Master's Thesis

Semantic approaches to citation recommendation

Tarek Saier

Examiners: Prof. Dr. Georg Lausen
Dr. Michael Färber

Albert-Ludwigs-University Freiburg

Faculty of Engineering

Department of Computer Science

Chair of Databases and Information Systems

April 30th, 2019

Writing Period

15. 10. 2018 – 30. 04. 2019

First Examiner

Prof. Dr. Georg Lausen

Second Examiner & Supervisor

Dr. Michael Färber

Master-Thesis

Semantic approaches to citation recommendation

Tarek Saier

Gutachter: Prof. Dr. Georg Lausen
Dr. Michael Färber

Albert-Ludwigs-Universität Freiburg

Technische Fakultät

Institut für Informatik

Lehrstuhl für Datenbanken und Informationssysteme

30. April 2019

Bearbeitungszeit

15. 10. 2018 – 30. 04. 2019

Erstgutachter

Prof. Dr. Georg Lausen

Zweitgutachter & Betreuer

Dr. Michael Färber

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work. I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date

Signature

Abstract

foo bar

Zusammenfassung

fu bar

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

1.1 Motivation

Citations are a central building block of scholarly discourse. They are the means by which scholars relate their research to existing work—be it in backing up claims, criticising, naming examples or engaging in any other way. Citing in a meaningful way requires an author to be aware of publications relevant to their work. Here, the ever increasing amount of new reseach publications per year poses a serious challenge. Even with academic search engines like Goolge Scholar and CiteSeerX at our disposal, identifying publications that are worthwhile to examine and appropriate to reference remains a time consuming task.

It is therefore not suprising that methods to aid researchers in these tasks have been and still are being actively researched.

While systems for recommending papers have existed since 1998 [1, 2]. The closely related field of citation analysis has an even longer history that spans multiple disciplices including applied linguistics, history and sociology of sience and information science [3, 4].

It is therefore not suprising that efforts to aid researchers in these tasks are no novelty. Systems for recommending papers have existed since 1998 [1, 2]. The closely related field of citation analysis has an even longer history that spans multiple disciplices including applied linguistics, history and sociology of sience and information science [3, 4].

there's peripheral fields like citation context analysis[5], using citations for document exploration[6], citation function[7, 8, 9, 10], citation based plagiarizm detection[11]

also there's different approaches to it (collab.fil. / content based fil.(input=paper / input=sentence / input=sentence+one citation[12] / input=abstract[13]) / graph based)

also there are efforts/ideas/... for using explicit semantic representations for scholarly discours[14, 15, 16]

therefore investigate how they can be applied to citation recommendation

1.2 Problem setting

1.3 Method

generate a data set fit for investigation of citation recommendation employing named entities and claims

devise citation recommendation approach implement and evaluate

1.4 Document structure

foo bar

1.5 Example Section

Copypasta of useful stuff below.

- Put a tilde (nbsp) in front of citations [7].

- **(TODO: Do this!)**
- **(EXTEND: Write more when new results are out!)**
- **(DRAFT: Hacky text!)**
- Chapter 1
- the colors of the Uni
  - UniBlue
  - UniRed
  - UniGrey
- a command for naming matrices $\mathbf{G}$, and naming vectors $\mathbf{a}$. This overwrites the default behavior of having an arrow over vectors, sticking to the naming conventions normal font for scalars, bold-lowercase for vectors, and bold-uppercase for matrices.
- named equations:

$$d(a, b) = d(b, a) \tag{1}$$

symmetry

- Use “these” for citing, not "these"
- If an equation is at the end of a sentence, add a full stop. If it’s not the end, add a comma: $a = b + c \quad (1)$,
- `https://en.wikipedia.org`
- Do not overuse footnotes[1] if possible.

[1] `https://en.wikipedia.org`

2 Related Work

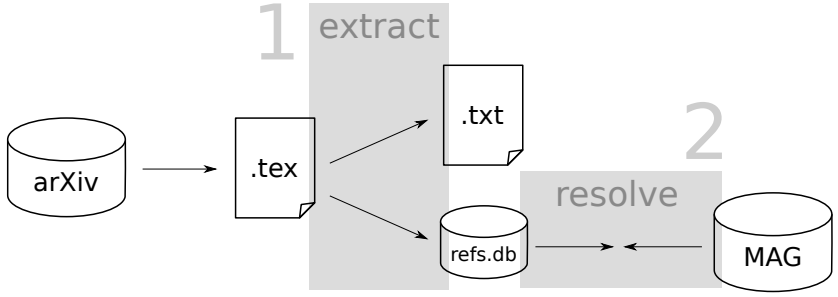
lots. pick wisely.

Leveraging Semantic Annotations to Link Wikipedia and News Archives[17]

Using NEL + dependency trees for music recommendation[18]

3 Background

explain all the things.



(a) subfloat

Figure 1: Caption that appears under the fig—do I want this in bold tho?

Algorithm 1 Stochastic Gradient Descent: Neural Network

Create a mini batch of m samples $\mathbf{x}_0 \dots \mathbf{x}_{m-1}$
foreach sample $\mathbf{x}$ **do**
  $\mathbf{a}^{\mathbf{x},0} \leftarrow \mathbf{x}$ ▷ Set input activation
  foreach Layer $l \in \{1 \dots L-1\}$ **do** ▷ Forward pass
    $\mathbf{z}^{\mathbf{x},l} \leftarrow \mathbf{W}^l \mathbf{a}^{\mathbf{x},l-1} + \mathbf{b}^l$
    $\mathbf{a}^{\mathbf{x},l} \leftarrow \varphi(\mathbf{z}^{\mathbf{x},l})$
  end for
  $\boldsymbol{\delta}^{\mathbf{x},L} \leftarrow \nabla_{\mathbf{a}} C_{\mathbf{x}} \odot \varphi'(\mathbf{z}^{\mathbf{x},L})$ ▷ Compute error
  foreach Layer $l \in L-1, L-2 \dots 2$ **do** ▷ Backpropagate error
    $\boldsymbol{\delta}^{\mathbf{x},l} \leftarrow ((\mathbf{W}^{l+1})^T \boldsymbol{\delta}^{\mathbf{x},l+1}) \odot \varphi'(\mathbf{z}^{\mathbf{x},l})$
  end for
end for
foreach $l \in L, L-1 \dots 2$ **do** ▷ Gradient descent
  $\mathbf{W}^l \leftarrow \mathbf{W}^l - \frac{\eta}{m} \sum_{\mathbf{x}} \boldsymbol{\delta}^{\mathbf{x},l} (\mathbf{a}^{\mathbf{x},l-1})^T$
  $\mathbf{b}^l \leftarrow \mathbf{b}^l - \frac{\eta}{m} \sum_{\mathbf{x}} \boldsymbol{\delta}^{\mathbf{x},l}$
end for

4 Data set

approach approach.

4.1 Existing data sets

and why a new one was necessary

MAG[19] (use/analysis: [20, 21, 22])

use of PMC OAS[23, 24, 25, 26] (PMC OAS problems: [23])

4.2 Data set creation

foo

survey paper on extraction of meta data (author, year, ...) and classification of sentences (method, goal, ...) from publications[27]

evaluation of reference string parsers[28], a dataset for reference string parsing[29]

4.3 Data set evaluation

bar

5 Semantic approaches to citation recommendation

types of citations (naming an entity, backing up a claim, etc.)

how citations are embedded in sentences (integral/non-integral[30, 31, 32, 33, 34])

5.1 Fields of Study as names entities

name name

5.2 Claims

5.2.1 Tools for extracting claims

tools tools

also: Survey on open information extraction[35]

context specific claim detection[36]

if only papers where semantically annotated as proposed in [14]

5.2.2 A model of aboutness closely tied to claim structure

unfeasibility of use of PredPatt output as is

resulting compromise model

predpatt[37, 38]

alternative view: model gives a very selective citation context derived from claim structure (cf. concept of reference scope as sub part of citation context sentence[39, 40]

6 Evaluation

evaluate evaluate

implemetation pain and bad evaluation scores[41]

6.1 Special considerations for citation recommendation

train/test splitting (per cited doc, temporal, ...), re-recommendation, number of contexts describing a recommendation item, ...

a cited doc's role (how it is cited) can develop over time[3, 42]

relevance of time[43]

candidates are only citations within current paper[44]

6.2 Offline evaluation

pre-filtering experiments (knn[26], lsi, lda, fos, ...)

different evaluation settings (all, CSonly, comparison to MAG, ...)

FoS alone, restrictively combined w/ BOW, only directly preceeding, ...

PP model alone, combined, ...

6.3 Online evaluation

Data set	#Papers	Cit. context	Disciplines	Full text	Ref. IDs
arXiv CS	90K	1 sentence	CS	yes	DBLP
CiteSeerX /RefSeer	1M	400 chars	all	no	no
PubMed Central OA[1]	2.3M	extractable	Biomed./Life Sci.	yes	mixed
Scholarly v2[2]	100K	extractable	CS	yes	no
ACL-ARC	11k	extractable	CS/comp. ling.	yes	no
ACL-AAN	18k	extractable	CS/comp. ling.	yes	no

Table 1: Table caption. foo bar...

[1] `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`
[2] `http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html`

7 Conclusion

conclude conclude.

8 Future work

As a first step identify types of citations more systematically.

For different types, different models.

Proper claim model. (that could also include assessing credibility[45])

Argumentative structures. (Argumentation mining[46, 47, 48])

Bibliography

[1] K. D. Bollacker, S. Lawrence, and C. L. Giles, "Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications," in *Proceedings of the Second International Conference on Autonomous Agents*, AGENTS '98, (New York, NY, USA), pp. 116–123, ACM, 1998.

[2] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, pp. 305–338, Nov 2016.

[3] J. Swales, "Citation analysis and discourse analysis," *Applied Linguistics*, vol. 7, no. 1, pp. 39–56, 1986.

[4] H. D. White, "Citation analysis and discourse analysis revisited," *Applied Linguistics*, vol. 25, no. 1, pp. 89–116, 2004.

[5] M. Hernández-Alvarez and J. M. Gomez, "Survey about citation context analysis: Tasks, techniques, and resources," *Natural Language Engineering*, vol. 22, no. 3, p. 327–349, 2016. "r".

[6] M. Berger, K. McDonough, and L. M. Seversky, "Cite2vec: Citation-driven document exploration via word embeddings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 1–1, 01 2016.

[7] M. J. Moravcsik and P. Murugesan, "Some results on the function and quality of citations," *Social Studies of Science*, vol. 5, no. 1, pp. 86–92, 1975.

[8] A. Abu-Jbara, J. Ezra, and D. Radev, "Purpose and polarity of citation: Towards nlp-based bibliometrics," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 596–606, Association for Computational Linguistics, 2013.

[9] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, (Stroudsburg, PA, USA), pp. 103–110, Association for Computational Linguistics, 2006.

[10] S. Teufel, A. Siddharthan, and D. Tidhar, "An annotation scheme for citation function," in *In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 2006.

[11] B. Gipp and J. Beel, "Citation based plagiarism detection: A new approach to identify plagiarized work language independently," in *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, (New York, NY, USA), pp. 273–274, ACM, 2010.

[12] Y. Kobayashi, M. Shimbo, and Y. Matsumoto, "Citation recommendation using distributed representation of discourse facets in scientific articles," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, (New York, NY, USA), pp. 243–251, ACM, 2018. r (ch 1-2).

[13] F. Ayala-Gómez, B. Daróczy, A. Benczúr, M. Mathioudakis, and A. Gionis, "Global citation recommendation using knowledge graphs," *Journal of Intelligent & Fuzzy Systems*, vol. 34, pp. 3089–3100, 05 2018.

[14] S. Buckingham Shum, E. Motta, and J. Domingue, "Scholonto: an ontology-based digital library server for research documents and discourse," *International Journal on Digital Libraries*, vol. 3, pp. 237–248, Oct 2000. r (ch 1-3).

[15] J. Schneider, T. Groza, and A. Passant, "A review of argumentation for the social semantic web," *Semant. web*, vol. 4, pp. 159–218, Apr. 2013.

[16] A. Kitamoto and Y. Nishimura, "Digital criticism platform for supporting evidence-based interpretation of sources," in *IPSJ SIG Computers and the Humanities Symposium 2015*, pp. 211–218, 12 2015. (in Japanese).

[17] A. Mishra and K. Berberich, "Leveraging semantic annotations to link wikipedia and news archives," in *Advances in Information Retrieval* (N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello, eds.), (Cham), pp. 30–42, Springer International Publishing, 2016. r (ch 1-3).

[18] M. Sordo, S. Oramas, and L. Espinosa-Anke, "Extracting relations from unstructured text sources for music recommendation," in *Natural Language Processing and Information Systems* (C. Biemann, S. Handschuh, A. Freitas, F. Meziane, and E. Métais, eds.), (Cham), pp. 369–382, Springer International Publishing, 2015.

[19] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, (New York, NY, USA), pp. 243–246, ACM, 2015. r.

[20] D. Herrmannova and P. Knoth, "An analysis of the microsoft academic graph," *D-Lib Magazine*, vol. 22, no. 9/10, 2016. r.

[21] B. Paszcza, "Comparison of microsoft academic graph with other scholarly citation databases," 11 2016. r (ch 1,""3"").

[22] S. E. Hug, M. Ochsner, and M. P. Brändle, "Citation analysis with microsoft academic," *Scientometrics*, vol. 111, pp. 371–378, Apr 2017. r.

[23] B. Gipp, N. Meuschke, and M. Lipinski, "Citrec : An evaluation framework for citation-based similarity measures based on trec genomics and pubmed central," in *iConference 2015 Proceedings*, iSchools, 2015.

[24] D. Duma, E. Klein, M. Liakata, J. Ravenscroft, and A. Clare, "Rhetorical classification of anchor text for citation recommendation," *D-Lib Magazine*, vol. 22, 2016.

[25] L. Galke, F. Mai, I. Vagliano, and A. Scherp, "Multi-modal adversarial autoencoders for recommendations of citations and subject labels," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, (New York, NY, USA), pp. 197–205, ACM, 2018.

[26] C. Bhagavatula, S. Feldman, R. Power, and W. Ammar, "Content-based citation recommendation," in *NAACL-HLT*, 2018.

[27] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Information extraction from scientific articles: a survey," *Scientometrics*, vol. 117, pp. 1931–1990, Dec 2018.

[28] D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel, "Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, (New York, NY, USA), pp. 99–108, ACM, 2018.

[29] S. Anzaroot and A. McCallum, "A new dataset for fine-grained citation field extraction," in *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

[30] J. Swales, *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

[31] K. Hyland, "Academic attribution: citation and the construction of disciplinary knowledge," *Applied Linguistics*, vol. 20, no. 3, pp. 341–367, 1999.

[32] P. Thompson, *A pedagogically-motivated corpus-based examination of PhD theses: Macrostructure, citation practices and uses of modal verbs.* PhD thesis, University of Reading, 2001.

[33] A. Okamura, "Citation forms in scientific texts: Similarities and differences in l1 and l2 professional writing," *Nordic Journal of English Studies*, vol. 7, no. 3, pp. 61–81, 2008.

[34] W. Lamers, N. J. v. Eck, L. Waltman, and H. Hoos, "Patterns in citation context: the case of the field of scientometrics," in *STI 2018 Conference proceedings*, pp. 1114–1122, Centre for Science and Technology Studies (CWTS), 2018.

[35] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A survey on open information extraction," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3866–3878, Association for Computational Linguistics, 2018.

[36] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim, "Context dependent claim detection," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, (Dublin, Ireland), pp. 1489–1500, Dublin City University and Association for Computational Linguistics, August 2014. r.

[37] A. S. White, D. Reisinger, K. Sakaguchi, T. Vieira, S. Zhang, R. Rudinger, K. Rawlins, and B. Van Durme, "Universal decompositional semantics on universal dependencies," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1713–1723, Association for Computational Linguistics, 2016.

[38] S. Zhang, R. Rudinger, and B. V. Durme, "An evaluation of predpatt and open ie via stage 1 semantic role labeling," in *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*, 2017.

[39] A. Abu-Jbara and D. Radev, "Reference scope identification in citing sentences," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, (Stroudsburg, PA, USA), pp. 80–90, Association for Computational Linguistics, 2012.

[40] R. Jha, A.-A. Jbara, V. Qazvinian, and D. R. Radev, "Nlp-driven citation analysis for scientometrics," *Natural Language Engineering*, vol. 23, no. 1, p. 93–130, 2017.

[41] J. Beel and S. Dinesh, "Real-world recommender systems for academia: The pain and gain in building, operating, and researching them [long version]," *CoRR*, vol. abs/1704.00156, 2017.

[42] J. He and C. Chen, "Temporal representations of citations for understanding the changing roles of scientific publications," in *Front. Res. Metr. Anal.*, 2018.

[43] J. Beel, "It's time to consider "time" when evaluating recommender-system algorithms [proposal]," *CoRR*, vol. abs/1708.08447, 2017.

[44] D. Duma and E. Klein, "Citation resolution: A method for evaluating context-based citation recommendation systems," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 358–363, 2014.

[45] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Credibility assessment of textual claims on the web," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, (New York, NY, USA), pp. 2173–2178, ACM, 2016.

[46] C. Stab and I. Gurevych, "Parsing argumentation structures in persuasive essays," *CoRR*, vol. abs/1604.07370, 2016.

[47] M. Lippi and P. Torroni, "Argumentation mining: State of the art and emerging trends," *ACM Trans. Internet Technol.*, vol. 16, pp. 10:1–10:25, Mar. 2016.

[48] I. Habernal and I. Gurevych, "Argumentation mining in user-generated web discourse," *Comput. Linguist.*, vol. 43, pp. 125–179, Apr. 2017.