Building Content-driven Entity Networks for Scarce Scientific Literature using Content Information

Reinald Kim Amplayo, Min Song
Department of Library and Information Science
Yonsei University
Seoul, Korea
{rktamplayo, min.song}@yonsei.ac.kr

Abstract

This paper proposes several network construction methods for collections of scarce scientific literature data. We define scarcity as lacking in value and in volume. Instead of using the paper's metadata to construct several kinds of scientific networks, we use the full texts of the articles and automatically extract the entities needed to construct the networks. Specifically, we present seven kinds of networks using the proposed construction methods: co-occurrence networks for author, keyword, and biological entities, and citation networks for author, keyword, biological, and topic entities. We show two case studies that applies our proposed methods: CADASIL, a rare yet the most common form of hereditary stroke disorder, and Metformin, the first-line medication to the type 2 diabetes treatment. We apply our proposed method to four different applications for evaluation: finding prolific authors, finding important bio-entities, finding meaningful keywords, and discovering influential topics. The results show that the co-occurrence and citation networks constructed using the proposed method outperforms the traditional-based networks. We also compare our proposed networks to traditional citation networks constructed using enough data and infer that even with the same amount of enough data, our methods perform comparably or better than the traditional methods.

1 Introduction

Large amounts of biomedical data can now be procured in the Internet. One of the more trustworthy source of data is from the scientific community where they do research on specific topics and publish them, which is then made available on the Internet. These vast amounts of data have been used successfully in a lot of areas in biomedicine (Margolis et al., 2014; Marx, 2013; Costa, 2014), from biocuration (Howe et al., 2008) to entity extraction (Rindflesch et al., 2000). In this paper, we focus on the application of the social and knowledge network construction to biomedical data.

One major yet unseen problem is the contradicting problem of *scarce data*. In this paper, we define scarcity in two-folds: lack of value and lack of volume. Lacking in value means that it lacks the necessary information to perform the method. In the case of constructing an author citation network, scarce data may not have the author and citing author information in its metadata. Lacking in volume means that it is not big enough to uncover important knowledge. In the case of constructing an author collaboration network, scarce data may not have enough scale to detect meaningful communities.

Both of these problems in scarcity exist in rare diseases since there are still very few research regarding these diseases. In this paper, we focus on a case study on the research area on Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy, also known as CADASIL. CADASIL (Chabriat et al., 2009) is the most common form of hereditary stroke disorder, yet is listed as one of the many rare diseases[1]. As of the time of writing, searching for research articles regarding CADASIL in Scopus[2] gives approximately only 1100 documents compared to, for example, the approx-

[1] http://globalgenes.org/rarelist

[2] http://www.ncbi.nlm.nih.gov/pmc/

Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), pages 20–29, Osaka, Japan, December 12th 2016.

imately 321 thousand lung cancer-related documents. Using current traditional network construction techniques on the CADASIL data may not work properly. Thus, it is necessary to create an alternative method to handle these kinds of data.

This paper proposes alternative methods to constructing social and knowledge networks to handle scarce data. Instead of using the metadata information, which may not be available, we use the full text of the paper to construct networks. More specifically, instead of using the unavailable author and abstract metadata information of the cited papers, we make use of the sentences where the in-text citations are located (which in this paper we call in-text citation context). Aside from it being able to handle scarce data, it also has some other advantages:

- It can discover **larger communities**, which can be subtopics of the subject at hand, or connections to other subjects which are related to the subject at hand.
- In case of constructing entity co-occurrence networks, it defines a much **clearer polarity** on whether the entities are more significant or less significant because the number of citations received by the entity is also reflected.
- In case of constructing entity citation networks, it makes **use of citation information** extensively. Only the part of the cited paper aimed to cite by the citing paper is included. This is an important distinction because even though the communities become larger and may include other subjects, only the related entities are extracted.

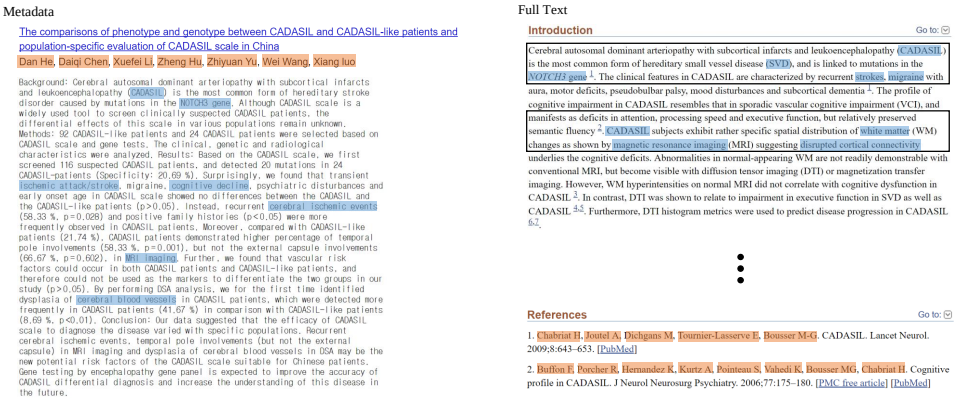
We apply our methods to four different tasks: finding influential authors, finding important biological entities, finding meaningful keywords, and discovering trendy topics. We also present a comparative experimental study on metformin, a drug for type 2 diabetes, which was used as a case study in Ding et al. (2013). We note that these tasks are presented to show comparisons between our proposed methods and the traditional methods in *constructing networks*. The novelty of the paper lies on the construction of entity networks through content-driven approaches.

2 Related work

In this section, we describe related research works on traditional social and knowledge networks and on methods that utilized in-text citation context.

After Newman (2001) introduced scientific collaboration networks, it has been used to analyze the patterns (Newman, 2004) and structure (Hou et al., 2008) of scientific collaboration and coauthorship inside a research community. Hou et al. (2008) also used author collaboration networks to identify prolific authors using the centrality measures. A more recent study by Song et al. (2014) used author collaboration networks to detect communities within the field. Interestingly, citation graphs where authors are the nodes are not used as much as compared to author collaboration networks. Author citation graphs have been used to define a scientist's weighting factor (Życzkowski, 2010) and to determine the citation strength of productive and highly cited authors (Ding, 2011). Entity-based networks, such as entity co-occurrence and entity citation networks, have also been constructed manually (Callon et al., 1991; Ding et al., 2001), using a dictionary (Pettigrew and McKechnie, 2001; Plake et al., 2006; Yan et al., 2013), and using a machine learning technique (Ding et al., 2013; Hahm and Song, 2015) to describe and measure the impact of the entity community or the entity itself and to detect the hidden knowledge between two entities.

Since there were enough data to do proper network analysis, all of the past works above used only meta information such as the paper's authors and abstracts. Only a few research works used the citation information, both the in-text citation context and the reference section of the paper (Yin et al., 2011; Jeong et al., 2014). Yin et al. (2011) used the in-text citation contexts to model linkage information to improve the retrieval of biomedical documents. Similar as ours, Jeong et al. (2014) takes the citation information and constructs a content-based co-citation author network. They constructed an author co-citation network that considers the two authors' contents' similarity when adding edges between the two authors. In this paper, on the other hand, we propose a method to the construction of co-occurrence and



(a) Traditional methods

(b) Full text-based methods

Figure 1: Entity extraction methods

citation networks for scarce data, where if the traditional methods are used to construct the network, network analysis is not possible.

3 Network construction

3.1 Traditional-based methods

This section introduces our approach to network construction compared to the traditional approaches. Figure 1a shows where the traditional methods extract the nodes or the entities used to construct the network. More generally, traditional methods get their entities from the metadata information. Author-based networks are constructed from the authors (highlighted orange in Figure 1a) of the paper and entity-based networks are constructed from the abstract of the paper. For example, the traditional method in constructing author collaboration networks creates edges between authors extracted from the author lists of the papers. Also, the traditional method in constructing entity-entity citation networks creates edges between entities (highlighted blue in Figure 1a) found in the two abstracts of the papers. The problem lies within the volume and the availability of these metadata information in scarce data. Networks constructed with little data cannot uncover important knowledge.

3.2 Full text-based methods

In this paper, we present a network construction method that uses the full texts instead of the available metadata information. Figure 1b shows the location where the full text-based methods extract their entities.

3.2.1 Co-occurrence networks

The full text-based method for constructing a co-occurrence network is similar to the traditional method. The only difference is the location where the entities are extracted. In the case of the author collaboration networks, instead of looking at the authors of the paper, we look at the reference section to extract the authors (highlighted orange in Figure 1b) from the references. One reference citation has one list of authors. A weighted edge is then created between two authors that belongs to one list. Note that if the edge already exists in the graph and another edge between two authors is created, the weight of the existing edge is increased by one. The main advantage of this method is that the constructed collaboration network reflects the number of citations the authors received. This makes it possible to define a much clearer polarity between prolific and non-prolific authors.

In the case of the entity co-occurrence networks, we look at the in-text citations to locate the chunks of text that the citing paper referenced the cited paper. From these chunks of text, we then extract the other entities such as the topic, the keywords, or the biological entities. After extraction, each chunk of text has its list of entities. A weighted edge is then created between two authors that belongs to one list.

3.2.2 Citation networks

The main disadvantage of the traditional method in constructing a citation network is the fact that it needs the citation information between two papers as a metadata information. If the author and abstract information of the cited papers are not available, citation networks cannot be constructed using the traditional method.

Our method does not need the citation information as a metadata information. In the case of the author citation networks, we create a directed weighted edge from each of the authors from the metadata information (orange in Figure 1a) to each of the authors from the reference section of the paper (orange in Figure 1b). In the case of the entity citation networks, we create a directed weighted edge from each of the entities extracted from the abstract (blue in Figure 1a) of the paper to each of the entities extracted from all the in-text citation contexts (blue in Figure 1b).

3.3 Entity extraction

The full text-based methods need to extract the entities to construct the networks. The authors are gathered from the author metadata information and from the reference section of the full text paper. All the other entities are extracted from the abstract and the in-text citation context.

3.3.1 Author extraction

In order to extract the authors from the reference section, it is necessary to take note of the many different styles of citations. Thus, we use an automatic machine learning method to extract the authors from the reference section. We sample a few reference section and manually tag the authors for each reference citation. We then feed them as input for our machine learning model. We use ABNER (Settles, 2005) to create a new linear-chain conditional random field (CRF) based entity extraction where the entity used is only the author. After training, the f1-score of the model is 99.3% with precision of 99.31% and recall of 99.29%. For papers with authors more than 11, we only extract the first 10 and the last author, following the sequence-determines-credit (SDC) and the first-last-author-emphasis (FLAE) approach to author credit contribution (Tscharntke et al., 2007). The author names are then formatted as FN LASTNAME where FN contains the first name initials and LASTNAME is the last name of the author.

3.3.2 Bio-entity extraction

There are multiple types of biological entities from diseases and genes to chemicals and proteins. We use PKDE4J (Song et al., 2015), a biological entity extraction text mining system that synthesized the extraction of 127 types of biological entities obtained from the UMLS semantic groups. Out of the two available methods, we make use of the machine learning-based entity extraction. Since the extracted entities are not preprocessed, we do simple preprocessing techniques by removing the non-alphanumeric symbols, removing multiple whitespaces, and lemmatizing the words using Stanford CoreNLP (Manning et al., 2014).

3.3.3 Keyword extraction

We also extract keywords from the text automatically by using the rapid automatic keyword extraction (RAKE) algorithm (Rose et al., 2010). RAKE is an unsupervised domain- and language-independent method for extracting keywords by making use of a generated stoplist which makes it usable for different domains and languages. In this paper, we use the SMART English stopword list provided by Salton et al. (1975) as the stoplist. After the extraction, we use the same techniques in Section 3.3.2 to preprocess the extracted keywords.

3.3.4 Topic extraction

Topics are extracted using the latent Dirichlet allocation (LDA) topic model (Blei et al., 2003). LDA is a topic modeling technique that infers each document its own topic given the words of each document and two Dirichlet priors α and β. We set the number of topics to 500 and the number of iterations to 5000. We set the Dirichlet priors $\alpha = 1$ and $\beta = 0.01$. The LDA topic model returns a document-topic distribution. From this distribution, we get the two topics with highest probabilities for each abstract and

Table 1: Dataset and network description

		author	bio-entity	keyword	topic
traditional co-occurrence	nodes	4,707	3,493	17,033	-
	edges	18,948	40,386	369,818	-
full text-based co-occurrence	nodes	84,180	21,987	142,319	-
	edges	295,066	89,298	846,269	-
full text-based citation	nodes	87,719	24,522	150,895	498
	edges	952,994	310,590	4,513,469	17,603

one topic for each citation context. We get two topics for the abstracts because the text is long and might be dealing with multiple topics.

4 Experiments

4.1 Dataset

We gather our datasets from PubMed Central (PMC). We use the query term *cadasil* to get the papers' author information and abstract from MEDLINE and PubMed Central IDs directly from PMC. Using the PMCIDs, we obtain the full text, excluding the abstract and including the reference section. From the full text, we extract the in-text citation context with the guidance from the reference section. The citation context contains at most 60 tokens: from the in-text citation, thirty tokens to the left or until the end of the paragraph, and thirty tokens to the right or until the end of the paragraph.

Multiple networks are then created using the methods described in Section 3. Table 1 shows the statistics of the networks created. There are a total of 10 networks: three traditional co-occurrence networks, three full text-based co-occurrence networks, and four full text-based citation networks. Since the paper's citation information is not available, citation networks using the traditional method is not possible. The difference in the size of the traditional and the full text-based networks can be clearly seen.

PageRank (Page et al., 1999) is then calculated for each node for each network. We follow Chen et al. (2007) in their use of $\delta = 0.5$ for PageRank in scientific documents, from the assumption that readers of scientific papers are more likely to jump randomly to a new document compared to web surfers.

We emphasize that the experiments below are shown to provide comparisons between the traditional-based network construction methods and our proposed methods.

4.2 Finding prolific authors

Collaboration networks and citation networks can be used to find prolific authors (Chiang et al., 2013; Garfield, 2006). *Prolific authors* are authors who stand out based on their research output and contributions (Hasselback et al., 2003). We compare the results of the three different author-based networks by sorting the nodes of each network by their PageRank scores in descending order. We then calculate two metrics to measure author prolificity based on the information on Scopus[3] h-index, a widely used author-level metric and the quotient of the total citations over the number of documents the author has (c/d metric). The second metric reflects prolificity more; an author is still influential if it has little documents with many citations. We then compute the average of the metrics of the first 10 authors for evaluation.

Table 2 shows the results of the experiments. It is shown clearly that the traditional co-occurrence network is inferior compared to the two full text-based networks in terms of the average h-index and the average c/d metric. In terms of the average h-index, the full text-based citation network is the more superior network. This means that author citation graph is better in finding prolific authors if we need to also consider productivity. In terms of the average c/d metric, the full text-based co-occurrence network is the more superior network. This means that the full text-based author collaboration network is better in finding prolific authors that emphasizes on the citation impact of the documents and does not consider productivity.

[3]https://www.scopus.com/search/submit/authorFreeLookup.uri

Table 2: Author collaboration and citation networks

(a) traditional co-occurrence

Author	h	c/d
HS MARK...	76	62.45
TR BARR...	29	38.09
AJ LAWR...	39	21.49
RG MORR...	61	46.19
M TRAYL...	10	19.31
C LAMBE...	8	14.38
P BENJA...	2	1.88
RL BROO...	7	9.64
S BEVAN...	22	40.41
B PATEL	8	9.45
average	*26.2*	*26.33*

(b) full text-based co-occurrence

Author	h	c/d
A JOUTEL	41	92.47
E TOURN...	57	59.28
MG BOUS...	87	58.19
H CHABR...	56	46.56
K VAHEDI	36	73.19
V DOMEN...	16	162.88
MM RUCH...	26	39.86
J WEISS...	112	154.07
E MAREC...	25	22.61
EA CABA...	23	13.64
average	47.9	**72.27**

(c) full text-based citation

Author	h	c/d
H CHABR...	56	46.56
A JOUTEL	41	92.47
MG BOUS...	87	58.19
M DICHG...	58	40.64
E TOURN...	57	59.28
K VAHEDI	36	73.19
HS MARK...	76	62.45
N PETERS	24	34.43
F FAZEK...	77	44.16
JM WARD...	71	34.00
average	**58.3**	54.54

Table 3: Extracted biological entities per method

traditional co-occurrence	**notch3**, **vascular dementia**, **stroke**, **hypertension**, alzheimer's disease, **migraine**, disease, vascular lesion, **ischemia**, **notch1**, multiple sclerosis, **amyloid angiopathy**, lacunar infarct, diabetes, single gene disorder, genetic disorder, **atherosclerosis**, allele, vascular, cortex
full text-based co-occurrence	**notch3**, **notch1**, **notch2**, **stroke**, alzheimer's disease, **hypertension**, multiple sclerosis, **vascular dementia**, dll4, **jag1**, **ischemic stroke**, **amyloid angiopathy**, **migraine**, disease, dll1, **fabry disease**, human disease, carasil, lacunar stroke, **atherosclerosis**
full text-based citation	**notch3**, **stroke**, **hypertension**, **caa**, alzheimer's disease, **notch1**, **migraine**, **atherosclerosis**, **vascular dementia**, lacunar infarct, disease, vascular lesion, **cvd**, diabetes, **notch2**, cortex, **ischemia**, dll4, skin, **brain atrophy**

4.3 Finding important biological entities

We can also find important biological entities using co-occurrence and citation networks (Plake et al., 2006; Ding et al., 2013). We compare the results of the three different bio-entity-based networks by sorting the nodes of each network by their PageRank scores in descending order. We then remove all the other bio-entities and leave only the genes and diseases. For evaluation, we compare the first 20 bio-entities to MalaCards (Rappaport et al., 2013), a disease database that records related genes and diseases.

Table 3 shows the results of the experiments. The bold-faced entities are the important bio-entities. The traditional co-occurrence network provides the least number of important bio-entities with only nine entities found. Both the full text-based co-occurrence and the full text-based citation network found 12 important bio-entities. Interestingly, the co-occurrence network found one more gene (jag1) than the citation network.

4.4 Finding meaningful keywords

The keywords automatically extracted by the RAKE algorithm (Rose et al., 2010) may be general keywords and/or are not specific to our CADASIL dataset. The networks can be used to find the most meaningful keywords among the extracted keywords. We compare the results of the three different keyword-based networks by sorting the nodes of each network by their PageRank scores in descending order. For evaluation, we compare the first 20 keywords to MalaCards (Rappaport et al., 2013), which also contains other information regarding CADASIL.

Table 4: Extracted keywords per method

traditional co-occurrence	homonymous visual field defect, **small vessel disease**, **vascular disease**, central retinal artery occlusion, intracranial pressure, optic disc edema, ischemic optic neuropathy, homonymous hemianopia, external carotid artery, ocular ischemic syndrome, **visual loss**, spontaneously, retinal ischemia, optic tract, retinal infarction, **cerebral white matter**, **central nervous system**, clinical presentation, cerebral atrophy, blood flow
full text-based co-occurrence	**cadasil**, **subcortical infarct**, **notch signaling**, risk factor, **vascular dementia**, **cognitive impairment**, **notch receptor**, **cerebral amyloid angiopathy**, **multiple sclerosis**, alagille syndrome, endothelial cell, **stroke**, **notch pathway**, **notch**, **alzheimer disease**, **cognitive decline**, risk, **notch signaling pathway**, disease, **small vessel disease**
full text-based citation	**notch signaling**, **cognitive impairment**, risk factor, endothelial cell, **cognitive decline**, **white matter**, risk, **alzheimer disease**, **notch receptor**, cognitive function, **cadasil**, cell, **stroke**, **subcortical infarct**, **ischemic stroke**, evidence, **notch**, **vascular risk factor**, previously, **notch signaling pathway**

Table 5: Influential topics using PageRank.

Topic 443	Topic 297	Topic 461	Topic 243	Topic 361
risk	cell	study	matter	study
factor	notch	disease	disease	matter
diabetes	stem	research	svd	brain
hypertension	signaling	approach	lesion	impairment
smoking	differentiation	datum	wmh	association
disease	progenitor	treatment	stroke	lesion
stroke	fate	review	lacunar	mri
study	development	result	hyperintensity	volume
age	pathway	patient	vessel	wmh
mellitus	role	disorder	mri	wml

Table 4 shows the results of the experiments. The bold-faced keywords are the meaningful extracted keywords. It is distinctly clear that the traditional-based method did not produce a lot of meaningful keywords, only extracting five. On the other hand, the full text-based co-occurrence network produced 14 meaningful keywords out of the 20 keywords extracted while the full text-based citation network produced 13 meaningful keywords out of the 20 keywords extracted.

4.5 Discovering influential research topics

Using the full text-based topic citation network, we can discover the top influential topics (Lee et al., 2016). Influential topics are topics that are frequently cited by other papers. In this paper, we present the influential topics in CADASIL research. Table 5 contains the top five influential topics based on PageRank. The most influential topic in CADASIL research is the research related to the cardiovascular disease (CVD) risk factors, such as high blood pressure, cholesterol, obesity, smoking, lack of physical ability and diabetes. The next most influential topic in CADASIL research is the research regarding notch signaling and how it regulates the differentiation of neural stem cells. The next three influential topics are case reports, research works on white matter hyperintensities (WMH) in small vessel diseases (SVD), and research works on cognitive impairment.

Table 6: Extracted genes per method

out-degree citation (Ding et al., 2013)	traditional co-occurrence	full text-based co-occurrence	full text-based citation
insulin	oglcnac	**slc2a4**	**slc2a4**
large	p78	gene	gene
impact	p180	**sirt1**	**sirt1**
lep	p202 ptp1b gene	**nfe2l2**	**nfe2l2**
tnf	trem1	met	ae
renin	**slc2a4**	glp1 ras	ppg
insulin receptor	dpp4	ppg	met
set	**pparg**	**tp53**	**pten**
mmp9	sglt2	ae	**tp53**
mmp2	ae	**pten**	sglt2

4.6 Metformin scarce data

In this section, we use Metformin as our data. Although Metformin is a widely research area in Medicine, we only use the first 1000 documents searched from the PubMed Central website to recreate a Metformin scarce data. We compare our methods to the traditional entity-entity citation network in Ding et al. (2013). They constructed the network using all the data found in the PMC website and used the abstracts of all the papers to extract the entities. Their results are then sorted using out-degree centrality. In this comparison, we use only the genes as the entities of our graph.

Table 6 shows the results of the experiments. The results in Ding et al. (2013) produced four related genes. It is clearly better compared to the traditional co-occurrence citation network with only two produced related genes. This is mainly because of the scarce data problem. However, both full text-based co-occurrence network and full text-based citation network produced five related genes, one more than the entity-entity citation network in Ding et al. (2013). This infers that even in the same setting with the same amount of data, the performance of the full text-based networks is comparable to or better than the performance of the traditional-based networks.

5 Conclusion

In this paper, we proposed an alternative method to constructing co-occurrence and citation networks. Instead of extracting entities from the given author and abstract metadata information, we proposed to look at the full text's reference section for the authors and the in-text citation context for the biological entities, keywords and topic. We especially recommend in using this to scarce data, where there is a lack in volume and in value. The advantages are three-fold: larger communities, clearer polarity, and citation emphasis.

We applied this method to research on CADASIL, a rare disorder. We constructed three co-occurrence networks (author, bio-entity, and keyword) and four citation networks (author, bio-entity, keyword, and topic) using the said method. We used it to different kinds of applications: finding prolific authors, finding important biological entities, finding meaningful keywords, and discovering influential topics. Compared to the traditional methods, full text-based methods perform noticeably better in finding significant entities. We also compared our method to the traditional-based entity-entity citation network in (Ding et al., 2013) and found out that even with the same quantity of data, the proposed full text-based network construction method is comparable to or better than the traditional-based network construction methods.

It is to note that looking at the full text instead of just the metadata information provides a more profound and defined analysis from the research articles. For future work, we can apply the methods and create a system to extract different kinds of entities from the full text and automatically construct the different kinds of networks given a set of research articles regarding a specific research area. This would further the research in biomedicine especially on rare diseases, genes, or chemicals.

Acknowledgements

This project is supported fully by Microsoft Research.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Michel Callon, Jean Pierre Courtial, and Francoise Laville. 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1):155–205.

Hugues Chabriat, Anne Joutel, Martin Dichgans, Elizabeth Tournier-Lasserve, and Marie-Germaine Bousser. 2009. Cadasil. *The Lancet Neurology*, 8(7):643–653.

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

Meng-Fen Chiang, Jiun-Jiue Liou, Jen-Liang Wang, Wen-Chih Peng, and Man-Kwan Shan. 2013. Exploring heterogeneous information networks and random walk with restart for academic search. *Knowledge and information systems*, 36(1):59–82.

Fabricio F Costa. 2014. Big data in biomedicine. *Drug discovery today*, 19(4):433–440.

Allan Peter Davis, Cynthia Grondin Murphy, Robin Johnson, Jean M Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Michael C Rosenstein, Thomas C Wiegers, et al. 2013. The comparative toxicogenomics database: update 2013. *Nucleic acids research*, 41(D1):D1104–D1114.

Ying Ding, Gobinda G Chowdhury, and Schubert Foo. 2001. Bibliometric cartography of information retrieval research by using co-word analysis. *Information processing & management*, 37(6):817–842.

Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, and Tamy Chambers. 2013. Entitymetrics: Measuring the impact of entities. *PloS one*, 8(8):e71416.

Ying Ding. 2011. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5(1):187–203.

Eugene Garfield. 2006. The history and meaning of the journal impact factor. *Jama*, 295(1):90–93.

Jung Eun Hahm and Min Song. 2015. Detection of hidden knowledge using a citation-based approach based on swanson's abc model. *Journal of the Korean Society for information Management*, 32(2):87–103.

James R Hasselback, Alan Reinstein, and Edward S Schwan. 2003. Prolific authors of accounting literature. *Advances in Accounting*, 20:95–125.

Haiyan Hou, Hildrun Kretschmer, and Zeyuan Liu. 2008. The structure of scientific collaboration networks in scientometrics. *Scientometrics*, 75(2):189–202.

Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, et al. 2008. Big data: The future of biocuration. *Nature*, 455(7209):47–50.

Yoo Kyung Jeong, Min Song, and Ying Ding. 2014. Content-based author co-citation analysis. *Journal of Informetrics*, 8(1):197–211.

Keeheon Lee, Hyojung Jung, and Min Song. 2016. Subject–method topic network analysis in communication studies. *Scientometrics*, pages 1–27.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Ronald Margolis, Leslie Derr, Michelle Dunn, Michael Huerta, Jennie Larkin, Jerry Sheehan, Mark Guyer, and Eric D Green. 2014. The national institutes of health's big data to knowledge (bd2k) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, 21(6):957–958.

Vivien Marx. 2013. Biology: The big challenges of big data. *Nature*, 498(7453):255–260.

Mark EJ Newman. 2001. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, 64(1):016131.

Mark EJ Newman. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205.

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

Karen E Pettigrew and Lynne EF McKechnie. 2001. The use of theory in information science research. *Journal of the American Society for Information Science and Technology*, 52(1):62–73.

Conrad Plake, Torsten Schiemann, Marcus Pankalla, Jörg Hakenberg, and Ulf Leser. 2006. Alibaba: Pubmed as a graph. *Bioinformatics*, 22(19):2444–2445.

Noa Rappaport, Noam Nativ, Gil Stelzer, Michal Twik, Yaron Guan-Golan, Tsippi Iny Stein, Iris Bahir, Frida Belinky, C Paul Morrey, Marilyn Safran, et al. 2013. Malacards: an integrated compendium for diseases and their annotation. *Database*, 2013:bat018.

Thomas C Rindflesch, Lorraine Tanabe, John N Weinstein, and Lawrence Hunter. 2000. Edgar: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 517. NIH Public Access.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text Mining*, pages 1–20.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Min Song, SuYeon Kim, Guo Zhang, Ying Ding, and Tamy Chambers. 2014. Productivity and influence in bioinformatics: A bibliometric analysis using pubmed central. *Journal of the Association for Information Science and Technology*, 65(2):352–371.

Min Song, Won Chul Kim, Dahee Lee, Go Eun Heo, and Keun Young Kang. 2015. Pkde4j: Entity and relation extraction for public knowledge discovery. *Journal of biomedical informatics*, 57:320–332.

Teja Tscharntke, Michael E Hochberg, Tatyana A Rand, Vincent H Resh, and Jochen Krauss. 2007. Author sequence and credit for contributions in multiauthored publications. *PLoS Biol*, 5(1):e18.

Erjia Yan, Ying Ding, Blaise Cronin, and Loet Leydesdorff. 2013. A bird's-eye view of scientific trading: Dependency relations among fields of science. *Journal of Informetrics*, 7(2):249–264.

Xiaoshi Yin, Jimmy Xiangji Huang, and Zhoujun Li. 2011. Mining and modeling linkage information from citation context for improving biomedical literature retrieval. *Information processing & management*, 47(1):53–67.

Karol Życzkowski. 2010. Citation graph, weighted impact factors and performance indices. *Scientometrics*, 85(1):301–315.