments $d \in D$ von Autor $j \in A$ ist und Ref_{da} die Anzahl der Referenzen eines Dokuments $d \in D$ von Zielautor $a \in A$. Für jeden Zielforscher werden so die ihm

Users), 4. Ein Cluster basiert auf CF und gleiche Tags in CiteULike (CULT = CiteULike Tags). Die Ähnlichkeiten der Autoren zum Zielautor a werden mit dem Cosinus-Maß gemessen. Zusätzlich können die Ähnlichkeiten zwischen allen Autoren untereinander gemessen werden (nur innerhalb der

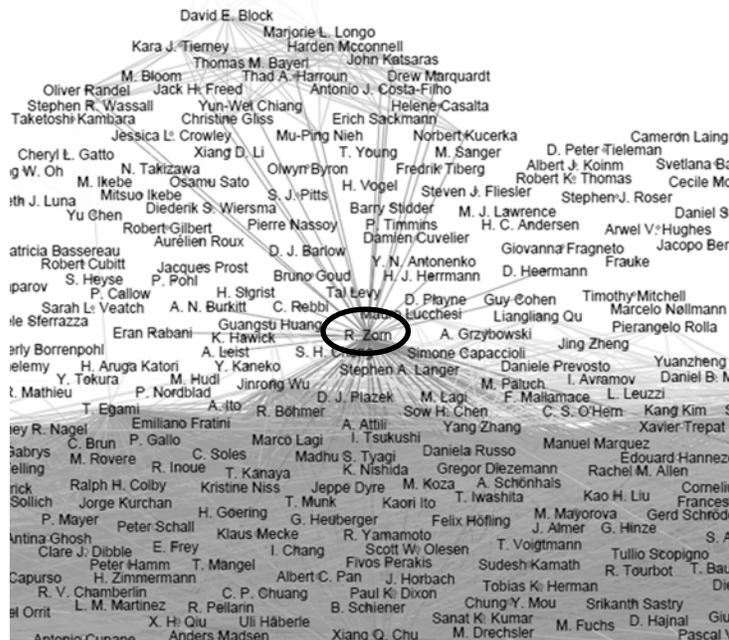


Abb. 4. Auszug aus CULU Graph, Kreis zeigt Testautor, Cosinus Intervall 0,99-0,49.

Datensets). Somit erhalten wir vier Cluster mit den Ähnlichkeiten aller Autoren untereinander. Diese Cluster werden als Graphen mit der Software Gephi¹ visualisiert und zeigen die Autorenrelationen (Abb. 1 bis 4). Die Größe der Knoten (=Autorenennamen) basiert auf den Relationen der einzelnen Autoren, die Größe der Kanten entspricht den Ähnlichkeits-Werten, d.h. die Stärke des Cosinus-Maß. Um die Graphen v.a. für die Evaluation übersichtlich zu halten, sind bei den größeren CiteULike Graphen Schwellenwerte beim Cosinus-Maß gesetzt worden. Zudem wurden bei diesen Graphen diejenigen Autorenpaare mit der höchsten Ähnlichkeit von $\text{sim} = 1$ herausgenommen, wenn deren Ähnlichkeit nur auf einem gemeinsamen Nutzer bzw. auf einem gemeinsamen Tag beruhte.

Bei einem der sechs Forscher wurden in CiteULike keine Nutzer und daher auch keine Tags gefunden, d.h. der Dienst selbst stellte die Artikel ins System. Daher konnten für diesen Autor (Autor 6 in Abb. 5) keine CiteULike Cluster und Graphen ermittelt werden. Bei einem weiteren Autor (Autor 5) hatten die Artikel in CiteULike keine Tags. In diesem Fall wurden in CiteULike alle Nutzer mit ihren gebookmarkten Ressourcen gesucht, die mindestens einen Artikel von Autor 5 in ihrer Bookmark-Liste haben.

Bei den Cosinus-Werten im BC Cluster fällt auf, dass diese um einiges höher sind bei in ACC, CULU und CULT, was daran liegen mag, dass die Anzahl der Referenzen der Autoren sehr hoch ist. Dementsprechend sind die Ähnlichkeitswerte in CULU und CULT verglichen mit BC und ACC höher, da diese auf entsprechend wenigen Nutzern und Tags basieren.

¹ <http://gephi.org/>

5. Evaluation

Die Evaluation in halb-strukturierten Interviews gliederte sich in drei Teile: Teil 1 bestand aus Fragen u.a. zum Rechercheverhalten und der Kooperationen mit Wissenschaftlern. In Teil 2 ging es um die empfohlenen Autoren: Von allen vier Clustern eines Testautors wurden jeweils die zehn ähnlichsten Autoren genommen und dem Physiker in einer Liste in alphabetischer Reihenfolge vorgelegt. Der Befragte sollte folgende Fragen beantworten:

1. Kennen Sie den Autor?
2. Haben Sie mit diesem Autor schon einmal kooperiert?
3. Ist dieser Autor Ihnen ähnlich bezogen auf den Forschungsschwerpunkt?
4. Wie wichtig ist der Autor für Ihre aktuelle Forschung (Bewertung von "nicht wichtig" (1) bis sehr wichtig (10))?
5. Mit wem der bekannten Autoren würden Sie kooperieren?
6. Was sind die Gründe für eine Kooperation bzw. Nicht-Kooperation?
7. Fehlen in der Liste Autoren, die für Ihre aktuelle Forschung wichtig sind?

Teil 3 beinhaltete die Evaluation der Graphen. Hierbei ging es um die richtige Verteilung der Autoren entsprechend ihren Forschungsgebieten, die Einordnung des Zielautors in die Graphen und die Relevanz der Graphen:

1. Entspricht Ihrer Einschätzung nach die Verteilung der Autoren der Realität, bezogen auf Forschungsähnlichkeit und Kooperationen?
2. Ist die in den Netzwerken dargestellte Zusammengehörigkeit richtig und sehen Sie sich an der richtigen Stelle?
3. Würde dieser Graph mit Empfehlungen von Autorenkollegen Ihnen helfen, z.B. bei der Organisation eines Workshops oder um Kooperationspartner zu finden?
4. Bezogen auf Frage 3: Wie relevant ist für Sie der Graph (Bewertung von „nicht relevant“ (1) bis „sehr relevant“ (10))?

Die Befragung in Teil 1 soll hier kurz zusammengefasst werden: Wie eine frühere Studie (Heck & Peters, 2010) schon zeigte, arbeiten viele Wissenschaftler und auch unsere Physiker größtenteils in kleinen Forschungsteams (oft 2 bis 5 Personen). Die Partner werden nach ihren Forschungsschwerpunkten ausgesucht. Gemeinsame Interessen unter den Forschern hielten die Befragten für sehr wichtig. Es ist aber auch nicht ausgeschlossen, dass sich Partner in ihren Disziplinen gegenseitig ergänzen. Die Forschungsgebiete sind dabei jedoch auch sehr nah. Sind die Interessen zu ähnlich oder die Forschungsrichtung genau gleich, sahen die Forscher die Kollegen in manchen Fällen eher als Konkurrenten. Konkurrenz und zu wenig thematische Übereinstimmung waren die beiden Gründe, die bei der Befragung genannt wurden, wenn ein Autor nicht für eine Kooperation in Frage kam (Teil 2, Frage 6).

In Teil 2 werden diejenigen Autoren als relevant und wichtig angesehen, die der Zielforscher mit einer Bewertung von mindestens 5 eingestuft hat, als auch die Autoren, die nicht in der Top 10 Liste der vier Cluster waren und vom Befragten nachträglich hinzugefügt und als sehr wichtig für seine jetzige

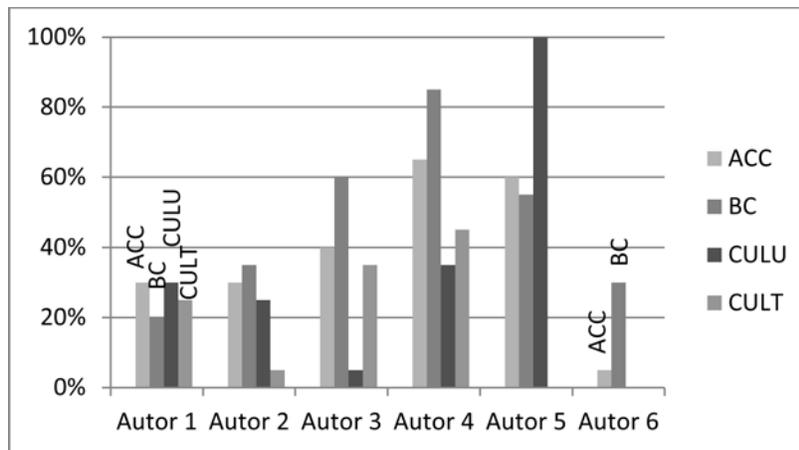


Abb. 5. Abdeckung der relevanten Autoren unter den Top 20 Autoren eines Clusters.

Forschung angesehen wurden. Abbildung 5 zeigt die Abdeckung der relevanten Autoren in den einzelnen Clustern gemessen an den ersten 20 ähnlichsten Autoren. Bei Zielautor 1 werden 30 % der Top 20 Autoren im Co-Zitations Cluster (ACC) als relevant eingestuft, beim Bibliographic Coupling Cluster (BC) sind es nur 20 %, im CiteULike Cluster basierend auf Nutzer (CULU) 30 % und beim CiteULike Cluster basierend auf Tags 25 %. Die Werte der sechs Probanden weisen teilweise große Unterschiede auf. Auffallend ist u.a., dass bei Zielautor 5 (keine Tags in CiteULike) CULU die volle Abdeckung hat, d.h. alle 20 Top Autoren aus CULU sind als relevant eingestuft worden. CULU und CULT schneiden in einigen

Tab. 1. Relevante Autoren in den Cluster am Beispiel von Zielautor 1. In Klammern: Autoren, die nur in einem Cluster gefunden wurden.

	ACC in Scopus	BC in WoS	CF in CiteULike
ACC in Scopus	11 (3)	0	8
BC in WoS	0	4 (3)	1
CF in CiteULike	8	1	15 (6)

Fällen auch besser ab, wenn alle Autoren eines Clusters betrachtet werden. Zum Beispiel sind sechs der von Zielautor 1 als relevant eingestufteten Autoren (insgesamt 29) nur in der über die Tags ermittelten CiteULike Datenbasis zu finden und nicht in den beiden anderen Clustern BC und ACC (Tab. 1). Bei Zielautor 2 sind es fünf von 19 wichtige Wissenschaftler, die nur in CiteULike gefunden wurden.

Der teils große Unterschied bei der Cluster Abdeckung mag daran liegen, dass einige der Befragten angegeben haben, heute anderen Forschungsinteressen nachzugehen oder sich ihre Schwerpunkte verlagert haben. Die Ähnlichkeitsberechnungen berufen sich auf die veröffentlichten Artikel der Forscher in den zurückliegenden fünf Jahren. Mögliche relevante Partner für die neuen Schwerpunkte können

daher nicht oder nur schwer gefunden werden, da die Forscher dazu noch wenig bis kaum Artikel veröffentlicht haben. Bei ACC kommt hinzu, dass die Artikel nicht gleich nach der Veröffentlichung zitiert werden, d.h. ältere Artikel mit einer schon höheren Zitationsrate kommen schwerer zum Tragen. Da wir die Zielforscher baten, nur die Autoren als wichtig einzustufen, die auch aktuell für ihre Forschung relevant sind, fallen die ehemals relevanten Wissenschaftler, die in den Clustern gefunden wurden, aus diesem Raster. Würden diese miteingerechnet, würde die Zahl der gefundenen wichtigen Forscher bei einigen Befragten leicht nach oben steigen.

Im dritten Evaluationsteil beurteilten die Physiker die Graphen. Die durchschnittliche Relevanz der Graphen (Bewertung von 1 bis 10) beträgt:

- ACC in Scopus: 5,08
- BC in WoS: 8,7
- CULU: 2,13
- CULT: 5,25

Die Werte von CULU und CULT basieren nur auf den Bewertungen von vier Forschern, da einmal keine Daten in CiteULike gefunden wurden und Zielautor 5 zu wenig Daten für den Graphen hatte. Zwei Forscher gaben BC und CULT gute Werte und schlugen vor, diese beiden Graphen zu kombinieren um möglichst viele relevante Kooperationspartner vorgeschlagen zu bekommen. In BC und ACC wurden von einigen wichtige Wissenschaftler vermisst. Zudem zeigten diese Graphen zwar relevante Forscher, aber diese seien zum großen Teil schon bekannt. In CULU und CULT dagegen würden sich auch Forscher finden, die vom Forschungsschwerpunkt zwar weiter von den Testautoren entfernt wären, aber z.B. für Workshops und Konferenzen auch interessant seien. Bei Betrachtung der Autorennetzwerke fielen fast allen Testforschern relevante Kollegen auf, die Ihnen bei der Befragung zuvor nicht eingefallen waren. Dies sei sehr hilfreich, da man sich nicht immer an alle Kollegen sofort erinnere, z.B. wenn es darum geht, wissenschaftliche Fragestellungen zu lösen oder Treffen zu organisieren. Bei den größeren CiteULike Graphen wurde auch angegeben, dass sich dort eventuell viele unbekannte, potentielle Partner finden ließen, was einen großen Mehrwert bringe. Um jedoch die Relevanz der unbekanntem, empfohlenen Wissenschaftler für ihre Forschung eindeutig zu bewerten, hätten sich die Zielforscher deren Publikationen ansehen müssen. Da die Autorenverteilungen in den Graphen aber als gut und der Realität entsprechend bewertet wurden, kann angenommen werden, dass auch die unbekanntem Autoren in der Regel ihrem Forschungsgebiet richtig eingeordnet wurden. Vor allem dann, wenn diese Autoren eindeutig einer Forschungsgruppe im Graphen zugehörig sind und der Zielautor die Gruppe eindeutig ihrer Forschung zuordnen konnte.

Ein wichtiger Aspekt für die Befragten war die gute Anordnung in den Graphen, d.h. Abspaltungen von Autorengruppen und breitere Verteilungen lieferten eine bessere Übersicht. Auch die Anzahl der gezeigten Autorennamen spielt eine Rolle. In den CiteULike Graphen (Abb. 3 u. 4) kam teilweise die geringe Datenmenge zum Tragen: Wenn es nur wenige Tags oder Nutzer gibt, zeigen die Graphen keine gute Verteilungen bzw. Abgrenzungen einzelner Autorengruppen auf. Für die Forscher war es dementsprechend schwieriger, relevante Informationen, sprich wichtige Autoren und Gruppen, zu

erkennen und zu bewerten. Hilfreich wäre hierbei eine zusätzliche Kategorisierung der Wissenschaftler, z.B. nach Tags, die Aufschlüsse über Forschungsgebiete brächten.

6. Ausblick

In diesem Ansatz haben wir drei verschiedene Modelle zu Autorenempfehlungen für Akademiker und Forscher analysiert. Zwei in der Scientometrie bekannte Methoden, BC und ACC, wurden mit CF verglichen. Erste Ergebnisse zeigen die Unterschiede auf, die darauf hindeuten, dass die Kombination dieser Methoden die Empfehlungen von relevanten Kooperationspartnern verbessern würde. Ähnlichkeitsberechnungen basierend auf Nutzern und Tags in CiteULike führte, trotz der teils geringen Datenmenge, zu relevanten Ergebnissen, die Berechnungen mit BC und ACC gut komplementieren könnten. Die Forscher bestätigten diese Resultate mit den Relevanzbewertungen der Graphen (CULT ist der am zweitbesten bewertete Graph). Autorenempfehlungen können helfen, schnell relevante Partner auf Basis wissenschaftlicher Reputation zu finden. Der Artikel zeigt hierzu vier neue Ansätze. Die weitere Herausforderung ist nun, diese Ansätze zu verbinden. Ein gewichteter Algorithmus, der die einzelnen Autorenrelationen berücksichtigt, sollte dabei den relevanten Autoren für einen Zielforscher die höchsten Ähnlichkeitswerte zuordnen. Weitere Ansätze, die berücksichtigt werden sollten, sind der FolkRank (Hotho et al., 2006) und „Expertise Analysis mit ‚SPEAR‘“ (Au Yeung et al., 2009).

7. Danksagung

Tamara Heck wird durch den Strategischen Forschungsfonds der Heinrich-Heine-University Düsseldorf gefördert. Danke an Oliver Hanraths für die Datenerhebung, Stefanie Haustein für hilfreiche Ratschläge und den Wissenschaftlern für die Teilnahme an der Evaluation.

8. Referenzen

Ahn, H. J. (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, 178, 37-51. doi:10.1016/j.ins.2007.07.024

Au Yeung, C. M., Noll, M., Gibbins, N., Meinel, C., & Shadbolt, N. (2009). On measuring expertise in collaborative tagging systems. *Proceedings of Web Science Conference: Society On-Line*. Athens, Greece. Retrieved from <http://journal.webscience.org/109>

Berkovsky, S., Kuflik, T., & Ricci F. (2007). Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18 (3), 245-286.

doi:10.1007/s11257-007-9042-9

- Blazek, R. (2007). Author-Statement Citation Analysis Applied as a Recommender System to Support Non-Domain-Expert Academic Research. Doctoral Dissertation. Fort Lauderdale, FL: Nova Southeastern University. Retrieved from <http://gradworks.umi.com/32/78/3278195.html>
- Bogers, T., & van den Bosch, A. (2008). Recommending scientific articles using CiteULike. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp. 287-290). Lausanne, Switzerland: ACM. doi:10.1145/1454008.1454053
- Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation. Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389-2404. doi: 10.1002/asi.21419
- Cabanac, G. (2010). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, 87(3), 597-620. doi:10.1007/s11192-011-0358-1
- Cacheda, F., Carneiro, V., Fernández, D., & Formoso, V. (2011). Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web*, 5(1), article 2. doi:10.1145/1921591.1921593
- Cai, X., Bain, M., Krzywicki, A., Wobcke, W., Kim, Y. S., Compton, P., & Mahidadia, A. (2011). Collaborative filtering for people to people recommendation in social networks. In J. Li (Ed.), *Lecture Notes in Computer Science, Vol. 6464. AI 2010: Advances in Artificial Intelligence* (pp. 476-485). Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-642-17432-2_48
- Desrosiers, C. , & Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. In F. Ricci, L. Rokach, B. Shapira & P.B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 197-144). New York, NY: Springer-Verlag.
- Gmur, M. (2003). Co-citation analysis and the search for invisible colleges. A methodological evaluation. *Scientometrics*, 57(1), 27-57. doi: 10.1023/A:1023619503005

- Heck, T. (2011). A comparison of different user-similarity measures as basis for research and scientific cooperation. *Proceedings of the Information Science and Social Media International Conference*, 117-128. Retrieved from <http://issuu.com/informationsvetenskap/docs/issome2011>
- Heck, T., & Peters, I. (2010). Expert recommender systems: Establishing Communities of Practice based on social bookmarking systems. *Proceedings of the 10th International Conference on Knowledge Management and Knowledge Technologies*, 458-464. Retrieved from <http://www.phil-fak.uni-duesseldorf.de/?id=12276>
- Herlocker, J. L., Konstan, J. A., Terveen L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), 5-53.
doi:10.1145/963770.963772
- Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Y. Sure & J. Domingue (Eds.), *Lecture Notes in Computer Science, Vol. 4011. The Semantic Web: Research and Applications* (pp. 411-426). Heidelberg, Germany: Springer-Verlag.
- Lee, D. H., & Brusilovsky, P. (2010a). Social networks and interest similarity. The case of CiteULike. In *Proceedings of the 21st ACM Conference on Hypertext & Hypermedia* (pp. 151-155). Toronto, Canada: ACM. doi:10.1145/1810617.1810643
- Lee, D. H., & Brusilovsky, P. (2010b). Using self-defined group activities for improving recommendations in collaborative tagging systems. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 221-224). Barcelona, Spain: ACM. doi:10.1145/1864708.1864752
- Linde, F. and Stock, W.G. (2011). *Information markets*. Berlin, Germany, New York, NY: De Gruyter Saur.
- Marinho, L. B., Nanopoulos, A., Schmidt-Thieme, L., Jäschke, R., Hotho, A., Stumme, G., & Symeonidis, P. (2011). Social tagging recommenders systems. In F. Ricci, L. Rokach, B. Shapira & P.B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 615-644). New York ,NY: Springer-Verlag.

- McNee, S. M., Kapoor, N., & Konstan, J.A. (2006). Don't look stupid. Avoiding pitfalls when recommending research papers. In *Proceedings of the 20th anniversary Conference on Computer Supported Cooperative Work* (pp. 171-180). Banff, Canada: ACM. doi:10.1145/1180875.1180903
- Meho, L. I., & Sugimoto, C. R. (2009). Assessing the scholarly impact of information studies. A tale of two citation databases – Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, 60(12), 2499-2508. doi:10.1002/asi.v60:12
- Parra, D., & Brusilovsky, P. (2009). Collaborative filtering for social tagging systems. An Experiment with CiteULike. In *Proceedings of the 3rd ACM Conference on Recommender Systems* (pp. 237-240). New York, USA : ACM. doi:10.1145/1639714.1639757
- Peters, I. (2009). *Folksonomies. Indexing and retrieval in Web 2.0*. Berlin, Germany: De Gruyter Saur.
- Ramezani, M., Bergman, L., Thompson, R., Burke, R., & Mobasher, B. (2008). Selecting and applying recommendation technology. *Proceedings of the International Workshop on Recommendation and Collaboration, in Conjunction with 2008 International ACM Conference on Intelligent User Interfaces*. Retrieved from <http://maya.cs.depaul.edu/~mobasher/pubs.html>
- Reichling, T., & Wulf, V. (2009). Expert recommender systems in practice. Evaluating semi-automatic profile generation. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (pp. 59-68). Boston, USA: ACM. doi: 10.1145/1518701.1518712
- Rendle, S., Marinho, L. B., Nanopoulos, A., & Schmidt-Thieme, L. (2009). Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 727-736). Paris, France: ACM. doi:10.1145/1557019.1557100
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (pp. 175-186). Chapel Hill, USA: ACM. doi:10.1145/192844.192905

- Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science and Technology*, 50(8), 639-651. doi: 10.1002/(SICI)1097-4571(1999)50:8<639::AID-ASI2>3.0.CO;2-C
- Shepitsen, A., Gemmell, J., Mobasher, B., & Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp.259-266). New York, USA: ACM. doi:10.1145/1454008.1454048
- Stock, W. G. (1999). Web of Science. Ein Netz wissenschaftlicher Informationen – gesponnen aus Fußnoten [Web of Science. A web of scientific information – cocooned from footnotes]. *Password*, no. 7+8, 21-25. Retrieved from http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/1/1197986684stock120_h.htm
- Zanardi, V. & Capra, L. (2008). Social ranking: Uncovering relevant content using tag-based recommender systems. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp.51-58). New York, USA: ACM. doi:10.1145/1454008.1454018
- Zhao, D., & Strotmann, A. (2011). Counting first, last, or all authors in citation analysis. Collaborative stem cell research field. *Journal of the American Society for Information Science and Technology*, 62(4), 654-676. doi:10.1002/asi.21495