Heterogeneous Networks and Their Applications: Scientometrics, Name Disambiguation, and Topic Modeling

Ben King, Rahul Jha
Department of EECS
University of Michigan
Ann Arbor, MI
{benking,rahuljha}@umich.edu

Dragomir R. Radev
Department of EECS
School of Information
University of Michigan
Ann Arbor, MI
radev@umich.edu

Abstract

We present heterogeneous networks as a way to unify lexical networks with relational data. We build a unified ACL Anthology network, tying together the citation, author collaboration, and term-cooccurence networks with affiliation and venue relations. This representation proves to be convenient and allows problems such as name disambiguation, topic modeling, and the measurement of scientific impact to be easily solved using only this network and off-the-shelf graph algorithms.

1 Introduction

Graph-based methods have been used to great effect in NLP, on problems such as word sense disambiguation (Mihalcea, 2005), summarization (Erkan and Radev, 2004), and dependency parsing (McDonald et al., 2005). Most previous studies of networks consider networks with only a single type of node, and in some cases using a network with a single type of node can be an oversimplified view if it ignores other types of relationships.

In this paper we will demonstrate *heterogeneous networks*, networks with multiple different types of nodes and edges, along with several applications of them. The applications in this paper are not presented so much as robust attempts to out-perform the current state-of-the-art, but rather attempts at being competitive against top methods with little effort beyond the construction of the heterogeneous network.

Throughout this paper, we will use the data from the ACL Anthology Network (AAN) (Bird et al., 2008; Radev et al., 2013), which contains additional metadata relationships not found in the ACL Anthology, as a typical heterogeneous network. The results in this paper should be generally applicable to other heterogeneous networks.

1.1 Heterogeneous AAN schema

We build a heterogeneous graph $G(V, E)$ from AAN, where V is the set of vertices and E is the set of edges connecting vertices. A vertex can be one of five semantic types: {paper, author, venue, institution, term}. An edge can also be one of five types, each connecting different types of vertices:

- author — [writes] — paper
- paper — [cites] — paper
- paper — [published in] — venue[1]
- author — [affiliated with] — institution[2]
- paper — [contains] — term

All of this data, except for the terms, is available for all papers in the 2009 release of AAN. Terms are extracted from titles by running TextRank (Mihalcea and Tarau, 2004) on NP-chunks from titles and manually filtering out bad terms.

We show the usefulness of this representation in several applications: the measurement of scientific impact (Section 2), name disambiguation (Section 3), and topic modeling (Section 4). The heterogeneous network representation provides a simple framework for combining lexical networks (like the term co-occurence network) with metadata relations from a source like AAN and allows us to begin to develop NLP-aware methods for problems like scientometrics and name disambiguation, which are not usually framed in an NLP perspective.

[1] For a joint meeting of venues A and B publishing a paper x, two edges (x, A) and (x, B) are created.

[2] Author-affiliation edges are weighted according to the number of papers an author has published from an institution.

Transactions of the Association for Computational Linguistics, 2 (2014) 1–14. Action Editor: Lillian Lee.
Submitted 3/2013; Revised 6/2013; Published 2/2014.

2 Scientific Impact Measurement

The study of scientometrics, which attempts to quantify the scientific impact of papers, authors, etc. has received much attention recently, even within the NLP community. In the past few years, there have been many proposed measures of scientific impact based on relationships between entities. Intuitively, a model that can take into account many different types of relationships between entities should be able to measure scientific impact more accurately than simpler measures like citation counts or h-index.

We propose using *Pagerank on the heterogeneous AAN* (Page et al., 1999) to measure scientific impact. Since changes in the network schema can affect the relative rankings between different types of entities, this method is probably not appropriate for comparing entities of two different types against each other. But between nodes of the same type, this measure is an appropriate (and as we will show, accurate) way to compare impacts.

We see this method as a first logical step in the direction of heterogeneous network-based scientometrics. This method could easily be extended to use a directed schema (Kurland and Lee, 2005) or a schema that is aware of the lexical content of citation sentences, such as sentiment-based signed networks (Hassan et al., 2012).

Determining the intrinsic quality of scientific impact measures can be difficult since there is no way to collect gold standard measurements for real-world entities. Previous studies have attempted to show that their measures give high scores to a few known high-impact entities, *e.g.* Nobel prize winners (Hirsch, 2005), or have performed a statistical component analysis to find the most important measures in a group of related statistics (Bollen et al., 2009). Our approach, instead, is to generate *realistic* data from *synthetic* entities whose impacts are known.

We had considered alternative formulations that did not rely on synthetic data, but each of them presented problems. When we attempted manual prominence annotation for AAN data, the inter-judge agreement (measured by Spearman correlation) in our experiments ranged from decent (0.9 in the case of institutions) to poor (0.3 for authors) to nearly random (0.03 for terms), far too low to use in most cases. We also considered evaluating prominence measures by their ability to predict future citations to an entity. Citations are often used as a proxy for impact, but our measurements have found that correlation between past citations and future citations is too high for citation prediction to be a meaningful evaluation[3].

2.1 Creating a synthetic AAN

In network theory, a common technique for testing network algorithms when judgments of real-world data are expensive or impossible to obtain is to test the algorithm on a synthetic network. To create such a synthetic network, the authors define a simple, but realistic generative process by which the real-world networks of interest may arise. The properties of the network are measured to ensure that it replicates certain observable behaviors of the real-world network. They can then test network algorithms to see how well they are able to recover the hidden parameters that generated the synthetic network. (Pastor-Satorras and Vespignani, 2001; Clauset et al., 2009; Karrer and Newman, 2011)

We take a two-step approach to generating this synthetic data, first generating entities with known impacts, and second, linking these entities together according to their latent impacts. Our heuristic is that high impact entities should be linked to other high impact entities and vice-versa. As in the network theory literature, we must show that this data reflects important properties observed in the true AAN.

One such property is that the number of citations per paper follows a power law distribution (Redner, 1998). We observe this behavior in AAN along with several other small-world behaviors, such as a small diameter, a small average shortest path length, and a high clustering coefficient in the coauthorship graph. We strive to replicate these properties in our synthetic data.

[3]Most existing impact measurements require access to at least one year's worth of citation information. The Spearman correlation between the number of citations received after one year and after five years is 0.79 with correlation between successive years as high as 0.99. Practically this means that the measures that best correlate with citations after five years are exactly those that best correlate with citations after one year.

Since scientific impact measures attempt to quantify the true impact of entities, we can use these measures to help understand how the true impact measures are distributed across different entities. In fact, citation counts, being a good estimate of impact, can be used to generate these latent impact variables for each entity. For each type of entity (papers, authors, institutions, venues, and terms), we create a latent impact by sampling from the appropriate citation count distribution. After sampling, all the impacts are normalized to fall in the $[0, 1]$ interval, with the highest-impact entity of each type having a latent impact of 1. Additive smoothing is used to avoid having an impact of 0.

Once we have created the entities, our method for placing edges is most similar to the Erdős-Réyni method for creating random graphs (Erdős and Rényi, 1960), in which edges are distributed uniformly at random between pairs of vertices. Instead of distributing links uniformly, links between entities are sampled proportionally to $I(a)I(b)(1 - (I(a) - I(b))^2)$, where $I(x)$ is the latent impact of entity x.

We tried several other formulae that failed to replicate the properties of the real AAN. The $I(a)I(b)$ part of the formula above reflects a preference for nodes of any type to connect with high-impact entities (*e.g.*, major conferences receive many submissions even though most submissions will be rejected), but the $1 - (I(a) - I(b))^2$ part also reflects the reality that entities of similar prominence are most likely to attach to each other (*e.g.*, well-known authors publish in major conferences, while less well-known authors may publish mostly in lesser-known workshops).

Using this distribution, we randomly sample links between papers and authors; authors and institutions; papers and venues; and papers and terms. The only exception to this was paper-to-paper citation links, for which we did not expect this same behavior to apply, as low-impact papers regularly cite high-impact papers, but not *vice-versa*. To model citations, we selected citing papers uniformly at random and cited papers in proportion to their impacts. (Albert and Barabási, 2002)

Finally, we generated a network equal in size to AAN, that is, with the exact same numbers of papers, authors, etc. and the exact same number of paper-author links, paper-venue links, etc. Table 1 compares the observed properties of the true AAN with the observed properties of this synthetic version of AAN. None of the statistics are exact matches, but when building random graphs, it is not uncommon for measures to differ by many orders of magnitude, so a model that has measures that are on the same order of magnitude as the observed data is generally considered to be a decent model (Newman and Park, 2003).

Relationship	**True value**	**Synth. value**
Paper-citations power law coeff.	1.82	2.12
Diameter	9	8
Avg. shortest path	4.27	4.05
Collaboration network clustering coeff.	0.34	0.26

Table 1: Network properties of the synthetic AAN compared with the true AAN.

2.2 Measuring impact on the synthetic AAN

This random network is, of course, still imperfect in some regards. First of all, it has no time aspect, so it is not possible for impact to change over time, which means we cannot test against some impact measures that have a time component like CiteRank (Maslov and Redner, 2008). Second, there are some constraints present in the real world that are not enforced here. Because the edges are randomly selected, some papers have no venues, while others have multiple venues. There is also nothing to enforce certain consistencies, such as authors publishing many papers from relatively few institutions, or repeatedly collaborating with the same authors.

We had also considered using existing random graph models such as the Barabási-Albert model (Barabási and Albert, 1999), which are known to produce graphs that exhibit power law behavior. These models, however, do not provide a way to respect the latent impacts of the entities, as they add links in proportion only to the number of existing links a node has.

We measure the quality of impact measures by comparing ranked lists: the ordering of the entities

Paper measure	Agreement
Heterogeneous network Pagerank	**0.773**
Citation network Pagerank	0.558
Citation count	0.642
Author measure	**Agreement**
Heterogeneous network Pagerank	**0.461**
Coauthorship network Pagerank	0.244
h-index (Hirsch, 2005)	0.292
Aggregated citation count	0.236
i10-index	0.235
Institution measure	**Agreement**
Heterogeneous network Pagerank	**0.373**
h-index (Mitra, 2006)	0.334
Aggregated citation count	0.327
Venue measure	**Agreement**
Heterogeneous network Pagerank	**0.449**
h-index (Braun et al., 2006)	0.425
Aggregated citation count	0.370
Impact factor	0.092
Venue citation network Pagerank (Bollen et al., 2006)	0.366

Table 2: Agreement of various impact measures with the true latent impact.

by their true (but hidden) impact against their ordering according to the impact measure. The agreement between these lists is measured by Kendall's Tau. Table 2 compares several well-known impact measures with our impact measure, Pagerank centrality on the heterogeneous AAN network. We find that some popular methods, such as h-index (Hirsch, 2005) are too coarse to accurately capture much of the underlying variation. There is a version of Kendall's Tau that accounts for ties, and while this metric slightly helps the coarser measures, Pagerank on the heterogeneous network is still the clear winner.

When comparing different ordering methods, it is natural to wonder which of entities the orderings disagree on. In general, non-heterogeneous measures like h-index or collaboration network Pagerank, which only focus on one type of relationship can suffer when the entity in question has an important relationship of another type. For example, if an author is highly cited, but mostly works alone, his

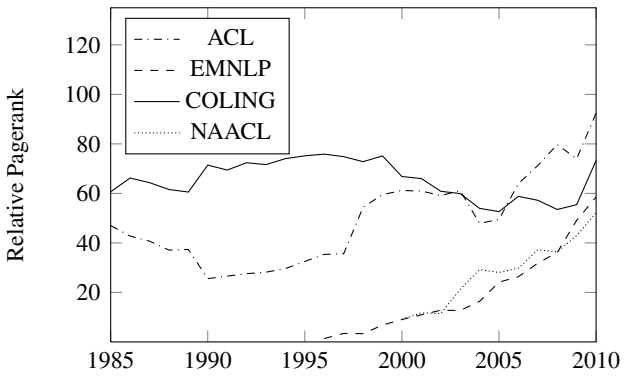

Figure 1: Evolution of conference impacts. The y-axis measures relative Pagerank, the entity's Pagerank relative to the average Pagerank in that year.

contribution would be undervalued in the collaboration network, but would be more accurate in the heterogeneous network.

The majority of the differences between the impact measures, though, tend to be in how they handle entities of low prominence. It seems that, for the most part, there is relatively little disagreement in the orderings of high-impact entities between different impact measures. That is, most highly prominent entities tend to be highly rated by most measures. But when an author or a paper, for example, only has one or two citations, it can be advantageous to look at more types of relationships than just citations. The paper may be written by an otherwise prominent author, or published at a well-known venue, and having many types of relations at its disposal can help a method like heterogeneous network Pagerank better distinguish between two low-prominence entities.

2.3 Top-ranked entities according to heterogeneous network PageRank

Table 3 shows the papers, authors, institutions, venues, and terms that received the highest Pagerank in the heterogeneous AAN. It is obvious that the top-ranked entities in this network are not simply the most highly cited entities.

This ranking also does not have any time bias toward the entities that are currently prominent, as some of the top authors were more prolific in previous decades than at the current time. We also see this effect with COLING, which for many of the early years, is the only venue in the ACL Anthology.

	Top Papers		Top Authors		Top Institutions		Top Venues		Top Terms
–	Building A Large Annotated Corpus Of English: The Penn Treebank	△ 15	Jun'ichi Tsujii	△ 8	Carnegie Mellon University	△ 1	COLING	–	translation
–	The Mathematics Of Statistical Machine Translation: Parameter Estimation	△ 7	Aravind K. Joshi	△ 1	University of Edinburgh	▽ 1	ACL	△ 3	speech
–	Attention, Intentions, And The Structure Of Discourse	△ 18	Ralph Grishman	▽ 2	University of Pennsylvania	△ 2	HLT	▽ 1	parsing
–	A Maximum Entropy Approach To Natural Language Processing	△ 75	Hitoshi Isahara	▽ 2	Massachusetts Institute of Technology	△ 4	EACL	▽ 1	machine translation
–	BLEU: a Method for Automatic Evaluation of Machine Translation	△ 20	Yuji Matsumoto	△ 12	Saarland University	△ 7	LREC	△ 3	generation
–	A Maximum-Entropy-Inspired Parser	△ 7	Kathleen R. McKeown	▽ 2	IBM T.J. Watson Research Center	–	NAACL	△ 3	evaluation
△ 2	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text	△ 13	Eduard Hovy	△ 39	CNRS	▽ 3	EMNLP	△ 6	grammar
▽ 1	A Systematic Comparison of Various Statistical Alignment Models	△ 10	Christopher D. Manning	△ 26	University of Tokyo	▽ 5	Computational Linguistics	△ 16	dialogue
△ 4	Transformation-Based Error-Driven Learning and Natural Language Processing: a Case Study in Part-of-Speech Tagging	△ 93	Yorick Wilks	▽ 4	Stanford University	△ 4	IJCNLP	△ 10	knowledge
△ 1	A Maximum Entropy Model for Part-of-Speech Tagging	▽ 9	Hermann Ney	△ 3	BBN Technologies	△ 1	Workshop on Speech and Natural Language	△ 1	discourse

Table 3: The entities of each type receiving the highest scores from the heterogeneous network Pagerank impact measure along with their respective changes in ranking when compared to a simple citation count measure.

One possible way to address this is to use a narrower time window when creating the graph, such as only including edges from the previous five years. We apply this technique in the following section.

2.4 Entity impact evolution

The heterogeneous graph formalism also provides a natural way to study the *evolution of impact* over time, as in (Hall et al., 2008), but at a much finer granularity. Hall et al. measured the year-by-year prominence of statistical topics, but we can measure year-by-year prominence for any entity in the graph.

To measure the evolution of impacts over the years, we iteratively create year-by-year versions of the heterogeneous AAN. Each of these graphs contains all entities along with all edges occurring in a five year window. Due to space, we cannot comprehensively exhibit this technique and the data it produces, but as a brief example, in Figure 1, we show how the impacts of some major NLP conferences changes over time.

The graph shows that NAACL and EMNLP have been steadily gaining prominence since their introductions, but also shows that ACL has had to make up a lot of ground since 1990 to surpass COLING. We also notice that all the major conferences have grown in impact since 2005, and believe that as the field continues to grow, the major conferences will continue to become more and more important.

3 Name Disambiguation

We frame network name disambiguation in a link prediction setting (Taskar et al., 2003; Liben-Nowell and Kleinberg, 2007). The problems of name disambiguation and link prediction share many characteristics, and we have found that if two ambiguous name nodes are close enough to be selected by a link-prediction method, then they likely correspond to the same real-world author.

We intend to show that the heterogeneous bibliographic network can be used to better disambiguate author names than the author collaboration network. The heterogeneous network for this problem contains papers, authors, terms, venues, and institutions. We compare several well-known network similarity measures from link prediction by transforming the

Network	Distance Measure	Precision	Recall	F1-score	Rand index	Purity	NMI
Heterogeneous	Truncated Commute Time	0.59	0.78	0.63	0.63	0.71	0.43
Heterogeneous	Shortest Path	**0.90**	0.79	0.83	**0.87**	**0.94**	0.76
Heterogeneous	PropFlow	0.89	**0.83**	**0.84**	**0.87**	0.93	**0.77**
Coauthorship	Truncated Commute Time	0.47	0.80	0.54	0.47	0.60	0.18
Coauthorship	Shortest Path	0.54	0.73	0.60	0.61	0.67	0.31
Coauthorship	PropFlow	0.57	0.76	0.64	0.66	0.71	0.43
Coauthorship	GHOST	0.89	0.60	0.69	0.81	**0.94**	0.63

Table 4: Performance of different networks and distance measures on the author name disambiguation task. The performance measures are averaged over the sets of two, three, and four authors. Rand index is from (Rand, 1971) and NMI is an abbreviation for normalized mutual information (Strehl and Ghosh, 2003)

similarities to distances and inducing clusters of authors based on these distances.

We compare three distance measures: shortest path, truncated commute time (Sarkar et al., 2008), and PropFlow (Lichtenwalter et al., 2010). *Shortest path distance* can be a useful metric for author disambiguation because it is small when two ambiguous nodes are neighbors in the graph or share a neighbor. Its downside is that it only considers one path between nodes, the shortest, and cannot take advantage of the fact that there may be many short paths between two nodes.

Truncated commute time is a variant of commute time where all paths longer than some threshold are truncated. The truncation threshold l should be set such that no semantically meaningful path is truncated. We use a value of ten for l in the heterogeneous graph and three in the coauthorship graph[4]. The advantage of truncated commute time over ordinary commute time is simpler calculation, as no paths longer than l need be considered. The downside of this method is that large branching factors tend to lead to less agreement between commute time and truncated commute time.

PropFlow is a quantity that measures the probability that a non-intersecting random walk starting at node a reaches node b in l steps or fewer, where l is again a threshold. As before, l should be a bound on the length of semantically meaningful paths, so we use the same values for l as with truncated commute time. Of course, PropFlow is not a metric, which is required for some clustering methods. We use the following equation to transform PropFlow to a metric: $d(a,b) = \frac{1}{PropFlow(a,b)} - 1$.

With each of the distance measures, we apply the same clustering method: partitioning around medoids, with the number of clusters automatically determined using the gap statistic method (Tibshirani et al., 2001). We create the null distribution needed for the gap statistic method by many iterations of randomly sampling distances from the complete distance matrix between all nodes in the graph. The gap statistic method automatically selects the number of clusters from two, three, or four author clusters.

We compare our methods against GHOST (Fan et al., 2011), a high-performance author disambiguation method based on the coauthorship graph.

3.1 Data

To generate name disambiguation data, we use the *pseudoword method* of (Gale et al., 1992). Specifically, we choose two or more completely random authors and conflate them by giving all instances of both authors the same name. We let each paper written by this pseudoauthor be an instance to be clustered. The clusters produced by any author disambiguation method can then be compared against the papers actually written by each of the two authors. This method, of course, relies on having all of the underlying authors completely disambiguated, which AAN provides.

This method is used to create 100 distambiguation sets with two authors, 100 for three authors, and 100 for four authors.

[4]This is a standard coauthorship graph with the edge weights equal to the number of publications shared between authors. The heterogeneous network does not have author-to-author links, as authors are linked by paper nodes.

3.2 Results

Table 4 shows the performance of author name disambiguation with different networks and distance metrics. F1-score is the measure that is most often used to compare author disambiguation methods. Both PropFlow and shortest path similarity on the heterogeneous network perform quite well according this measure, as well as the other reported measures. While comparable recall can be achieved using only the coauthorship graph, the heterogeneous graph allows for much higher precision.

4 Random walk topic model

Here we present a topic model based entirely on graph random walks. This method is not truly a statistical model as there are no statistical parameters being learned, but rather a topic-discovery and -assignment method, attempting to solve the same problem as statistical topic models such as probabilistic latent semantic analysis (pLSA) (Hofmann, 1999) or latent Dirichlet allocation (LDA) (Blei et al., 2003). In the absence of better terminology, we use the name *random walk topic model*.

While this method does not have the robust mathematical foundation that statistical topic models possess, in its favor it has modularity, simplicity, and interpretability. This language model is *modular* as it completely separates the discovery of topics from the association of topics with entities. It is *simple* because it requires only a clustering algorithm and random walk algorithms, instead of complex inference algorithms. The method also does not require any modification if the topology of the network changes, whereas statistical models may need an entirely different inference procedure if, *e.g.*, author topics are desired in addition to paper topics. Thirdly this method is easily *interpretable* with topics provided by clustering in the word-relatedness graph and topic association based on random walks from entities to topics.

4.1 Topics from word graph clustering

From the set of ACL anthology titles, we create two graphs: (1) a *word relatedness graph* by creating a weighted link between each pair of words corresponding to the PropFlow (Lichtenwalter et al., 2010) measure between them on the full heterogeneous graph and (2) a *word co-occurence graph* by creating a weighted link between each pair of words corresponding to the number of titles in which both words occur.

Both of these graphs are then clustered using Graph Factorization Clustering (GFC). GFC is a soft clustering algorithm for graphs that models graph edges as a mixture of latent node-cluster association variables. (Yu et al., 2006)

Given a word graph G with vertices V and adjacency matrix $[w]_{ij}$, GFC attempts to fit a bipartite graph $K(V, U)$ with adjacency matrix $[b]_{ij}$ onto this data, with the m nodes of U representing the clusters. Whereas in G, similarity between two words i and j can be measured with w_{ij}, we can similarly measure their similarity in K with $w'_{ij} = \sum_{p=1}^{m} \frac{b_{ip} b_{jp}}{\lambda_p}$ where $\lambda_p = \sum_{i=1}^{n} b_{ip}$ is the degree of vertex $p \in U$.

Essentially the bipartite graph attempts to approximate the transition probability between i and j in G with the sum of transition probabilities from i to j through any of the m nodes in U. Yu, et al. (2006) present an algorithm for minimizing the divergence distance $\ell(\mathbf{X}, \mathbf{Y}) = \sum_{ij}(x_{ij} log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij})$ between $[w]_{ij}$ and $[w']_{ij}$.

We run GFC with this distance metric and $m = 100$ clusters on the word graph until convergence (change in log-likelihood $< 0.1\%$). After convergence, the nodes in U become the clusters and the weights b_{ip} (constrained to sum to 1 for each cluster) become the topic-word association scores.

Examples of some topics found by this method are shown in Table 5. From manual inspection of these topics, we found them to be very much like topics created by statistical topic models. We find instances of all the types of topics listed in (Mimno et al., 2011): chained, intruded, random, and unbalanced. For an evaluation of these topics see Section 4.3.1.

4.2 Entity-topic association

To associate entities with topics, we first create the heterogeneous network as in previous sections, adding links between papers and their title words, along with links between words and the topics that were discovered in the previous section. Word-topic links are also weighted according to the weights

Word sense induction	sense disambiguation word induction unsupervised clustering senses based similarity chinese
CRFs + their applications	entity named recognition random conditional fields chinese entities biomedical segmentation
Dependency parsing	parsing dependency projective probabilistic incremental deterministic algorithm data syntactic trees
Tagging	models tagging model latent markov conditional random parsing unsupervised segmentation
Multi-doc summarization	summarization multi document text topic based query extractive focused summaries
Chinese word segmentation	word segmentation chinese based alignment character tagging bakeoff model crf
Lexical semantics	lexical semantic distributional similarity wordnet resources lexicon acquistion semantics representation
Cross-lingual IR	cross lingual retrieval document language linguistic multi person multilingual coreference
Generation for summar.	sentence based compression text summarization ordering approach ranking generation
Spoken language	speech recognition automatic prosodic tagging spontaneous news broadcast understanding conversational
French function words	de la du des le automatique analyse une en pour
Question answering	question answering system answer domain retrieval web based open systems
Unsupervised learning	unsupervised discovery learning induction knowledge graph acquisition concept clustering pattern
SVMs for NLP	support vector machines errors space classification correcting word parsing detecting
MaxEnt models	entropy maximum approach based attachment model models phrase prepositional disambiguation
Dialogue systems	dialogue spoken systems human conversational multi interaction dialogues utterances multimodal
Semantic role-labeling	semantic role labeling parsing syntactic features ill dependency formed framenet
SMT	based translation machine statistical phrase english approach learning reordering model
Coreference resolution	resolution coreference anaphora reference pronoun ellipsis ambiguity resolving approach pronominal
Semi- and weak-supervision	learning supervised semi classification active data clustering approach graph weakly
Information retrieval	based retrieval similarity models semantic space model distance measures document
Discourse	discourse relations structure rhetorical coherence temporal representation text connectives theory
CFG parsing	context free grammars parsing linear probabilistic rewriting grammar systems optimal
Min. risk train. and decod.	minimum efficient training error rate translation risk bayes decoding statistical
Phonology	phoneme conversion letter phonological grapheme rules applying transliteration syllable sound
Sentiment	sentiment opinion reviews classification mining polarity analysis predicting product features
Neural net speech recog.	speech robust recognition real network time neural networks language environments
Finite state methods	state finite transducers automata weighted translation parsing incremental minimal construction
Mechanical Turk	mechanical turk automatic evaluation amazon techniques data articles image scientific

Table 5: Top 10 words for several topics created by the co-occurence random walk topic model. The left column is a manual label.

Topic 59		Topic 82	
translation	0.1953	parsing	0.1715
machine	0.1802	dependency	0.1192
statistical	0.0784	projective	0.0138
Machine Translation	0.0018	K-best Spanning Tree Parsing	0.0025
Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability	0.0016	Pseudo-Projective Dependency Parsing	0.0024
Filtering Antonymous, Trend- Contrasting, and Polarity-Dissimilar Distributional Paraphrases for Improving Statistical Machine Translation	0.0015	Shift-Reduce Dependency DAG Parsing	0.0017
Knight, Kevin	0.0083	Nivre, Joakim	0.0120
Koehn, Philipp	0.0074	Johnson, Mark	0.0085
Ney, Hermann	0.0072	Nederhof, Mark-Jan	0.0064
RWTH Aachen University	0.0212	Vaxjo University	0.0113
Carnegie Mellon University	0.0183	Brown University	0.0107
University of Southern California	0.0177	University of Amsterdam	0.0094
Workshop on Statistical Machine Translation	0.0590	ACL	0.0512
EMNLP	0.0270	EMNLP	0.0259
COLING	0.0173	CoNLL	0.0223

Table 6: Examples of entities associated with selected topics.

determined by GCF. We then simply take random walks from topics to entities and measure the proportion at which the random walk arrives at each entity of interest. These proportions become the entity-topic association scores.

For example, if we wanted to find the authors most associated with topic 12, we would take a number of random walks (say 50,000) starting at topic 12 and terminating as soon as the random walk first reaches an author node. Measuring the proportion at which random walks arrive at each allows us to compute an association score between topic 12 and each author.

A common problem in random walks on large graphs is that the walk can easily get "lost" between two nodes that should be very near by taking a just a few steps in the wrong direction. To keep the random walks from taking these wrong steps, we adjust the topology of the network using directed links to keep the random walks moving in the "right" direction. We design the graph such that if we desire a random walk from nodes of type s to nodes of type t, the random walk will never be able to follow an outgoing link that does not decrease its distance from the nodes of t.

As shown in section 2.3, there are certain nodes at which a random walk (like Pagerank) arrives at more often than others simply because of their positions in the graph. This suggests that there may be stationary random walk distributions over entities, which we would need to adjust for in order to find the most *significant* entities for a topic.

Indeed this is what we do find. As an example, if we sample topics uniformly and take random walks to author nodes, by chance we end up at Jun'ichi Tsujii on 0.3% of random walks, Eduard Hovy on 0.2% of walks, etc. These values are about 1000 times greater than would be expected at random.

To adjust for this effect, when we take a random walk from a topic x to an entity type t, we subtract out this stationary distribution for t, which corresponds to the proportion of random walks that end at any particular entity of type t by chance, and not by virtue of the fact that the walk started at topic x. The resulting distribution yields the entities of t that are most significantly associated with topic x. Table 6 gives examples of the most significant entities for a couple of topics.

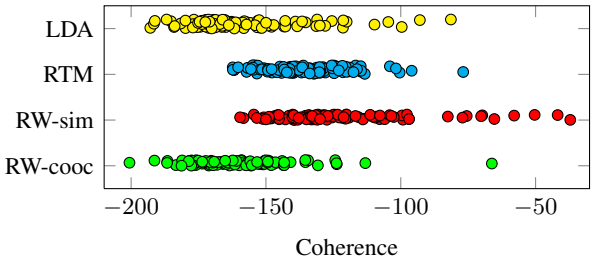

Figure 2: Distribution of topic coherences for the four topic models.

4.3 Topic Model Evaluation

We provide two separate evaluations in this section, one of the topics alone, and one extrinstic evaluation of the entire paper-topic model. The variants of random walk topic models are compared against LDA and the relational topic model (RTM), each with 100 topics (Chang and Blei, 2010). As RTM allows only a single type of relationship between documents, we use citations as the inter-document relationships.

4.3.1 Topic Coherence

The *coherence* of a topic is evaluated using the coherence metric introduced in (Mimno et al., 2011). Given the top M words $V^{(t)} = (v_1^{(t)}, ..., v_M^{(t)})$ for a topic t, the coherence of that topic can be calculated with the following formula:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \left(\frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \right),$$

where $D(v)$ is the number of documents containing v and $D(v, v')$ is the number of documents containing both v and v'.

This measure of coherence is highly correlated with manual annotations of topic quality, with a higher coherence score corresponding to a more coherent, higher quality topic. After calculating the coherence for each of the 100 topics for RTM and the random-walk topic model, the average coherence for RTM topics was -135.2 and the average coherence for word-similarity random walk topics was -122.2, with statistical significance at $p < 0.01$. Figure 2 demonstrates this, showing that the word similarity-based random walk method generates several highly coherent topics. The average coherence for the LDA and the co-occurence random walk model were significantly lower.

4.3.2 Extrinsic Evaluation

One difficulty in evaluating this random-walk topic model intrinsically against a statistical topic model like RTM is that existing evaluation measures assume certain statistical properties of the topic, for example, that the topics are generated according to a Dirichlet prior. Because of this, we choose instead to evaluate this topic model extrinsically with a downstream application. We choose an information retrieval application, returning a ranked list of similar documents, given a reference document.

We evaluate five different methods: citation-RTM, LDA, the two versions of the random-walk topic model, and a simple word vector similarity baseline. Similarity between documents with the topic models are determined by cosine similarity between the topic vectors of the two documents. Word vector similarity determines the similarity between documents by taking the cosine similarity of their word vectors. From these similarity scores, a ranked list is produced.

The document set for this task is the set of all papers appearing at ACL between 2000 and 2011. The top 10 results returned by each method are pooled and manually evaluated with a relevance score between 1 and 10. Thirty such result sets were manually annotated. We then evaluate each method according to its discounted cumulative gain (DCG) (Järvelin and Kekäläinen, 2000).

Performance of these methods is summarized in Table 7. The co-occurence-based random walk topic model performed comparably with the best performer at this task, LDA, and there was no significant difference between the two at $p < 0.05$.

Going forward, an important problem is to reconcile the co-occurence- and word-similarity-based formulations of this topic model, as the two formulations perform very differently in our two evaluations. Heuristically, the co-occurence model seems to create good human-readable topics, while the word-similarity model creates topics that are more mathematically-coherent, but less human-readable.

Method	DCG
Word vector	1.345 ± 0.007
LDA	3.302 ± 0.008
RTM	3.058 ± 0.011
Random-walk (cooc)	3.295 ± 0.006
Random-walk (sim)	2.761 ± 0.007

Table 7: DCG Performance of the various topic models and baselines on the related document finding task. A 95% confidence interval is provided.

5 Related Work

Heterogeneous networks have been studied in a number of different fields, such as biology (Sioson, 2005), transportation networks (Lozano and Storchi, 2002), social networks (Lambiotte and Ausloos, 2006), and bibliographic networks (Sun et al., 2011). These networks are also sometimes known by the name complex networks or multimodal networks, but both these terms have other connotations. We prefer "heterogeneous networks" as used by Sun et al. (2009).

There has also been some study of these networks in general, in community detection (Murata, 2010), clustering (Long et al., 2008; Sun et al., 2012), and data mining (Muthukrishnan et al., 2010), but there has not yet been any comprehensive study. Recently, NLP has seen several uses of heterogeneous networks (though not by that name) for use with label propagation algorithms (Das and Petrov, 2011; Speriosu et al., 2011) and random walks (Toutanova et al., 2004; Kok and Brockett, 2010).

Several authors have proposed the idea of using network centrality measures to rank the impacts of journals, authors, papers, etc. (Bollen et al., 2006; Bergstrom et al., 2008; Chen et al., 2007; Liu et al., 2005), and it has even been proposed that centrality can be applicable in bipartite networks (Zhou et al., 2007). We propose that Pagerank on any general heterogeneous network is appropriate for creating ranked lists for each type of entity. Most previous papers also lack a robust evaluation, demonstrating agreement with previous methods or with some external awards or recognitions. We use a random graph that replicates the properties of the real-world network to show that Pagerank on the heterogeneous network outperforms other methods.

Name disambiguation has been studied in a number of different settings, including graph-based settings. It is common to use the coauthorship graph (Kang et al., 2009; Fan et al., 2011), but authors

have also used lexical similarity graphs (On and Lee, 2007), citation graphs (McRae-Spencer and Shadbolt, 2006), or social networks (Malin, 2005). Almost all graph methods are unsupervised.

There have been some topic models developed specifically for relational data (Wang et al., 2006; Airoldi et al., 2008), but both of these models have limitations in the types of relational data they are able to model. The group topic model described in (Wang et al., 2006) is able to create stronger topics by considering associations between words, events, and entities, but is very coarse in the way it handles the behavior of entities, and does not generalize to multiple different types of entities. The stochastic blockmodel of (Airoldi et al., 2008) can create blocks of similar entities in a graph and is general in the types of graphs it can handle, but produces less meaningful results on graphs that have specific schemas.

6 Conclusion and Future Directions

In this paper, we present a heterogeneous network treatment of the ACL Anthology Network and demonstrate several applications of it. Using only off-the-shelf graph algorithms with a single data representation, the heterogeneous AAN, we are able to very easily build a scientific impact measure that is more accurate than existing measures, an author disambiguation system better than existing graph-based author disambiguation systems, and a random-walk-based topic model that is competitive with statistical topic models.

While there are many other tasks, such as citation-based summarization, that could likely be approached using this framework with the appropriate addition of new types of nodes into the heterogeneous AAN network, there are even some potential synergies between the tasks described in this paper that have yet to be explored. For example, we may consider that the methods of the author disambiguation or topic modeling tasks could be to find the highest-impact papers associated with a term (for survey generation, perhaps) or high-impact authors associated with a workshop's topic (to select good reviewers for it). We believe that heterogeneous graphs are a flexible framework that will allow researchers to find simple, flexible solutions for a variety of problems.

Acknowledgments

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2008. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014.

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

A.L. Barabási and R. Albert. 1999. Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Carl T. Bergstrom, Jevin D. West, and Marc A. Wiseman. 2008. The eigenfactor metrics. *The Journal of Neuroscience*, 28(45):11433–11434.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC08)*, pages 1755–1759.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. 2006. Journal status. *CoRR*, abs/cs/0601030.

Johan Bollen, Herbert Van de Sompel, Aric Hagberg, and Ryan Chute. 2009. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022.

Tibor Braun, Wolfgang Glänzel, and András Schubert. 2006. A hirsch-type index for journals. *Scientometrics*, 69(1):169–173.

Jonathan Chang and David M Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.

Peng Chen, Huafeng Xie, Sergei Maslov, and Sid Redner. 2007. Finding scientific gems with googles pagerank algorithm. *Journal of Informetrics*, 1(1):8–15.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609.

Paul Erdős and Alfréd Rényi. 1960. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479.

Xiaoming Fan, Jianyong Wang, Xu Pu, Lizhu Zhou, and Bing Lv. 2011. On graph-based name disambiguation. *J. Data and Information Quality*, 2(2):10:1–10:23, February.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, volume 54, page 60.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 363–371. ACL.

Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. 2012. Extracting signed social networks from text. *TextGraphs-7*, page 6.

Jorge E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, 102(46):16569.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48. ACM.

In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. 2009. On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97.

Brian Karrer and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 145–153. ACL.

Oren Kurland and Lillian Lee. 2005. Pagerank without hyperlinks: Structural reranking using links induced by language models. In *SIGIR '05*.

Renaud Lambiotte and Marcel Ausloos. 2006. Collaborative tagging as a tripartite network. *Computational Science–ICCS 2006*, pages 1114–1117.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

R.N. Lichtenwalter, J.T. Lussier, and N.V. Chawla. 2010. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM.

Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. 2005. Co-authorship networks in the digital library research community. *Information processing & management*, 41(6):1462–1480.

Bo Long, Zhongfei Zhang, and Tianbing Xu. 2008. Clustering on complex graphs. In *Proc. the 23rd Conf. AAAI 2008*.

Angelica Lozano and Giovanni Storchi. 2002. Shortest viable hyperpath in multimodal networks. *Transportation Research Part B: Methodological*, 36(10):853–874.

Bradley Malin. 2005. Unsupervised name disambiguation via social network similarity. In *Workshop on Link Analysis, Counterterrorism, and Security*, volume 1401, pages 93–102.

Sergei Maslov and Sidney Redner. 2008. Promise and pitfalls of extending google's pagerank algorithm to citation networks. *The Journal of Neuroscience*, 28(44):11103–11105.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530. ACL.

Duncan M. McRae-Spencer and Nigel R. Shadbolt. 2006. Also by the same author: Aktiveauthor, a citation graph approach to name disambiguation. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 53–54. ACM.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4, pages 404–411. Barcelona, Spain.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of HLT-EMNLP*, pages 411–418. ACL.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. ACL.

Panchanan Mitra. 2006. Hirsch-type indices for ranking institutions scientific research output. *Current Science*, 91(11):1439.

Tsuyoshi Murata. 2010. Detecting communities from tripartite networks. In *Proceedings of the 19th international conference on World wide web*, pages 1159–1160. ACM.

Pradeep Muthukrishnan, Dragomir Radev, and Qiaozhu Mei. 2010. Edge weight regularization over multiple graphs for similarity learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 374–383. IEEE.

Mark E.J. Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122.

Byung-Won On and Dongwon Lee. 2007. Scalable name disambiguation using multi-level graph partition. In *Proceedings of the 7th SIAM international conference on data mining*, pages 575–580.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: bringing order to the web.

Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, pages 1–26.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

S. Redner. 1998. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134.

P. Sarkar, A.W. Moore, and A. Prakash. 2008. Fast incremental proximity search in large graphs. In *Proceedings of the 25th international conference on Machine learning*, pages 896–903. ACM.

Allan A. Sioson. 2005. *Multimodal networks in biology*. Ph.D. thesis, Virginia Polytechnic Institute and State University.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63, Edinburgh, Scotland, July. ACL.

Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617.

Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. 2009. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM.

Yizhou Sun, Rick Barber, Manish Gupta, and Jiawei Han. 2011. Co-author relationship prediction in heterogeneous bibliographic networks.

Yizhou Sun, Charu C. Aggarwal, and Jiawei Han. 2012. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *Proceedings of the VLDB Endowment*, 5(5):394–405.

Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. 2003. Link prediction in relational data. In *Neural Information Processing Systems*, volume 15.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Kristina Toutanova, Christopher D Manning, and Andrew Y Ng. 2004. Learning random walk models for inducing word dependency distributions. In *Proceedings of the twenty-first international conference on Machine learning*, page 103. ACM.

Xuerui Wang, Natasha Mohanty, and Andrew McCallum. 2006. Group and topic discovery from relations and their attributes. Technical report, DTIC Document.

Kai Yu, Shipeng Yu, and Volker Tresp. 2006. Soft clustering on graphs. *Advances in Neural Information Processing Systems*, 18:1553.

Ding Zhou, Sergey A. Orshanskiy, Hongyuan Zha, and C. Lee Giles. 2007. Co-ranking authors and documents in a heterogeneous network. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 739–744. IEEE.