1 Summary on Publications citing SUMO, 2002-2012

Daniel Krajzewicz;
German Aerospace Center, Germany

1.1 Abstract

Having a first release after the success of internet search engines, and being a scientific tool which usage results are usually published, both the development as well as the usage of the open source traffic simulation package "SUMO" can be followed and exploited using commonly available tools to a high degree. In the following, an evaluation of publications which cite or name SUMO is given. Different aspects are shown, such as the development of the publications number over years or the publications' topics.

Keywords: Document Processing, open source Development, Bibliometrics, Traffic Science.

1.2 Introduction

While co-authoring an open source scientific software like the traffic simulation suite "SUMO" [1] [2], one may get interested in its usage, usability, and acceptance in the field it is designed for. Direct interaction with users is one possibility, but it often covers discussions on current issues only and the final results of the users' work are rarely disseminated using this channel. Several circumstances simplify, not to say allow, to collect information necessary for answering questions on a software's usage and acceptance, nowadays. First to name is of course the evolvement of internet search engines, in combination with the fact that SUMO was made available for the public after the search engines' omnipresence. The second is that SUMO mainly targets a scientific audience who presents obtained results in publications of different kind, usually – as a good practice – citing the used tools. Such publications are often indexed by search engines and can be found using appropriate search terms. Additionally, SUMO is a specific tool, reducing the audience and, by doing this, limiting the number of references to a manageable size.

Along the development years, the number of found publications naming SUMO – named "the collection" in the following – grew to a size which is assumed to be high enough to exploit them in a statistical way, not by discussing every paper for itself, but by looking at the tendencies in naming and using SUMO. The collection's basic bibliographic information already allows to examine the development of the number of publications or their types. Additionally, the collection's documents were classified manually by the addressed research topic, the organisations the authors belong to, and the role SUMO plays within the publication, allowing further evaluations. As most documents are available as .pdf-files, it was also partially possible to retrieve the text and to evaluate it.

This document concentrates mainly on listing the generated numbers, distributions, or classes that summarize the collection itself or the work described in the documents the collection consists of. To some degree, it is similar in both, the scientific context of work and in the used

technique, to the evaluation of "IVC[1] Simulation Studies" performed by Stefan Joerer, Christoph Sommer and Falko Dressler [3].

It should also be stated that the work presented here is not intended to bring any further insight into topics such as document processing, traffic science, bibliometrics, or science theory, even these topics are touched and/or their tools are partially used. Trying to gain some knowledge about citation theories, the author stumbled across [4], what he thinks is worth to be read.

This document is structured as follows: at first, the collection is described, including the process of retrieving the documents it consists of, information how the documents were classified and information about the statistical relevance and the data quality. Then, different views on the documents are given, mainly split by the regarded information. This document ends with acknowledgements and a summary/outlook.

Figure 1-1: Frequency of words within the collection's titles, generated with Wordle™ [25]

1.3 The Collection

This section introduces the documents collection, describing how it was generated, first. The manual classification performed on the collection is presented afterwards. The section closes with a discussion about the statistical relevance and the assumed quality of the collection.

1.3.1 Collecting the Documents

A large part of the documents was collected by searching the web using Google [5] and other web search engines. The search was performed between the years 2002 and 2012 repeatedly, albeit spontaneously, not following any fixed frequency or any certain event. Usually, the search was performed using combinations of the terms "SUMO", "traffic simulation", "Simulation of Urban Mobility", and "Krajzewicz"[2], as neither "SUMO", "traffic simulation", nor "Simulation of Urban Mobility" yields in sets containing references to the object of research only. Starting in 2011, additional Google Scholar "alerts" [6] have been used. When established, a Google Scholar alert sends mails to the person who has set it up as soon as Google Scholar receives a new document containing certain words. Two Google

[1] IVC: inter-vehicle communication

[2] The author's name was used not only for some obvious personal reasons, but also because the author is the (co-)author of most publications about SUMO; additionally, the name is occurring relatively seldom, making it a good discriminator.

Scholar alerts were set up, one tracking the search term "'traffic simulation' SUMO", and one tracking "Krajzewicz".

Though SUMO users were asked to send information about their results and/or publications via the mailing list and this request is also included at SUMO's home page, only few citations were reported, which are also enclosed in the list. SUMO was also described by its authors. Publications authored by SUMO's core development team members[3] only are not included in the collection, as they do not resemble SUMO's visibility and usage in the user community. Publications, which the author has contributed to were not added to the collection explicitly. Nonetheless, the database contains five documents co-authored by him.

The links sent by Google Scholar often point to digital library portals, such as Springer Link [7], IEEE Xplore [8], ScienceDirect [9], or ACM Digital Library [10]. Usually, these portals do not only allow to download the publication itself, but also its bibliographic information. In such cases, the bibliographic information was stored as a BibTeX-entry [11] [12] into a BibTeX-database. The entry was afterwards extended by a reference to the downloaded full-text file. In some cases, the Google Scholar link points to the publication itself. In such cases, and for the documents collected before 2011, which were available as .pdf-files, the title of the publication was extracted and used as search term in Google Scholar [13]. From the results, the entry with the same title and the same authors as the original document was chosen as the one representing the document. No ambiguities were observed during this preparation step and the result was a set of BibTeX-databases containing the documents' bibliographic references.

From the database, eleven documents were removed, as it was not possible to obtain their full-texts, disallowing their further processing. Ten of those documents were published in the year 2012, the remaining one in the year 2011.

1.3.2 Manual Classification

Three classification schemes were applied to each of the documents in the collection. This work was done using JabRef [14], an open source literature management application working natively with BibTeX-files. The following classification schemes were applied:

1) SUMO's role in the reported research

The first classification scheme resembles the role of SUMO within the publication. Here, the following classes were used: "mentioned", "used", "extended", and "contributed", where "contributed" is a sub-set of "extended". The classification is assumed to be valid to a high degree, even despite the fact that it was not always possible to determine whether SUMO was extended or not. The classes were defined before classifying the documents.

2) The topic of the publication

The second classification scheme tries to sort the publications by their main research topic. It should indeed be named as a "try", as it is the semantically most weak of the used classification schemes. Imagine a document which reviews mobility models for VANET simulation (such as [15]). During the classification, the document was assigned to the subtopic "mobility models" of the major topic "V2X". On the contrary, a publication

[3] explicitly: Michael Behrisch, Laura Bieker, Jakob Erdmann, Marek Heinrich, Melanie Knocke, Daniel Krajzewicz, Yun-Pang Flötteröd, Peter Wagner

targeting on traffic simulations ([16], e.g.) was assigned to the subtopic "traffic simulations" of the major topic "mobility models". The main motivation behind this scheme was to quickly recognize in which scientific area SUMO is used or mentioned.

In several cases, a document was assigned to more than one topic. An example may be a thesis, where a simulation system is presented, which is used in subsequent steps to evaluate a V2X[4] message routing protocol as well as applications based on the communication. Such a document would be assigned the subtopics "simulation software", "routing protocols", and "applications" of the major topic "V2X". Although a coarse overview on the topics SUMO is used for was known before the classification, the topics were not defined before the document's classification. Instead, new topics or sub-topics were added during the classification process, if needed. After classifying all of the collection's documents, the topics were revisited for balancing them, mainly by joining less occupied topics into classes named "other" at different depths of the topics tree. The chosen topics, sub-topics as well as the terms used to describe them are surely dictated by the author's experience.

3) Organisation(s) the author(s) belong to

Finally, the organisation or organisations the authors of a publication belong to was determined, which was almost always possible using the information contained in the document itself. The classification is hierarchic: on the first level, the country an organisation is located in, was used. On the second level, the organisation – mostly an university – was given. On the third level, if available, the department and/or institute was used. For five publications, it was not possible to completely assign the authors to institutes. These publications were assigned to the group "unknown". Three reports, all from projects co-founded by the European Commission are not indexed, mainly because resolving the partner organisations' abbreviations given in the document was assumed to be too time consuming.

1.3.3 Statistical Relevance

One should pose the questions whether the collection contains all publications citing SUMO or represents the set of all publications citing SUMO properly. The first question should be negated, starting with the fact that the core developer's publications are not included. But, e.g., Master theses are not always made publicly available and so cannot be found using web search engines. This is assumed to count for other types of publications as well.

The second question can be hardly answered without knowing all publications. Nonetheless, based on following SUMO's usage via both, users' publications, as well as the interaction on the mailing list, it is assumed that it is a fair overview of how SUMO is used. In addition, the documents collecting steps were performed using unbiased search terms, concentrating on finding information about SUMO only.

Of a rather academic nature is the question whether all groups that use SUMO, cite it. This is probably not the case, starting with the fact that SUMO is often used as a part of multi-simulator network architectures, such as "Veins" [17], "MOVE" [18], "VSimRTI" [19], or "iTETRIS" [20].

[4] The term "V2X" is a simplifying abbreviation of "vehicle-to-vehicle and vehicle-to-infrastructure"

1.3.4 Data Quality

The BibTeX-descriptions obtained from library portals are of good quality. The bibliographic information from Google Scholar was often erroneous. Both, Master and Doctoral theses were very often not assigned to the proper BibTeX-type. This was corrected manually. Often, the publication year was not given or one could find the name of a month in the according field, instead. This was corrected manually, but the information could not be found for six documents.

Another predictable issue were the author names. Dots were missing at the first name abbreviations, non-Latin characters as German Umlaute (äöü) were encoded in different ways, first names were only sometimes abbreviated, and middle names not always given. This was corrected manually by inspection of similar names and/or from knowing the listed authors. It is assumed to be not reliable to 100%.

The collection includes 362 documents. Table 1-1 summarizes the known quality issues.

Table 1-1: Summary on the collection's quality issues.

Property	Given	Estimated Correctness
document number	362	
BibTeX: type	362	Is mandatory in BibTeX, but is probably not correct for many of the documents as discussed.
BibTeX: Publication Year	356	Assumed to be correct
BibTeX: Journal / BibTeX: Booktitle		Assumed to be incorrect and complicated and time consuming to be corrected; neglected
BibTeX: Authors	362	Assumed to be correct to a large degree
Classification: Role	362	Assumed to be correct to a large degree
Classification: Topic	362	Assumed to be correct
Classification: Institutions	359(-5)	Assumed to be correct to a large degree, despite the limitation described in section 1.3.2.

1.4 Evaluation

In the following, the collection's properties are given. The evaluation was done using the Python [21] programming language and the matplotlib [22] module for visualisation.

1.4.1 Titles

A very coarse, statistical view at the titles, shown in Figure 1-1, already implies what will be discussed in section 3.5 about research topics of the collection: the dominance of research on vehicular communications. Figure 1-1 was generated by collecting the words from titles, for which the stem was determined using Porter stemming [23] [24]. As stems obtained from the Porter algorithm are usually pruned so much that no real English word is obtained, a mapping from the stem to the shortest complete word that was found within the titles was used to get existing English words. The resulting list of occurrences per word was then visualised using Wordle™ [25]. For stemming, the Python Porter stemming implementation from Vivake Gupta and Danny Yoo was used [26].

1.4.2 Development over Time

Figure 1-2 shows the development of the annual publications number over time, distinguishing the publications' types. It shows a fair increase over the years, albeit with some dents.

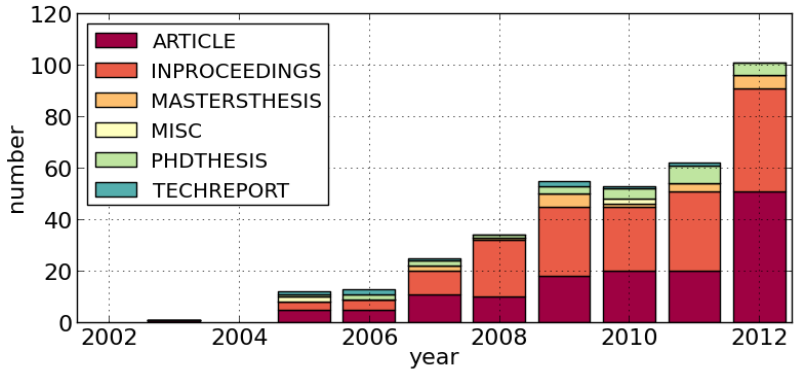

Figure 1-2: The development of publications over the years, divided by publication type

1.4.3 Authors and Authorship

The publications are authored by 863 persons in sum; the development is shown in Figure 1-3. "all" denotes the set of all authors that have co-authored a document in a given year, "new" is the subset of "all" containing authors who have not (co-)authored an earlier publication from the collection and "single" is a subset of "new", which contains authors who have contributed to one publication within the collection only.

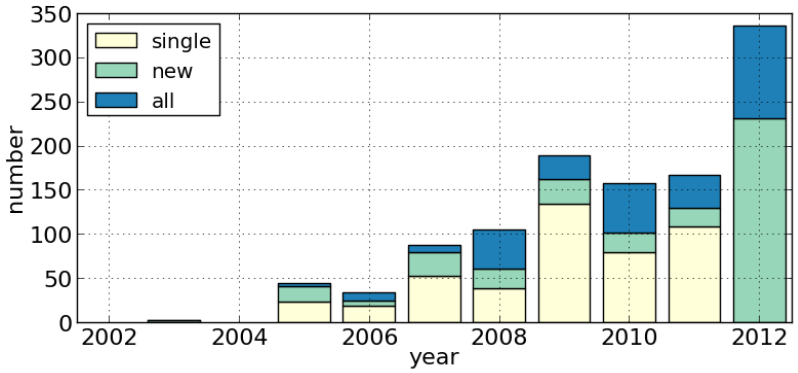

Figure 1-3: The development of authors over the years, classified by the continuity of their publications from the collection

Albeit the number of "all" authors is growing, the majority of the authors has participated in one publication only. This could be interpreted as a minor acceptance, but is rather supposed to have its reasons in a high number of co-authors. Often, a document describes a larger project and only few of the documents' authors have used SUMO by themselves.

1.4.4 SUMO's Role

The manual classification of documents by the role of SUMO within the reported research is maybe the best indicator for SUMO's acceptance in the scientific community. As shown in Figure 1-4, not only the number of documents which report about using SUMO is increasing over time, but also their percentage, in comparison to documents which mention SUMO only.

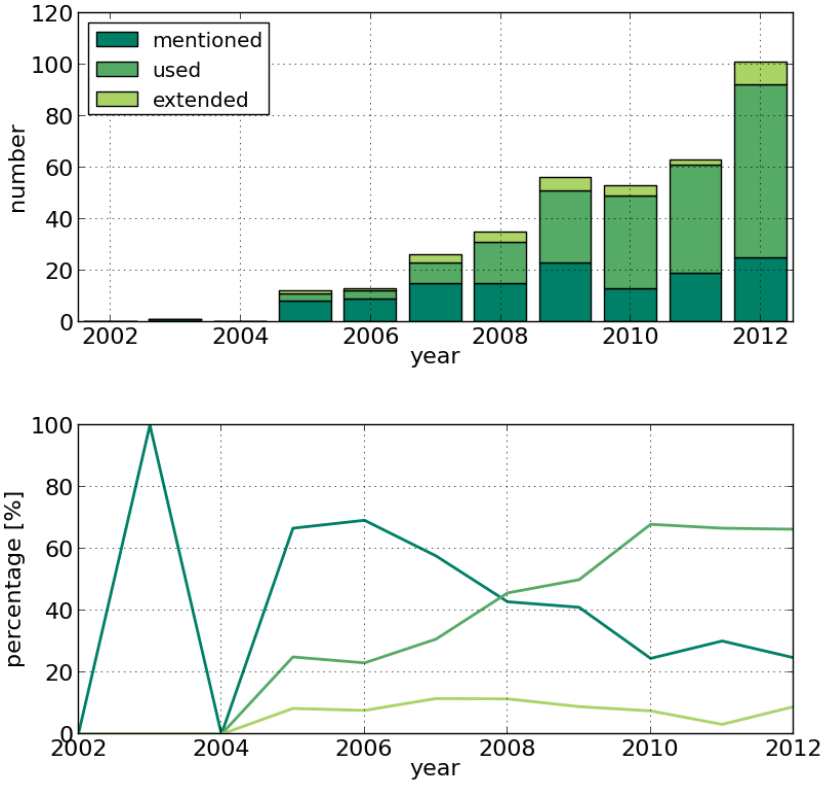

Figure 1-4: Role of SUMO within the evaluated publications. Top: absolute numbers, bottom: percentage, both along the years

Also remarkable is the almost constant percentage of documents where extensions to SUMO are reported. It should be noted, that after introducing the TraCI application control interface in 2008, which allows an on-line interaction with the simulation, the need to extend SUMO was reduced as the logic to evaluate can be embedded into an external application since that time. Although one could assume a reduction of extensions to SUMO since that time, this is not visible within the collection.

1.4.5 Research Topics

As described in section 1.3.2, a document may be assigned to more than one research topics. For the discussion, it may be interesting to know which and how many topic combinations exist within the collection. Figure 1-5 shows the co-occurrence of topics. Topics which do occur in combination with other topics are not shown, herein.

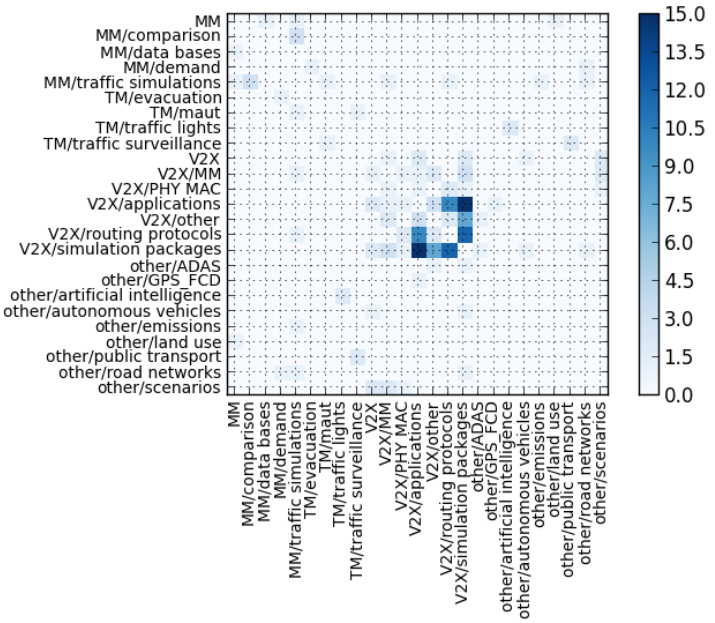

Figure 1-5: Co-occurrences of topics within the collection. Topics with no co-occurrence are not shown. The abbreviations are: MM: mobility models, TM: traffic management

The development of addressed topics within the collection over time is shown in Figure 1-6. Clearly evident is the domination of documents which describe work on "V2X". In sum, about 70 % of the collection were classified into this topic, about 12 % present work on "mobility models" and each of both topics "traffic management" and "other" is discussed in about 9 % of the publications.

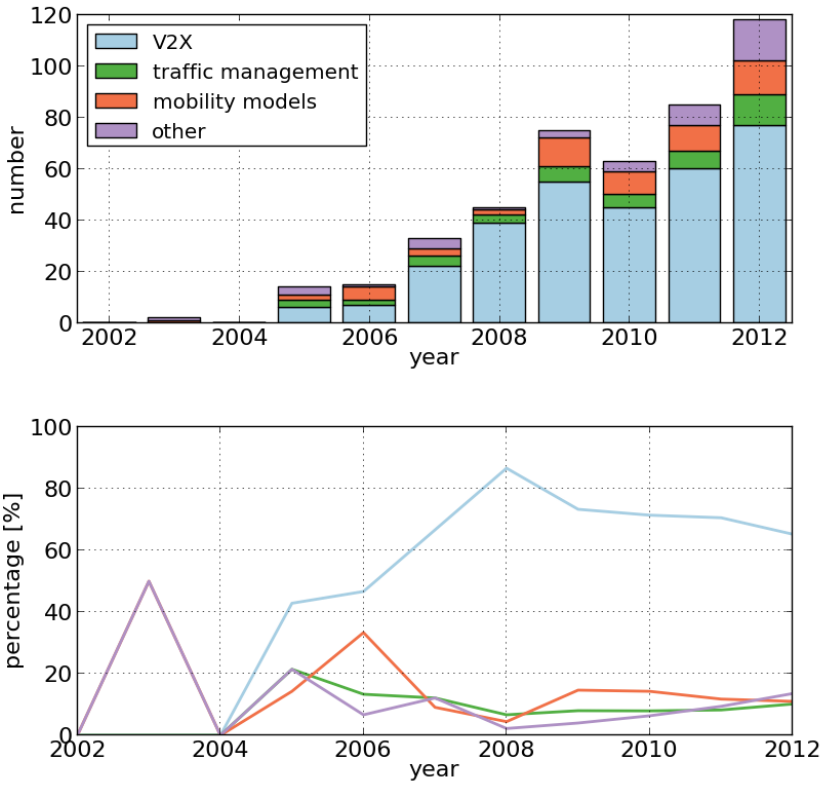

Figure 1-6: Development of the publications' major topics. Top: absolute numbers, bottom: percentage, both along the years

The dominance of V2X research within the SUMO user community was already observed and reported, see [1]. The evaluation of the topics proves it. It may be also noted, that, vice-versa, SUMO is the most often used traffic simulation tool in V2X research, as reported in [3]. The reasons can only be guessed. The first documents targeting this topic occur in the year 2005. One of those, "MOVE: A MObility model generator for VEhicular network" by Feliz Kristianto Karnadi, Zhi Hai Mo, and Kun-Chan Lan [28], may be the reason for SUMO's prominence in this research field. [28] presents "MOVE", a complete system for generating vehicular traces that can be used in the ns2 simulator, which was the state-of-the-art for communication simulation at that time. Screenshots of different scenarios show the system's variability in use.

Whether [28] was the seed to SUMO's popularity within the research on V2X or not, may be provable by evaluating the citations of following papers, trying to determine how often [28] is cited. This was not done so far.

As "V2X" covers about 70 % of the publications, it may be investigated using the same approach. Figure 1-7 shows the development of sub-topics of "V2X" along the years. The overall frequencies of "V2X" sub-topics are as following: "simulation packages": 30.9 %, "routing protocols": 24.7 %, "applications": 22.7 %, "PHY/MAC" and "mobility models": 5.2 %, and "other" cover 11.3 %.

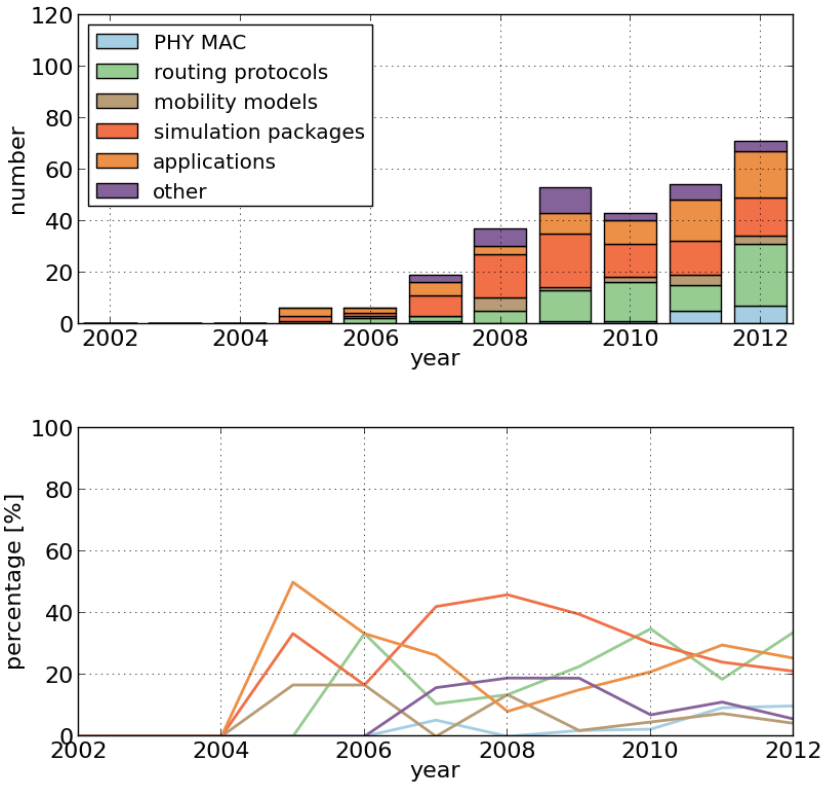

Figure 1-7: Development of the publications' sub-topic within the "V2X" topic. Top: absolute numbers, bottom: percentage, both along the years

It is interesting to note that "simulation packages" – presentation of tools for research – is dominant to such a high degree. On the other hand, one should take into regard that the presentation of a "simulation package" occurs often in combination with some further evaluation of V2X functionality, may it be a "routing protocol", or an "application", see also Figure 1-5. Nonetheless, the author finds the number of reports on "simulation packages" quite high. Whether such "simulation packages" find their way into a broader use, or are only used once, and whether the necessity to develop new ones exist, given the high number of existing ones, is matter of a different kind of research.

A promising fact is the increase in using SUMO for evaluating V2X-based applications, as in most cases, such research requires a joint operation of a traffic simulation and a communication simulation, usually employing a middleware instance. The increase of publications on this topic shows that available middleware solutions are accepted and can be used for scientific work. Within the work on "routing protocols", SUMO is usually used to generate vehicular "traces" only, which are then exported into a format readable by the used communication simulator.

The document sets of the remaining major topics are too small for a meaningful insight into the development over time. Figure 1-8 shows the distribution of the sub-topics within the major topics "mobility models" and "traffic management".

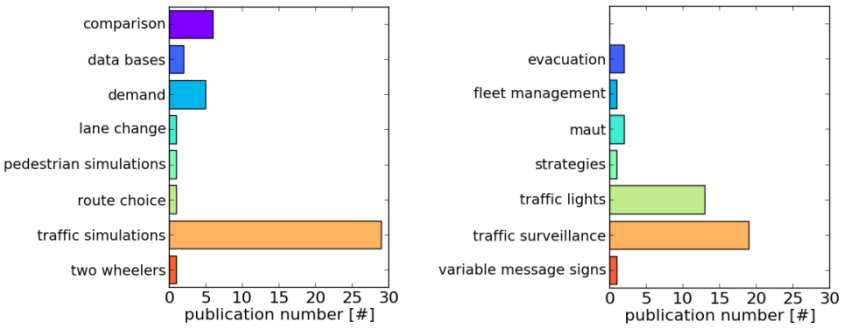

Figure 1-8: Subtopics and their relative frequency for the topics "mobility models" (left) and "traffic management" (right)

1.4.6 Countries and Organisations

The categorization into the institutions the authors belong to is shown on the top-most, national level only, herein. Figure 1-9 shows a matrix of international co-authoring of the collection's papers, whereas Figure 1-10 shows the numbers of publications per country.

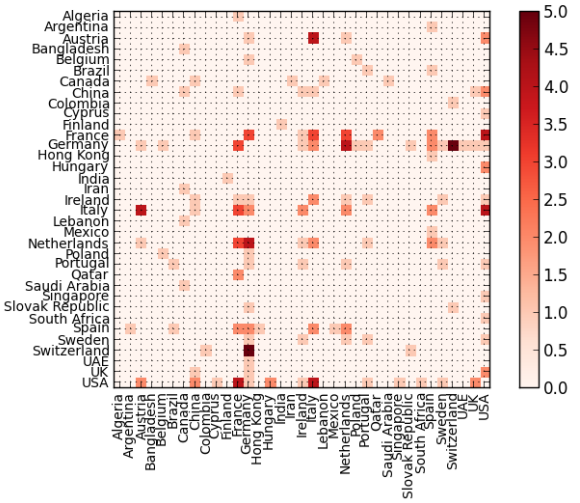

Figure 1-9: Co-occurrence of countries within the collection

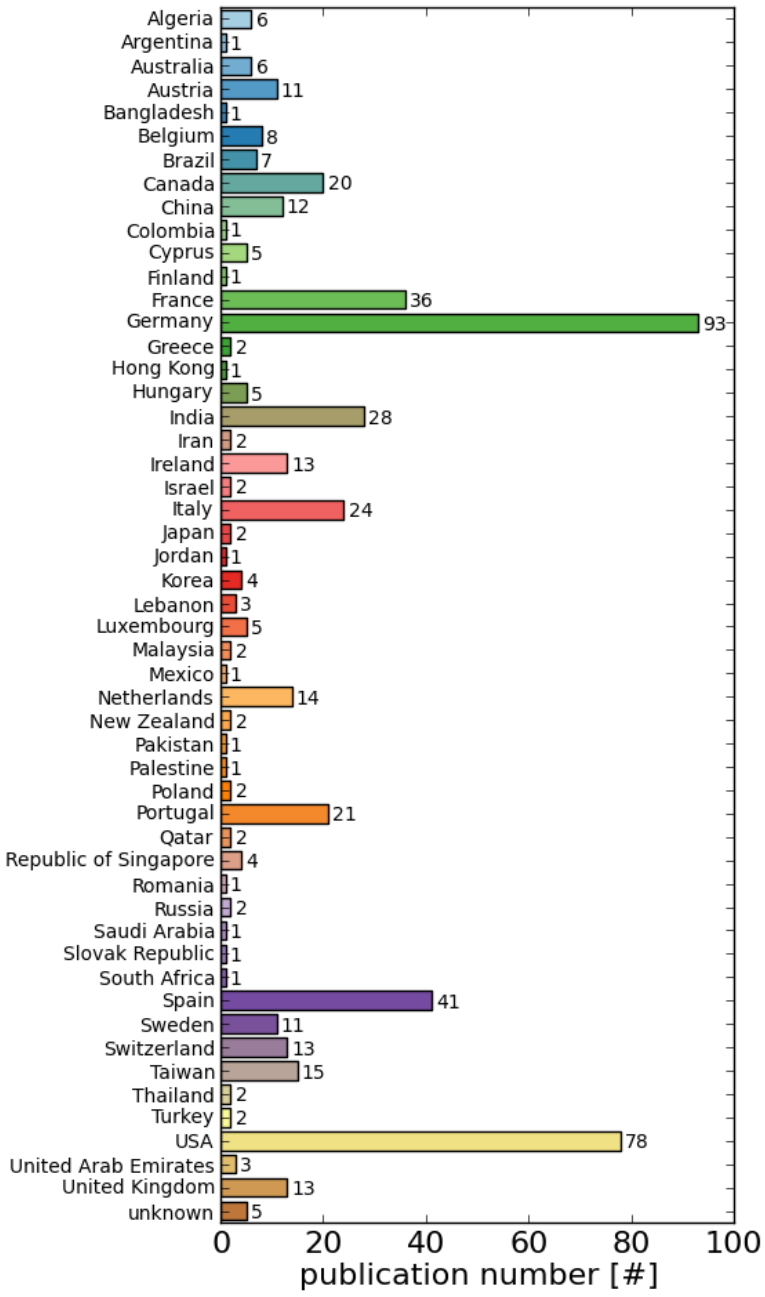

Figure 1-10: The number of publications per country

1.4.7 Content Statistics

Parallel to classification of the documents, the reference to the pdf-document was added to each entry. In a later step, "pdfminer" [29], an open source Python library for pdf parsing, was used to retrieve an xml-representation of the document as well as the text contained in it. At the current time, only some coarse values were extracted, including the number of pages, the number of word, and the number of characters. Figure 1-11 shows these values – the colouring is based on publication type as used in Figure 1-2. More in-depth evaluation of the contents may be the topic of later work.

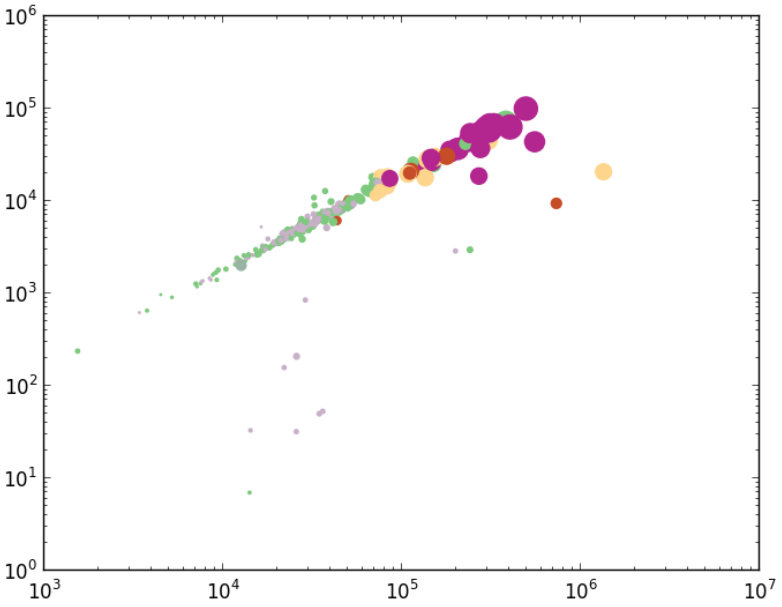

Figure 1-11: Word number (y) over character number (x) with log-scales for x- and y-axes; colour: BibTeX-type as in previous figures; size: page number

1.5 Summary and Outlook

Different views on a collection of documents which name the open source traffic simulation SUMO were presented. The document's bibliographic references, a manual classification performed on them and text processing were used to obtain the presented results. The evaluations show the progress of using SUMO, the major topics it is applied for, as well as other aspects.

Summarizing, it is possible to state that the number of publications increases and that no hints for a change in this development are visible. SUMO itself is accepted as a tool useful for research, indicated by the growing number of work that report its usage. The researches which cite or use SUMO come from all over the world, albeit European countries and the USA dominate. Research on vehicular communication is the major application topic to be found within the papers and, even if extrapolating the numbers, one should assume that it states at this position for the next time.

The shown distributions and developments over time are a coarse look at the data set only. It is assumed that the data set allows further evaluations, but most of such would be based on per-author evaluations what was not wanted for the presentation performed here.

As one could assume, the collection itself is of a high value, pointing to already done, classified work, or to parties working on certain topics. Nonetheless, the work on the collection – collecting, scanning, reading, and classifying the documents – gave probably more insight and surprises than a view on the complete collection presented here.

1.6 Acknowledgements

The author wants to thank all authors of the evaluated documents, and especially those who informed the SUMO development team about their work. Further thanks go to Alexandra Eßl for her work on the collection.

The work would be not possible without free search engines and library portals mentioned earlier. Also, acknowledgements go to the developers of the open source tools used in this research, namely JabRef, Python, matplotlib, and pdfminer.

1.7 References

[1] Krajzewicz, D., Erdmann, J., Behrisch, M., Bieker, L.: Recent *Development and Applications of SUMO - Simulation of Urban MObility.* In: International Journal On Advances in Systems and Measurements, 5 (3&4), pp.128-138. ISSN 1942-261x (2012)

[2] DLR and contributors: SUMO homepage. http://sumo.sourceforge.net/ (2013)

[3] Joerer, S., Sommer, C., Dressler, F.: *Towards Reproducibility and Comparability of IVC Simulation Studies--A Literature Survey.* In: Communications Magazine, IEEE, 50 (10), pp.82-88. ISSN 0163-6804 (2012)

[4] Smith, L. C.: Citation analysis. Library trends, 30. Jg., Nr. 1, pp. 83-106 (1981)

[5] Google. Google web search engine. http://www.google.com; last visited on 08.04.2013.

[6] Google. Google Scholar alerts. http://googlesystem.blogspot.de/2010/05/email-alerts-for-google-scholar.html, last visited on 08.04.2013.

[7] Springer. Springer Link. http://link.springer.com, last visited on 08.04.2013.

[8] IEEE. IEEE Xplore. http://ieeexplore.ieee.org/, last visited on 08.04.2013.

[9] ScienceDirect. ScienceDirect. http://www.sciencedirect.com/, last visited on 08.04.2013.

[10] ACM. ACM Digital Library. http://dl.acm.org/, last visited on 08.04.2013.

[11] BibTeX.org. http://www.bibtex.org/, last visited on 08.04.2013.

[12] Wikipedia. BibTeX description. http://en.wikipedia.org/wiki/BibTeX, last visited on 08.04.2013.

[13] Google. Google Scholar. http://scholar.google.com/, last visited on 08.04.2013.

[14] JabRef development team. JabRef web site. http://jabref.sourceforge.net/; last visited on 08.04.2013

[15] Härri, J., Fiore, M., Filali, F., Bonnet, C., Chiasserini, C., Casetti, C.: *A realistic mobility simulator for vehicular ad hoc networks.* Research report RR-05-150, EURECOM (2005)

[16] Kotushevski, G., Hawick, K.: *A review of traffic simulation software.* In: Conf. Information and Mathematical Sciences, New Zealand (2009)

[17] Sommer, C., Dressler, F.: *Progressing toward realistic mobility models in VANET simulations.* In: Communications Magazine, IEEE, 46, pp. 132-137 (2008)

[18] Karnadi, F., Mo, Z., Lan, K.-c.: *Rapid generation of realistic mobility models for VANET.* In: Wireless Communications and Networking Conference, pp. 2506-2511 (2007)

[19] Queck, T., Schünemann, B., Radusch, I., Meinel, C.: *Realistic simulation of v2x communication scenarios.* In: Proceedings of Asia-Pacific Services Computing Conference, 2008. APSCC 08. IEEE, 1623-1627 (2008)

[20] Rondinone, M., Maneros, J., Krajzewicz, D., Bauza, R., Cataldi, P., Hrizi, F., Gozalvez, J., Kumar, V., Röckl, M., Lin, L., Lazaro, O., Leguay, J., Härri, J., Vaz, S., Lopez, Y., Sepulcre, M., Wetterwald, M., Blokpoel, R., Cartolano, F.: *ITETRIS: a modular simulation platform for the large scale evaluation of cooperative* ITS applications. In: Simulation Modelling Practice and Theory (2013)

[21] Python Software Foundation: Python Programming Language – Official Website. http://www.python.org/, last visited on 08.04.2013.

[22] Hunter, J., Dale, D., Firing, E., Droettboom, M. and the matplotlib development team: matplotlib web site. http://matplotlib.org/, last visited on 08.04.2013.

[23] Wikipedia: "Stemming". http://en.wikipedia.org/wiki/Porter_stemmer, last visited on 08.04.2013.

[24] Porter, M.F.: An algorithm for suffix stripping. In: Program, 14(3), S. 130-137 (1980)

[25] Feinberg, J.: WordleTM. http://www.wordle.net/, last visited on 08.04.2013.

[26] Gupta, V., Yoo, D.: Python Porter stemming implementation. https://hkn.eecs.berkeley.edu/~dyoo/python/py_lovins/, last visited on 08.04.2013.

[27] Cottingham, D. N., Davies, J. J., Beresford, A. R.: Congestion-Aware Vehicular Traffic Routing Using WiFi Hotspots, In: Proc. Communications Innovation Institute Workshop (2005)

[28] Karnadi, F.; Mo, Z., Lan, K.-c. MOVE: A MObility model generator for VEhicular network (2005)

[29] Shinyama, Y.: pdfminer web site. http://www.unixuser.org/~euske/python/pdfminer/, last visited on 08.04.2013.