Semantometrics: Fulltext-based Measures for Analyzing Research Collaboration

Drahomira Herrmannova[1] and Petr Knoth[2]

[1] *d.herrmannova@open.ac.uk,* [2]*petr.knoth@open.ac.uk*
Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes (United Kingdom)

Introduction

The aim of this article is to demonstrate some of the possible uses of a novel set of metrics called *Semantometrics* in relation to the role of "bridges" in scholarly publication networks. In contrast to the existing metrics such as Bibliometrics, Altmetrics or Webometrics, which are based on measuring the number of interactions in the scholarly network, Semantometrics build on the premise that full-text is needed to understand scholarly publication networks and the value of publications.

Up to date many studies of scientific citation, collaboration and coauthorship networks have focused on the concept of cross-community ties (Shi et al., 2010; Guimerà et al., 2005; Silva et al., 2014). It has been observed that in citation networks, bridging or cross-community citation patterns are characteristic for high impact papers (Shi et al., 2010). This is likely due to the fact that such patterns have the potential of linking knowledge and people from different disciplines. Likewise, in collaboration and coauthorship networks, it has been shown that newcomers in a group of collaborators can increase the impact of the group (Guimerà et al., 2005).

The studies up to date have been focusing on analysing citation and collaboration networks without considering the content of the analysed publications. Our work has focused on analysing scholarly networks using semantic distance of the publications in order to gain insight into the characteristics of collaboration and communication within communities. Our hypothesis states that the information about the semantic distance of the communities will allow us to better understand the importance and the types of the cross-community ties (bridges).

More specifically, in order to gain insight into the type of collaboration between authors we are currently investigating the possibility of utilising semantic distance in a coauthorship network together with the concept of *research endogamy*. In social sciences, endogamy is the practice or tendency of marrying within a social group. This concept can be transferred to research as collaboration with the same authors or collaboration among a group of authors. The concept of research endogamy has been previously used to evaluate conferences (Montolio et al., 2013) as well as journals and patents (Silva et al., 2014).

Furthermore, in (Knoth & Herrmannova, 2014) we have introduced and tested the first Semantometric measure which we call *contribution(p)* and which can be used to estimate research publication contribution. Our results suggested that measuring semantic similarity of publications can be utilised to provide meaningful information about the value of a research publication, which is not captured by traditional bibliometric measures.

Types of research collaboration in a coauthorship network

We are currently investigating the possibility of combining semantic distance and research endogamy in the publication's collaboration network. The rationale behind this approach is based on how research collaboration happens. In case the authors of a publication come from different disciplines, their research is likely to link the two disciplines and to build a bridge between them. This bridge can help to provide vision and ideas otherwise unseen and help to transfer knowledge between the disciplines.

We propose to measure the semantic distance of coauthors of a publication based on semantic distance of all pairs of the coauthors, where the distance of a pair of authors can be expressed similarly as the *contribution(p)* measure (Knoth & Herrmannova, 2014). This situation is depicted in Figure 1, where the sets A and B correspond to the publication records of the two authors.

Table 1. Types of research collaboration based on semantic distance and research endogamy.

	High endogamy	**Low endogamy**
High distance	Established interdisciplinary collaboration	New interdisciplinary collaboration
Low distance	Expert group	New expert collaboration

In order to distinguish between emerging, short-term and established research collaboration, we propose to combine the semantic distance with research endogamy value of the publication as defined in (Silva et al., 2014). We assume that based on the combination of semantic distance and research endogamy the types of research

collaboration can be divided into four groups (Table 1).

We believe this classification is a useful tool in characterising the types of research collaboration that goes beyond the traditional understanding of the concept of bridges as used in scholarly communication networks. While semantic distance allows distinguishing between inter- and intra-disciplinary collaboration, research endogamy allows differentiating between emerging and established research collaborations.

Using semantic distance to measure research contribution in a citation network

A similar Semantometric approach based on the concept of semantic distance can be applied in citation networks. We have used this approach in (Knoth & Herrmannova, 2014) to develop a measure which we call *contribution(p)*. This measure is based on a hypothesis, which states that the added value of publication *p* can be estimated based on the semantic distance from the publications cited by *p* to the publications citing *p*. This situation is depicted in Figure 1.

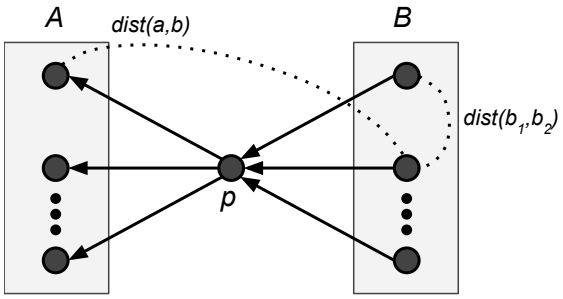

Figure 1. Explanation of contribution(p) calculation.

This hypothesis is based on the process of how research builds on the existing knowledge in order to create new knowledge on which others can build. A publication, which in this way creates a bridge between existing knowledge and something new, which will be developed based on this knowledge, brings a contribution to science. A publication has a high contribution if it connects more distant areas of science. Building on these ideas, we have developed a formula, which can be used for assessing research contribution of a publication. In order to adjust the contribution value to a particular domain and publication type, the metric uses a normalisation factor, which is based on the semantic distance of publications within the set of publications citing *p* and the publications cited by *p*. The measure and our experiments are in detail described in (Knoth & Herrmannova, 2014).

Conclusion

In this paper we proposed to apply the Semantometric idea of using full-texts to recognise types of scholarly collaboration in research coauthorship networks. We have applied semantic distance combined with research endogamy to classify research collaboration into four broad classes. This classification can be useful in research evaluation studies and analytics, e.g. to identify emerging research collaborations or established expert groups. Furthermore, we have presented another Semantometric measure, which we call *contribution(p)* and which is based on the idea of the importance of bridges in a citation network.

While bridges have been the concern of many research studies, their identification has been limited to the structure of the interaction networks. In contrast to these approaches, our approach takes into account both the interaction network (coauthorship, citations) as well as the semantic distance between research papers or communities. This provides additional qualitative information about the collaboration, which hasn't been previously considered.

References

Bornmann, L. & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1).

Guimerà, R., Uzzi, B., Spiro, J. & Nunes Amaral, L. A. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, *308*(April), 697–702.

Knoth, P. & Herrmannova, D. (2014). Towards Semantometrics: A new Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine*, *20*(11).

Montolio, S. L., Dominguez-Sal, D. & Larriba-Pey, J. L. (2013). Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, *42*(2), 11–16.

Priem, J. & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, *15*(7).

Seglen, P. O. (1992). The Skewness of Science. *Journal of the American Society for Information Science*, *43*(9), 628–638.

Shi, X., Leskovec, J. & Mcfarland, D. A. (2010). Citing for High Impact. Proceedings of the 10th Annual Joint Conference on Digital Libraries - *JCDL '10* (p. 49). New York, New York, USA.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira Jr., W. & Laender, A. H. F. (2014). Community-based Endogamy as an Influence Indicator. *Digital Libraries 2014 Proceedings.* London, United Kingdom.