

# Entwicklung von Testkollektionen für P2P Information Retrieval

Gregor Heinrich, Sven Teresniak, Hans Friedrich Witschel  
Universität Leipzig  
Institut für Informatik  
Augustusplatz 10-11  
04109 Leipzig  
{heinrich,teresniak,witschel}@informatik.uni-leipzig.de

**Dieser Beitrag stellt laufende Arbeiten an Testkollektionen vor, die für die realistische Simulation semantischer Suche in P2P-Netzwerken verwendet werden können. Er entwickelt eine allgemeine Vorgehensweise für die Erstellung solcher Simulationsdaten. Mithilfe eines probabilistischen Modells wird die semantische Verteilung von Peer-Dokumentenbeständen und Suchanfragen abgedeckt, wobei die verwendeten Dokumente keine Vorklassifizierung benötigen.**

## 1 Einleitung

Ein zentrales Problem bei der Simulation von P2P Information Retrieval-Systemen ist die Modellierung der Nutzercommunity. Speziell bei Ansätzen, welche semantische Ähnlichkeiten für Routing verwenden (siehe z.B. [WB04]), ist die Verwendung realistischer statistischer Verteilungen semantischer Parameter wichtig, um für Simulationen ausreichend genaue Vorannahmen über den Zustand eines P2P-Netzwerkes zu Simulationsbeginn treffen zu können. Zumeist sind simulative Experimente vonnöten, da – gerade bei Neuentwicklungen – in der Regel keine ausreichend große Nutzergruppe für Messungen verfügbar ist.

Überraschenderweise gibt es keine Referenzmodelle bzw. Testkollektionen, die das komplette Verhalten der Community beschreiben. Traditionelle Testkollektionen, wie sie beispielsweise für verschiedene Information-Retrieval-Tasks von der Text Retrieval Conference (TREC) zur Verfügung gestellt werden, beinhalten eine Menge von Textdokumenten und einen Satz Anfragen mit zugehörigen Relevanzbewertungen. Diese Daten können jedoch bestenfalls einen geringen Teil des benötigten Modells abbilden.

Im Unterschied zu verfügbaren Test-Kollektionen (z.B. TREC) entstehen zunächst zwei zusätzliche Anforderungen an eine Test-Kollektion für P2P Information Retrieval:

- Verteilung von Dokumenten auf die Peers,
- Verteilung von Suchanfragen auf Peers; diese sind i.a. von den lokalen Dokumenten semantisch abhängig, enthalten aber oft auch neue Aspekte.

Ziel dieses Beitrags ist es, Grundlagen eines Modells für Nutzerverhalten zu entwickeln, das die semantische Verteilung von Inhalten und Anfragen in einer Peer-Community allgemein beschreiben kann und es ermöglicht, aus vorhandenen Corpora eine Testkollektion zu generieren, die in Simulationen auch großer Netzwerke verwendet werden kann. Dazu soll ein Vorgehen definiert werden, das es erlaubt,

1. Verteilungen in einem Modellbetrieb eines P2PIR-Systems zu messen und
2. diese Verteilungen auf frei verfügbare Corpora bzw. herkömmliche Testkollektionen zu übertragen, um so zu neuen P2P-Testkollektionen zu gelangen.

Ein Ergebnis dieses Vorgehens ist eine (möglichst einfache) Vorschrift, mit welcher sowohl die Dokumente als auch die Suchanfragen einer TREC-Kollektion auf eine beliebige Anzahl Peers verteilt werden.

Weitere wichtige Aspekte bei der Simulation von P2P-Netzwerken ergeben sich aus der Dynamik solcher Systeme: Frequenz von Anfragen eines Peers, Downloadverhalten, Netzverhalten bei Überlastsituationen sowie Offlinestatus („Churn“). Solche Probleme sollen in diesem Beitrag jedoch nicht diskutiert werden. Aspekte der Dynamik können in einer späteren Phase hinzugenommen werden bzw. sollen zunächst der individuellen Gestaltung bei der Erstellung einer Simulationsumgebung überlassen werden.

## 2 Verwandte Arbeiten

Vorarbeiten zum Thema Verteilung von Dokumenten und Anfragen auf Peers finden sich in Crespo und Garcia-Molina [CGM02] und Schlosser et al. [SCK03]. Dort werden Verteilungen von Themen vorgeschlagen, die die Interessen jedes Peers beschreiben und anhand derer Anfragen gestellt werden. Die Annahmen für diese Themenverteilungen sind dort jedoch stark vereinfacht und erfordern vorklassifizierte Daten. Zur eigentlichen Modellierung von Inhalten wird auf frei verfügbare Corpora zurückgegriffen (z.B. Open Directory oder CiteSeer), deren Themenverteilungen auf Peers übertragen werden – ein Ansatz, der für den Aufbau von Testkollektionen prinzipiell sinnvoll erscheint.

Auch Neumann et al. [NBMW06] legen sich auf eine bestimmte Testkollektion fest (Wikipedia), benutzen allerdings keine feste Klassifizierung der Daten. Diese werden stattdessen unter Ausnutzung der Verlinkung von Wikipedia-Artikeln mittels eines Graph-Cluster-Algorithmus geclustert. Die Verteilung der Artikel auf Peers geschieht durch eine weitere Unterteilung der Cluster in „Chunks“, die mittels eines Sliding Window auf Peers verteilt werden, um eine Überlappung der Dokumentenbestände zu erreichen.

Der Ansatz schlägt weiterhin vor, Queries aus dem Google-*Zeitgeist*-Archiv zu verwenden, für die allerdings keine echten Relevanzurteile vorhanden sind. Neben dem Fehlen von Relevanzurteilen ist bei diesem Ansatz vor allem problematisch, dass die Überlappung von Dokumenten nachträglich hergestellt werden muss; ein *fuzzy* Clustering würde diese Überlappung hingegen in natürlicher Weise beinhalten.

Schließlich ist noch der Ansatz von Cooper zu erwähnen [Coo04], welcher zunächst das Anlegen von Statistiken in Form von Histogrammen vorsieht, unter anderem über die Anzahl relevanter Dokumente pro Query, den Replikationsgrad von Dokumenten oder die durchschnittliche Anzahl relevanter Dokumente pro Query, die auf demselben Peer liegen. Die Histogramme werden dann – durch Multiplikation der diskreten Messwerte mit einer Konstanten – auf die gewünschte Größe skaliert. Aus den so gewonnenen Kenngrößen wird eine künstliche Testkollektion erzeugt, indem für eine Menge künstlicher Queries, Dokumente und Peers künstliche Relevanzurteile (Query-Dokument-Zuordnung) und die Zuordnung von Dokumenten zu Peers solange variiert wird, bis die künstliche Verteilung der realen gleicht.

Hierbei entsteht allerdings keine reale Testkollektion, sondern lediglich eine, deren Verhalten in Bezug auf die oben genannten Statistiken einer realen ähnlich ist. Hinzu kommt, dass die oben erwähnten Statistiken oft schwer zu beschaffen sind, da dies voraussetzt, dass für Dokumente und Queries einer realen P2P-Community Relevanzurteile bzw. Statistiken darüber vorhanden sind.

Um eine flexible Modellierung der Inhalte zu erlauben und trotzdem mit realen Daten arbeiten zu können, wird in diesem Beitrag die Verwendung latenter Konzepte als Träger semantischer Informationen über die Peers vorgeschlagen.

### 3 Modellbildung

Das hier vorgestellte Vorgehen zur Erzeugung einer P2PIR-Testkollektion beruht auf der Idee, die Parameter eines Modells durch empirische Datenerhebung zu spezifizieren und das so gewonnene Modell auf eine bestehende Testkollektion anzuwenden. Es ergibt sich folgendes Ein-/Ausgabe-Verhalten des zu entwerfenden Algorithmus:

- Eingabe: Modell  $M$  mit Parametersatz  $P$ , Testkollektion  $T$ , Anzahl Peers  $N$ , Kollektion  $U$  aus realem P2P-Betrieb, zugehörige Verteilung  $V$  der Dokumente bzw. Konzepte auf Peers in diesem realen Szenario.
- Ausgabe: Werte der Parameter aus  $P$ , Verteilung  $V'$  der Dokumente und Anfragen aus  $T$  über die  $N$  Peers derart, dass der Unterschied zwischen  $V$  und  $V'$  minimal ist.

Wichtig ist dabei die Skalierbarkeit und Generalisierbarkeit: Die Anzahl der Peers und Dokumente in  $T$  soll sowohl quantitativ als auch semantisch variabel sein, d.h. das im realen Betrieb empirisch parametrisierte Modell soll sich auf andere Peernetzwerkgrößen mit anderen Corpusgrößen und -inhalten übertragen lassen.

Als Grundlage des Modells werden Verfahren der latenten Konzeptanalyse vorgeschlagen. Dies erscheint sinnvoll, da im realen Nutzerbetrieb eines P2P-Systems meist keine Klassifikationsdaten vorhanden sind, wie sie beispielsweise der Ansatz von Schlosser et al. [SCK03] voraussetzt.

Bei Verfahren der latenten semantischen Analyse (LSA) [DDL+90] werden aus Textdaten sog. latente Konzepte nicht-überwacht extrahiert. Latente Konzepte sind

semantische Einheiten, deren gewichtete Kombinationen Dokumente und Wörter in Corpora gleichermaßen repräsentieren können und die zu einem gewissen Grade unabhängig von linguistischen Phänomenen wie Polysemie und Synonymie sind.

Die Fähigkeit latenter Konzepte, Semantik in einem Raum mit wählbarer, niedriger Dimension zu beschreiben, wird als geeignet angesehen, generalisierbare Modelle für Nutzerverhalten zu entwickeln, indem zunächst die Verteilung der latenten Konzepte untersucht wird und aus dieser die Verteilung der anderen, sichtbaren Objekte (z.B. der Dokumente) abgeleitet wird. Die nicht-überwachte Funktionsweise erlaubt dabei einerseits die Modellierung unbekannter, verteilter Corpora, wie sie im P2P-Betrieb zu erwarten sind, andererseits schließt sie Subjektivität bei der Klassifikation aus.

In Abb. 1a ist die Zuordnungsstruktur eines P2P-Netzwerkes als bipartiter Graph zwischen Dokumenten und Peers dargestellt. Ziel einer latent-semantischen Modellierung ist die Zuordnung von latenten Konzepten zu Peers und Dokumenten, dargestellt in Abb. 1b.

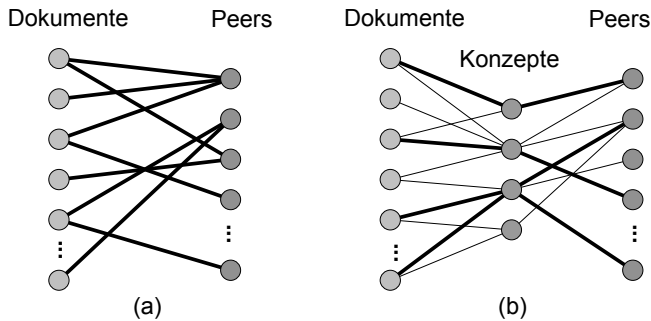


Abb 1. Zuordnung von (a) Dokumenten zu Peers bzw. (b) von Dokumenten und Peers zu Konzepten.

Zur Ermittlung dieser Konzepte haben sich – aufbauend auf dem ursprünglichen Ansatz von LSA<sup>1</sup> [DDL+90] – probabilistische Verfahren durchgesetzt, die verbesserte Retrievalergebnisse zeigen und relativ einfach erweiterbar sind. Deren wichtigster Vertreter ist Latent Dirichlet Allocation (LDA) [BNJ03], das Konzepte als latente multinomiale Wahrscheinlichkeitsvariablen in einem Bayesschen Netzwerk [Mur01] modelliert.

Für jedes Dokument  $d$  wird im LDA-Modell eine Verteilung  $p(z|d)$  von Konzepten betrachtet. Ein Konzept  $z$  wiederum besteht aus einer Verteilung  $p(w|z)$  über das Vokabular. Die Verteilung über die Wörter eines Dokuments ergibt sich als Mischmodell über alle Konzepte:  $p(w|d) = \sum_z p(w|z) p(z|d)$ , wobei die Wörter bei gegebenem Dokument als statistisch unabhängig angenommen werden. Die Parameter der Verteilungen  $p(z|d)$  und  $p(w|z)$  können aus realen Dokumenten mit geeigneten Verfahren geschätzt werden (Variational EM [BNJ03], MCMC [GS04], vgl. auch [Hei05]).

<sup>1</sup> LSA beruht auf Singulärwertzerlegung der Term-Dokument-Matrix und Semantikvergleich mittels Cosinusmaß und kann als Spezialfall des Vektorraummodells angesehen werden.

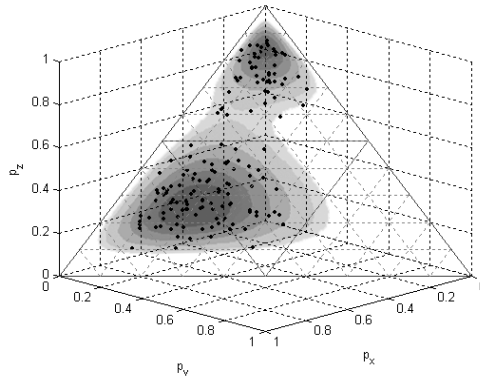


Abb. 2: Multinomiale Konzeptverteilungen einer Dokumentenmenge (Punkte) und geschätzte Dirichlet-Mischverteilung mit 2 Komponenten (Grauwerte).

Der einfachste Ansatz zur gleichzeitigen Modellierung von Peers und Dokumenten ist die Berechnung des Schwerpunkts der Multinomialparametervektoren der Peer-Dokumente („Centroid-LDA“ oder CLDA). Außerdem können algorithmische Erweiterungen von LDA betrachtet werden, die Konzepte zu Gruppierungen von Dokumenten im Modell selbst abbilden, um z.B. Autorenaktivität [SSR+04][BG05] und Email-Kommunikation [MCE+04] zu modellieren.

Das Author-Topic-Modell (ATM) [SSR+04] ersetzt die Zuordnung von Konzepten zu Dokumenten  $p(z|d)$  durch eine solche zu deren Autoren  $p(z|a)$ , was im vorliegenden Fall für eine Konzeptermittlung für Peers genutzt werden kann und aufgrund anderer Dokumentengewichtungen nicht äquivalent zu CLDA ist. Mit einer eindeutigen Zuordnung zwischen Dokumenten und Autoren ist das ATM identisch zu LDA:  $p(z|d) = p(z|a(d))$ . Mit dieser Identität lässt sich das ATM so anpassen, dass sowohl Konzepte für Peers als auch Dokumente ermittelt werden können, also ein *Peer-Document-Topic-Model* (PDTM) entsteht. Die Menge realer Peers wird hierfür um eine Menge künstlicher Peers erweitert, von denen jeder jeweils ein Dokument enthält. Die Konzepte dieser künstlichen Peers entsprechen den Konzepten der Dokumente.

Ein alternativer Ansatz für gleichzeitige latent-semantische Modellierung von Dokumenten und Peers ist der Ansatz von PHITS [CH01], welcher in seiner Standardform sowohl Dokumente als auch Links zwischen Dokumenten mithilfe derselben latenten Variablen modelliert. Dieses Verfahren ließe sich analog zu oben auf ein P2P-Netz übertragen, indem man Links durch Peers ersetzt („PPHITS“).

Allerdings besteht bei allen genannten latent-semantischen Ansätzen das Problem, dass Dokumente oder Peers jeweils nur durch einen Konzeptmittelwert  $p(z|d)$  repräsentiert werden, ohne Aussagen über die semantische Streuung des Inhalts oder gar Clusterstrukturen zuzulassen. In Abb. 2 werden dreidimensionale Multinomialverteilungen auf einer Simplexfläche graphisch als Punkte dargestellt<sup>2</sup>. Jeder Punkt stellt ein Dokument dar. Um einen Peer zu beschreiben, liefern CLDA, ATM bzw.

<sup>2</sup> Ein Simplex ergibt sich als Wertebereich wegen der Normierungsbedingung von Wahrscheinlichkeitsverteilungen.

PDTM einen Punkt. Wichtige Eigenschaften eines Peers wie Streuung und Clusterstruktur von Peer-Inhalten gehen in dieser Repräsentation verloren.

Abhilfe kann eine alternative Modellierung des Corpus und seiner Peer-Zuordnung schaffen, die wie CLDA zweistufig aufgebaut ist: Mittels LDA werden in einem ersten Schritt die Dokumenten-Konzepte  $\theta_d = p(z|d)$  ermittelt. In einem zweiten Schritt wird die Streuung und Clusterstruktur der Peers gefunden, indem die Parameter  $y$  einer Verteilung  $p(\theta_{d(a)}|y)$  über die dem Peer (oder einer anderen Dokumentenmenge)  $a$  zugeordneten Multinomialverteilungen  $\theta_{d(a)}$  geschätzt werden. Hierfür schlagen wir analog zu Gaussian Mixture Models die Verwendung einer Dirichlet-Mischverteilung vor<sup>3</sup>. Abb. 2 zeigt eine solche Dirichlet-Verteilung für zwei Komponenten.

## 4 Alignierung mit realen Daten

Eines der zentralen Probleme bei der Erstellung von Testkollektionen ist die Übertragung einer Verteilung latenter Variablen auf Peers  $p(z|a)$ , welche man für ein reales P2P-System extrahieren kann, auf eine Testkollektion, beispielsweise TREC – die „Alignierung“ von  $T$  mit  $U$ .<sup>4</sup>

In diesem Beitrag wird der Alignierungsprozess zunächst für den Fall des im letzten Abschnitt beschriebenen Verfahrens CLDA umrissen sowie beschrieben, wie sich eine gefundene Alignierung bewerten lässt. Eine Entwicklung des eigentlichen Optimierungsprozesses für die Verteilung  $V'$  ist Gegenstand laufender Arbeiten und wird nur skizziert.

Der Vergleich einer gegebenen Verteilung von Dokumenten einer Testkollektion auf  $N$  Peers ( $N$  beliebig) mit einer realen Verteilung lässt sich mit Hilfe von Maßen realisieren, die auf dem Gebiet des (weichen) Clusterings eingesetzt werden: In [Mei02] wird ein Maß vorgeschlagen, „variation of information“ (VI), welches die Mutual Information zweier weicher Clusterings  $p(z_1|x)$ ,  $p(z_2|x)$  von Beobachtungen  $x$  benutzt, um diese zu vergleichen.

Dabei wird nicht vorausgesetzt, dass beide Clusterings aus der gleichen Anzahl von Clustern bestehen. Um nun Peer-Verteilungen  $V$  und  $V'$  mit unterschiedlichen Peeranzahlen zu vergleichen, kann das VI-Maß „reziprok“ verwendet werden, indem Clusterings  $p(a_U|z_U)$ ,  $p(a_T|z_T)$  mit identischen Konzeptzahlen verglichen werden, d.h. Peers werden als weiche Cluster von Konzepten aufgefasst. Dies beruht auf der Annahme, dass eine Zuordnung zwischen den Konzepten  $z_U$  und  $z_T$  existiert<sup>5</sup>.

Für diese Zuordnung schlagen wir ein frequenzbasiertes Verfahren vor: Für beide Kollektionen  $U$  und  $T$  wird das Verfahren LDA ausgeführt, so dass sich in beiden Kollektionen Dokumente als Mischung von Konzepten  $p(z|d)$  ergeben. Eingangs werden die Konzepte  $z$  miteinander aligniert, indem man sie nach ihrer Stärke  $\#z$  ordnet:  $\#z = \#\{d : p(z|d) > u\}$ , also die Anzahl der Dokumente, die das Konzept  $z$  mit einer

<sup>3</sup> Die Dirichlet-Verteilung ist die zur Multinomialverteilung konjugierte Verteilung [Mur01].

<sup>4</sup> Die Alignierung von Anfragen ist analog zu der von Dokumenten.

<sup>5</sup> Um Generalisierbarkeit zu erreichen, werden die Wortverteilungen  $p(w_U|z_U)$  und  $p(w_T|z_T)$  außer Acht gelassen.

Mindestwahrscheinlichkeit  $u$  enthalten. Anschließend werden Konzepte beider Kollektionen einander absteigend nach  $\#z$  zugeordnet.

Als nächstes müssen Peers zu Clustern  $z$  zugeordnet werden. Für das Corpus  $U$  kann diese Zuordnung direkt geschehen, nachdem die Schwerpunkte der Peer-Dokumentenmengen  $p(z_U|a_U)$  im zweiten Schritt des CLDA-Ansatzes gefunden wurden. Der Bayessche Satz liefert  $p(a_U|z_U) = \lambda p(z_U|a_U) p(a_U)$  mit Normalisierungskonstante  $\lambda$  und Prior  $p(a_U)$ , der abhängig von der Dokumentzahl  $\#d(a_U)$  des Peers gewählt wird. Für die Testkollektion  $T$  muss die Zuordnung zwischen Dokumenten und Peers  $V' = \{(d_T, a_T)\}$  gefunden werden, deren Konzeptrepräsentation  $p(a_T|z_T)$  mit  $p(a_U|z_U)$  verglichen werden kann.

Die Suche nach derjenigen Zuordnung  $V' = \{(d_T, a_T)\}$ , welche das Maß  $VI(V, V')$  minimiert, stellt das eigentliche Alignierungsproblem dar, das durch abwechselndes Mutieren von  $V'$  und Messung von  $VI$  gelöst werden kann. Der Suchraum kann hierbei durch geeignete Randbedingungen bei der Zuordnungsauswahl stark beschränkt werden. Als sinnvoll sehen wir den Ansatz an, hier die Popularität von Dokumenten  $\#a(d_U)$  und Peergrößenverteilung  $\#d(a_U)$  aus  $V$  einzubringen, die das  $VI$ -Maß beeinflussen. Mit den Konzepten populärer Dokumente in  $U$  lassen sich Dokumente mit ähnlichen Konzeptverteilungen in  $T$  finden, die durch Zuordnung zu vielen Peers  $\#a(d_T)$  populär gemacht werden können. Mutation erfolgt durch Verschieben von Dokumenten in  $V'$ .

Abb. 3 fasst den gesamten Ablauf der Alignierung noch einmal zusammen.

1. Ausführung LDA auf Kollektionen  $U$  und  $T$ .  
Ergebnis: Dokumente als Mischung von Konzepten  $p(z|d)$
2. Alignierung von Konzepten in  $T$  mit Konzepten aus  $U$
3. Vorgabe einer initialen Dokumentenverteilung  $V'$  auf Peers in  $T$ .
4. Berechnung von Konzeptschwerpunkten für alle Peers in  $U$  und  $T$ .
5. Mutation der Verteilung  $V'$  aus 3.
6. Falls Wert des  $VI$ -Maßes kleiner geworden: gehe zu 4.; sonst: Stop

*Abb. 3: Ablauf der Alignierung mit realen Daten.*

## 5 Zusammenfassung und Ausblick

In diesem Beitrag wurde eine grundlegende Vorgehensweise zur Erstellung von P2PIR-Testkollektionen vorgestellt, welche auf der Verwendung von latenten Konzepten (anstelle harter Kategorisierungen) basiert. Zur Lösung des Problems der Alignierung von Konzeptverteilungen in realen Systemen mit Testkollektionen wurde ein Maß für den Vergleich von Verteilungen vorgeschlagen.

Neben der Entwicklung und Implementierung eines effizienten Algorithmus zur Optimierung dieses Maßes sind weitere Fragestellungen für die Zukunft interessant:

- Praktische Probleme bei der Erhebung der Daten aus der realen Nutzercommunity: Verhältnis zwischen Anfragen eines Peers und eigenem Dokumentenbestand, Duplikate und populäre Dokumente.
- Modellierung von Dynamik: Churn, Abfrageverhalten über die Zeit hinweg, Downloadverhalten, Verschiebung von Nutzerinteressen über die Zeit, etc.

## 6 Literaturverzeichnis

- [BNJ02] Blei, D.; Ng, A. & Jordan, M. Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems 14*, MIT Press, 2002.
- [BG05] Bhattacharya, I. & Getoor, L. A Latent Dirichlet Model for Unsupervised Entity Resolution. *The 6th SIAM Conference on Data Mining (SIAM SDM-06)*
- [CG02] Crespo, A. & Garcia-Molina, H. Semantic Overlay Networks for P2P Systems. Technical report, Computer Science Department, Stanford University, 2002.
- [CH01] Cohn, D. & Hofmann, T. The Missing Link: A Probabilistic Model of Document Content and Hypertext Connectivity. *Advances in Neural Information Processing Systems*, MIT Press 2001.
- [Coo04] Cooper, B.F. A content model for evaluating peer-to-peer searching techniques. *ACM/IFIP/USENIX 5th International Middleware Conference*, Toronto, 2004.
- [DDL+90] Deerwester, S.C.; Dumais, S.T.; Landauer, T.K.; Furnas, G.W. & Harshman, R.A. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 1990, 41, 391-407.
- [GS04] Griffiths, T.L. & Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004, 101, 5228-5235.
- [Hei05] Heinrich, G. Parameter estimation for text analysis. Web: <http://www.arbylon.net/publications/text-est.pdf>, 2005.
- [MCW04] McCallum, A.; Corrada-Emmanuel, A. & Wang, X. The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. Technical Report, University of Massachusetts, Amherst, 2004.
- [Mei02] Meila, M. Comparing clusterings. Technical Report 418, Department of Statistics, University of Washington, 2002.
- [Mur01] Murphy, K. An introduction to graphical models, Web: [http://www.ai.mit.edu/~murphyk/Papers/intro\\_gm.pdf](http://www.ai.mit.edu/~murphyk/Papers/intro_gm.pdf), 2001
- [NBMW06] Neumann, T., Bender, M., Michel, S. & G. Weikum. A reproducible benchmark for P2P retrieval. *Proc. First Int. Workshop on Performance and Evaluation of Data Management Systems, ExpDB 2006*, 1-8.
- [SCK03] Schlosser, M. T., Condie, T. E. & Kamvar, S. D. Simulating a File-Sharing P2P Network. In *First Workshop on Semantics in P2P and Grid Computing*, 12th WWWConference, Budapest, 2003.
- [SSR+04] Steyvers, M.; Smyth, P.; Rosen-Zvi, M. & Griffiths, T. Probabilistic Author-Topic models for information discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.
- [WB05] Witschel, H.F. & Böhme, T. Evaluating profiling and query expansion methods for p2p information retrieval. In *Proc. of the 2005 ACM Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR)*, 2005.