MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-832

CiteGraph: A Citation Network System for MEDLINE Articles and Analysis

Qing Zhang[a], Hong Yu[b,c]

[a] *Department of Electrical Engineering and Computer Science, University of Wisconsin- Milwaukee, Milwaukee, WI, USA*
[b] *University of Massachusetts Medical School, Worcester, MA, USA*
[c] *VA Central Western Massachusetts, Northampton, MA*

Abstract

This paper details the development and implementation of CiteGraph, a system for constructing large-scale citation and co-authorship networks from full-text biomedical articles. CiteGraph represents articles and authors by uniquely identified nodes, and connects those nodes through citation and co-authorship relations. CiteGraph network encompasses over 1.65 million full-text articles and 6.35 million citations by 1.37 million unique authors from the Elsevier full-text articles. Our evaluation shows 98%~ 99% F1-score for mapping a citation to the corresponding article and identifying MEDLINE articles. We further analyzed the characteristics of CiteGraph and found that they are consistent with assumptions made using small-scale bibliometric analysis. We also developed several novel network-based methods for analyzing publication, citation and collaboration patterns. This is the first work to develop a completely automated system for the creation of a large-scale citation network in the biomedical domain, and also to introduce novel findings in researcher publication histories. CiteGraph can be a useful resource to both the biomedical community, and bibliometric research.

Keywords:
Information Science, Information Storage and Retrieval, Databases, Bibliographic, MEDLINE.

Introduction

With the volume of full-text biomedical articles being increasingly available online, citation recognition and analysis can benefit many text-mining applications. Citation plays an important role for both the rhetorical structure [1] and the semantic content of articles [2] and has proven beneficial to many text mining tasks, including information retrieval, extraction, summarization, and question answering [3]. Much in the same way hyperlinks transformed the World Wide Web from a set of static documents into a vibrant and interesting network, a citation network can be a scientific knowledge resource with utility far beyond the sum of its parts.

In this paper, we develop and evaluate CiteGraph, a fully implemented pipeline system that builds a large-scale citation network from a large collection of biomedical full-text articles. Currently, the CiteGraph network encompasses 4.23 million articles taken from the Elsevier collection spanning disciplines ranging from physics to economics. In particular 1.65 million MEDLINE indexed articles are identified from them. The CiteGraph network aligns co-authorship, bibliographical, institutional and citation information into a single cohesive network resource. It also assigns PMIDs, the unique identifier in the Medline collection, to the corresponding articles. The data of the Elsevier and the Medline collections are shown in Figure 1. Since the articles in the Elsevier collection contain full citation information, the overlap between Elsevier and MEDLINE, named as EMedline here represents a subset of 1.65 million biomedical articles from which we built the EMedline Citegraph network.

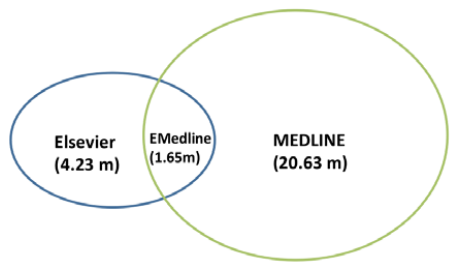

Figure 1 - The overview of the Elsevier, EMedline and MEDLINE datasets.

The illustration of the CiteGraph network is shown in Figure 2. For the purpose of this work, we define an *incite* of A as an article that cites A and an *outcite* of A as an article A cites.

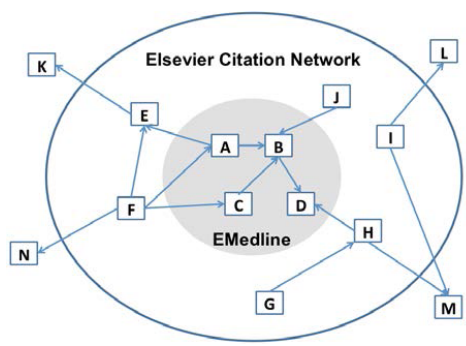

Figure 2 - An illustration of the CiteGraph network. Each square represents a unique article. A directed link represents a citation relation. The CiteGraph network encapsulates citation relations extracted from the Elsevier data set (outer circle). The set of articles contained in the MEDLINE dataset and the links between them is EMedline network (the inner circle).

We evaluated the CiteGraph on linking articles by citation and identifying MEDLINE articles. The system achieved F-1 scores of 0.99 and 0.98 respectively. We also present further analysis of properties of the network and its citation-related characteristics.

Related Work

Citation networks have been built in many domains and accordingly, the literature of citation analysis is rich. Bilke and Peterson (2001) [4] analzyed a citation network built from publications (1975─1989) of high-energy physics, observing that the number of citations follows a power law distribution. Redner (2005) [5] analyzed citations from 110 years of Physics Review publications, a citation network consisting of 353,268 articles. Chen and Rener (2010) [6] used the same network to study the community structure of physics subfields by applying modularity maximization on well-cited articles. In the legal domain, arguments often cite previous case law to support a particular legal opinion. These citations are categorized by legal issues and defined by guidelines. Zhang and Koppaka (2007) [7] developed a legal citation network tool capable of creating legal citations networks for a given issue. Kajikawa and Takeda (2009) [8] analyzed citation network of research papers in the field of organic light-emitting diodes. In that study, they used topological clustering methods to cluster articles based on citation relations and to identify research disciplines. Their methods were able to identify emerging research topics by recursive clustering. Calero-Medina and Noyons (2008) [9] explored "main paths" in a citation network to study research stream and diversity.

Work has been done to address general properties of citation networks. Sen (2005) [12] studied the distribution of citations, and observed that the number of outcites has an exponential distribution, while the number of incites has a power-law distribution. Small (1999)[13] used data from the Information Science Institute to create a map of scientific articles covering 23 disciplines by using co-citation clustering. Discipline subtopics and interdisciplinary pathways were also identified. Greenberg (2009) [14] observed citation bias, amplification and invention in a citation network of the MEDLINE articles. The work shows that above citation distortions may occur with literature that has been accumulated over long periods. Link analysis methods such as HITS and PageRank [15] have been applied to citation networks to rank articles [16,17,18].

Co-authorship networks have also been studied. De Castro and Grossman (1999) [19] defined a co-authorship as a link between the nodes of any two author who have jointly published an article. They further defined distance as the number of nodes in the shortest path between two author nodes. In a mathematical co-authorship network, it was shown that the distance between an author and the famous mathematician Paul Erdos was inversely correlated with that author's impact and reputation [19]. Nascimento et al (2003) [20] extended this work to calculate the "centrality" score for each author, which was defined as the average of distance between the author and any other author in the co-authorship network. Co-citation networks, which link authors that are cited by the same article, were explored to determine semantic similarity of author topics [21] .

Several large-scaled literature repositories have been built. CiteSeer^X[22] is a digital library of scientific literature, and search engine, primarily focusing on computer and information science domain. It autonomously creates a citation index, and provides citation statistics including the number of citations for a given paper, and a list of top cited articles and authors. CiteSeer^X also allows searching of articles that cite a given paper. Google Scholar [23] is another popular literature search engine that collects papers from multiple disciplines. The citations of an article can be displayed, and the number of incites is used as part of its ranking algorithm.

Methods

In this study, we describe a system, called CiteGraph to build the citation and co-authorship network. It consists of a matching algorithm that maps each citation to its article and the corresponding PMID. The implementation of the matching algorithm resulted in two distinct components: *citation mapping* matches a reference to its corresponding article and subsequently establishes the citation links between two articles in the CiteGraph network and *PMID mapping* assigns a unique PMID to each article node. In addition to the two components, *author name disambiguation* disambiguates authors.

Data and Preprocessing

The data that CiteGraph used comprises of 4.23 million full-text articles from the Elsevier data and 20.63 million MEDLINE records. Each record comes with an XML file with fields including *Title, Journal, Author,* and *Year,* as well as the MeSH terms assigned to the record. Every Elsevier article also comes with an XML file. Unlike the MEDLINE, the Elsevier article does not include MeSH terms, but has citations— with fields similar to the ones in MEDLINE—that the article cites. Accordingly, CiteGraph parsed both the MEDLINE articles and the Elsevier XML files by field, assigned the parsed citations to their files, and then indexed them with the open source information retrieval tool Apache Lucene [24].

Mapping Algorithm

We match a citation to its corresponding article as well as an article to its corresponding PMID; this is an important step because significant variations exist in article parsing as well as citation expression. The matching algorithm identifies whether the fields (e.g., *title*, *journal*, and *year*) of two entities (a citation, an article or a MEDLINE record) match. The details of the algorithm are described as follows.

Title: We found title varies in its expression. For example, named entities including chemical and gene names were often represented differently in a citation when comparing it to the title in the original published article. We therefore developed an approximation approach for matching two titles. Specifically, two title fields are considered equal if one of following conditions is met: 1) the set of tokens contained in one title field is a subset of the tokens in the other, or 2) the number of tokens common to both fields is more than 80% of the size of the larger of the two fields. This ad-hoc approximation approach works quite well, as demonstrated in our preliminary evaluation.

Author List: To identify whether the two author lists are identical, we first compared the surnames of the authors; we discarded first names as we found them to introduce significant noise. The author list fields are considered equal only if the set of surnames are equivalent or if the set of surnames in one field is fully contained in the surname set of the second.

Journal: Due to the fact that many journal citations are given using the journal initials, or abbreviated names (e.g., "Mech. Dev", "J. Neurosci." and "JAMA"), journal initials were compared rather than full titles. Stop words, such as "of" and "the" were removed. If the number of common initials in the journal titles was greater than 80% of the tokens in the longer

journal name, they were considered equivalent. The figure of 80% was determined through empirical evaluation.

Citation Mapping Component

Following the citation matching algorithm, we implemented the citation-article mapping component that matches a citation to its article. For each citation, the title, author list and journal name were extracted and used to create a query. Using the Lucene indices, we retrieve top 20 candidate articles. The matching algorithm is then applied to match the citation to each of the 20 retrieved article and outputted the one with the highest matching score.

PMID Mapping Component

As stated earlier, we identify the MEDLINE articles in our network created by using the Elsevier data so that the intersection can be analyzed separately as the EMedline network, limiting the citation network to the biomedical domain. The PMID Mapping component matches an Elseiver article to its corresponding MEDLINE record, and subsequently assigns the corresponding PMID. Similar to the citation mapping component implementation, we used the parsed fields of each article to retrieve top 10 MEDLINE records using the Lucene index and outputted the best match.

Author Name Disambiguation Component

Author names are frequently ambiguous; the same name may refer to different authors. We therefore developed the author name disambiguation component. For implementation, we used the Author-ity database [25], which has disambiguated all authors in the MEDLINE database. With the Author-ity database, we identified a total of 1.37 million unique authors in our EMedline network.

The CiteGraph Networks and Evaluation

The aforementioned three CiteGraph components allow us to create citation networks by linking articles and co-authors. Two networks were created by CiteGraph: the Elsevier network, consisting of all article nodes and citations extracted from the Elsevier dataset; and the EMedline network, a subset of the Elsevier network consisting of only the articles which were assigned PMIDs and the citations between them, as illustrated in Figure 2.

Using both our Elsevier and the Medline datasets, the Elsevier network contains 4.22 million articles, and 106.25 million citations, while the EMedline network has 1.65 million articles, and 6.35 million citations. The average EMedline network node cites to 3.85 EMedline and 32.34 Elsevier nodes. EMedline article nodes are also on average cited by 3.85 EMedline nodes and by 26.66 Elsevier nodes. Table 1 shows the characteristics of the EMedline network.

Table 1 - EMedline Network Statistics

	EMedline Incite	EMedline Outcite	Total Incite	Total Outcite
Average	3.85	3.85	26.66	32.34
Max	4977	829	37065	3730
Min	0	0	0	0
Std	10.94	5.87	78.64	32.03

The ratio of internal citations (links between EMedline nodes) to total citations (links to EMedline nodes from all Elsevier nodes) in the EMedline network is 0.14. This is similar to the results found in existing work(6) which uses the 110 years of Physics Review (PR) article collections, where the internal-citation ratio is as small as 0.2 for well-cited elementary–particle physics publications.

There is no gold standard for citation mapping; therefore the evaluations of citation mapping and PMID mapping are carried out by human judges. Author name disambiguation performance is determined by the quality of Author-ity database so no evaluation is conducted in this study. In each of the two evaluations, seven human evaluators, all of whom have PhD in either computer science or biomedical informatics, and none of whom participated in our study. Each evaluator was provided with entity (article, MEDLINE or citation) pairs, and was tasked with determining whether the two entities refer to the same article. Every evaluator provides judgments on 20 instances of each task. 25% of the instances are double annotated in order to evaluate inter-annotator agreement.

Citation Mapping: For each citation in the network, a list of potential article mappings is created by selecting Elsevier articles that have either their *title* or *author list* field equal to the corresponding field of the citation, as determined by the matching algorithm described in Method section. Each evaluator is presented with a list of 20 citations randomly selected from the set of citations with at last one potential mapped article. For each citation, the list of potential mapped articles is presented, and the evaluator must select which article, if any, corresponds to the citation in question. This establishes a set of 110 user-curated citation mappings. Precision is measured as the number of citations correctly mapped divided by the number of citations presented that receive a mapping. Recall is calculated as the number of citations correctly mapped divided by the total number of correct mappings found by evaluators. Precision and recall for this task were calculated as 1.00 and 0.96 respectively, resulting in an F -1 score of 0.98.

PMID mapping: The evaluation for PMID mapping is preformed using the same tool as the citation mapping. Evaluators are presented with information for 20 articles that had at least one candidate PMID mapping, and the list of potential PMID mappings. The evaluators are then asked to choose which PMID, if any, corresponds to the given article. For this task, both precision and recall were found to be 0.99, for a resulting F-1 score of 0.99.

The high performances of both tasks are due to two reasons. The first one is the well-formatted data source. The fields such as *title*, *journal* in both collections are clearly tagged. Second, the use of index helps to find potentially matches before sophisticated comparison.

Table 2 - Evaluation Results

Task	Precision	Recall	F1	Inter-annotator agreement (Kappa)
Citation mapping	1	0.96	0.98	1
PMID mapping	0.99	0.99	0.99	1

Network Analysis

Two types of analysis are performed on the network. The first one is article network analysis, which reveals the power law distribution of articles over the number of citation received.

The second one is co-authorship network analysis to show the connectivity and temporal dynamics of researchers.

Article Network Analysis

The growth and preferential attachment are two characteristics in real world networks. The probability of a new node to connect with a given node is not uniform. Instead it's more likely to connect to the node that already has a large number of connections. According to the formulation of [27] suppose there are y articles and each of which received citations $x > 0$. Then x and y satisfy $\log y = \alpha - \beta * \log x$, where e^{α} is the number of articles with incites 1, and x^{β} is the decreasing rate of the number of the articles with incites x. Figure 3 shows the plot of the EMedline network, which demonstrates α=0.85, and β=-2.40 for the straight line.

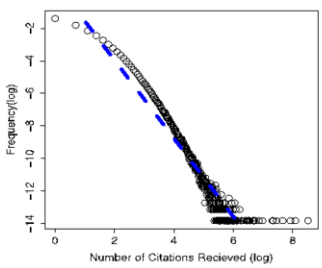

Figure 3 - The EMedline citation frequency distribution. The dashed line is the linear regression of the plot.

Co-authorship Network Analysis

Basic Analysis

The average clustering coefficient is 0.68, which is similar with the coefficient 0.67 and 0.69 of digital library research community co-authorship network [28] and ACM SIGMOD network [20]. It is likely a small world graph according to other study [29], but a random graph is needed to verify it, which is not included in this study.

Table 3 - Statistics of Network Measures.

Measure	Mean	Median	Std	Max	Min
Component size.	3.18	3	1.83	35	2
Clustering Coef.	0.68	0.8095	0.35	1	0
Num. of Co-authors	11	6	14	671	0
Co-authorship year span	1.52	1	1.57	35	1

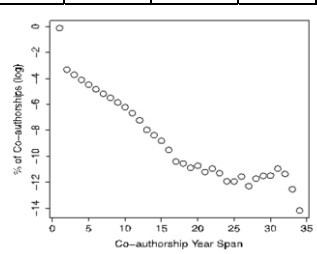

Figure 4 - Co-authorship time span. It is defined as the time difference of the first time collaboration to the last time.

We study the connected component of the co-authorship network, and found 25,510 components. The largest one: component #1 has size 1.27 millions, therefore 92.7% authors in the network are connected in this giant component. Similar phenomenon has also been found in other networks such as the ACM SIGMOD [20], whose largest component consists of 38.2% of the nodes.

Temporal Analysis

The time span of co-authorships is an indicator of the strength of the collaborations. As shown in Figure 4, co-authorship spans from 1 to 35 years, while 83.7% of author pairs just appear once. There are 766,834 and 113,640 co-authorships with five and ten years span respectively. We also analyze the network for its characteristics of author-author and author-article relations. Specifically, we show the relations of author, publications over the career such as average number of co-authors for an author in the i^{th} year of his/her publication history. There are only 0.24% authors who have a publication history longer than 16 years in this collection, so the data is not as representative as the rest. Therefore we use year offset from 0 to 16 for the analysis. The results are shown in Figure 5. Average number of co-authors per publication is generally increasing, which suggests an author tends to have more co-authors as he/her gets senior (a). Number of publications every year is also increasing, and peaks at 14. It shows an author gets more and more productive along his/her career (c). Similarly the percentage of co-authors outside of institution is increasing, suggesting in general authors have broader collaborations when becoming senior (b). Number of citations per publication decreases, and the reason could be that earlier works have more time to receive citations than the newer ones (d). The funding level, policy, and culture changes over time are also possible reasons for the trend in (a) and (b).

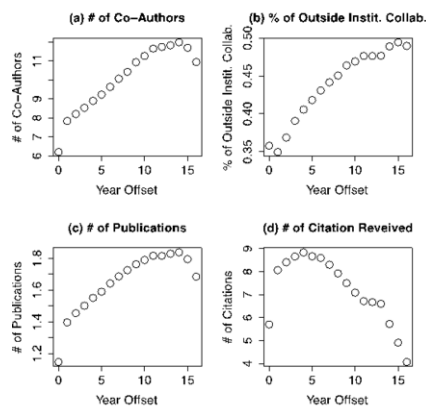

Figure 5 - Author temporal analysis.

Conclusion and Future Work

We described our efforts on building large citation networks using both the Elsevier and the MEDLINE datasets. We developed a citation matching algorithm and implemented three components that match a citation to its corresponding article, identifies the MEDLINE indexed articles, and disambiguates author names. Our system─ named CiteGraph─ incorporates over 4 million Elsevier articles and over 20 million MEDLINE records. The evaluation demonstrated a F-1 score ranging from 98% - 99% for different components. With the CiteGraph networks, we subsequently conducted preliminary graph analysis, including citation frequency over publications and co-authorship network clustering coefficient. Our analysis

demonstrates that the CiteGraph networks reflect the general characteristics of existing networks in other domains. In addition, the temporal analysis shows the researcher tends to have more co-authors per publication and more outside institution collaboration along the career. The limitation of this work is due to the incomplete dataset, as CiteGraph is built on a subset of MEDLINE.

The CiteGraph network provides the medical informatics community a new resource on text mining. In the future we would like to combine the citation network and co-authorship network together to analyze links between them. We speculate that such combination and analyses may lead to important discovery, including document ranking, author ranking, and author collaboration pattern detection.

Acknowledgement

Research reported in this publication was supported in part by 1R01GM095476 to Yu, by a start-up fund from University of Massachusetts Medical School (to Yu) and by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR000161. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] Agarwal S, Choubey L, Yu H. Automatically Classifying the Role of Citations in Biomedical Articles. AMIA Annual Symposium Proceedings. 2010. pp. 11.

[2] Nakov PI, Schwartz A, Hearst M. Citances: Citation sentences for semantic analysis of bioscience text. Proceedings of the SIGIR'04 Workshop on Search and Discovery in Bioinformatics. 2004.

[3] Garfield E. Citation analysis as a tool in journal evaluation. Science. 1972. 178:471-79.

[4] Bilke S, Peterson C. Topological properties of citation and metabolic networks. Physical Review E. 2001; 64(3):036106.

[5] Redner S. Citation statistics from 110 years of Physical Review. Physics Today. 2005;58(6):49–54.

[6] Chen P, Redner S. Community structure of the physical review citation network. Journal of Informatrics. 2010;4(3):278–90.

[7] Zhang P, Koppaka L. Semantics-based legal citation network. Proceedings of the 11th International Conference on Artificial Intelligence and Law. 2007. pp. 123–30.

[8] Kajikawa Y, Takeda Y. Citation network analysis of organic LEDs. Technological Forecasting and Social Change. 2009;76(8):1115–23.

[9] Calero-Medina C, Noyons E. Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. Journal of Informatrics. 2008;2(4):272–9.

[10] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM). 1999;46(5):604–32.

[11] Csárdi G, Strandburg KJ, Zalányi L, Tobochnik J, Érdi P. Modeling innovation by a kinetic description of the patent citation system. Physica A: Statistical Mechanics and its Applications. 2007;374(2):783–93.

[12] Sen P. Directed accelerated growth: application in citation network. Physica A: Statistical Mechanics and its Applications. 2005;346(1):139–46.

[13] Small H. Visualizing science by citation mapping. Journal of the American Society for Information Science. 1999;50(9):799–813.

[14] Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. British Medical Journal. 2009;339(jul2003):b2680.

[15] Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Information Processing & Management. 2008;44(2):800–10.

[16] Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Computer Networks. 2000;33(1-6):387–401.

[17] Ng AY, Zheng AX, Jordan MI. Link analysis, eigenvectors and stability. International Joint Conference on Artificial Intelligence. 2001: pp. 903–10.

[18] Ng AY, Zheng AX, Jordan MI. Stable algorithms for link analysis. Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval. 2001. pp. 258–66.

[19] De Castro R, Grossman JW. Famous trails to Paul Erdős. The Mathematical Intelligencer. 1999;21(3):51–3.

[20] Nascimento MA, Sander J, Pound J. Analysis of SIGMOD's co-authorship graph. ACM SIGMOD Record. 2003;32(3):8–10.

[21] White HD, Griffith BC. Author cocitation: A literature measure of intellectual structure. Journal of the American Society for Information Science. 1981;32(3):163–71.

[22] CiteSeerX. Available from: http://citeseerx.ist.psu.edu/

[23] Google Scholar. http://scholar.google.com/

[24] Apache Lucene. http://lucene.apache.org/core/

[25] Torvik VI, Smalheiser NR. Author name disambiguation in MEDLINE. ACM TKDD. 2009;3(3):11.

[26] Kleinberg J, Kumar R, Raghavan P, Rajagopalan S, Tomkins A. The web as a graph: Measurements, models, and methods. Computing and Combinatorics. 1999;1–17.

[27] Aiello W, Chung F, Lu L. A random graph model for massive graphs. Proceedings of the thirty-second annual ACM Symposium on Theory of Computing. 2000. pp. 171–80.

[28] Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. Information Processing & Management. 2005; 41(6):1462–80.

[29] Kleinberg J. The small-world phenomenon: an algorithm perspective. Proceedings of the thirty-second annual ACM Symposium on Theory of Computing. 2000. pp. 163–70.

Address for correspondence:

Hong Yu, AS6-2071 ASC-QHS, 368 Plantation Street, Worcester, MA 01605. hong.yu@umassmed.edu