Hagen CHRAPARY, Wolfgang DALITZ & Wolfram SPERBER, Berlin

Software Knowledge Management und swMATH

Mathematische Software ist heute ein weitverbreitetes Werkzeug in der Forschung, aber auch in der mathematischen Bildung. Die Anzahl mathematischer Softwareprodukte wächst, nicht zuletzt durch die Anforderungen der Anwender, stark. Anders als für mathematische Publikationen sind für das Management mathematischer Software viele Fragen offen und werden noch diskutiert, etwa die Standardisierung von Software Zitationen. Der Aufbau einer effizienten Infrastruktur für Software ist eine notwendige Voraussetzung für die Überprüfung von Forschungsergebnissen, die mittels Software erzielt worden sind und wichtig für die Entscheidung, ob eine bestimmte Software zur Lösung eines Problems verwendet werden soll.

Das swMATH[1] Portal gibt einen weitgehend vollständigen Überblick über die existierende mathematische Software. Es ist ein Dienst entstanden, der automatisiert die im Web vorhandenen Informationen zu einer Software identifiziert, auswertet und verfügbar macht. Der Dienst soll insbesondere durch die Verknüpfung mit anderen Quellen, etwa dem Internet Archive, persistente Informationen über ein Software Produkt und dessen Versionen liefern.

Suche nach mathematischer Software

Wenn man im Web nach mathematischer Software sucht, geschieht das häufig mittels einer anderen universellen Suchmaschine. Eine Anfrage nach "mathematical software" liefert etwa bei Google über 25 Millionen Treffer, was eine gewisse Relevanz zum Ausdruck bringt. Allerdings zwingt es dazu, entweder die Trefferliste abzuarbeiten oder die Suchanfrage zu spezifizieren. Einer der ersten Treffer der Anfrage "mathematical software education" listet die Wikipedia Seite "List of educational software"[2]. In der Rubrik "Mathematics" finden sich derzeit Verweise auf weniger als 20 Softwareprodukte und vier weitere Wikipedia Listen. Die Liste ist sehr heterogen und umfasst universelle Softwareprodukte (Maple, Mathematica, Matlab), die sowohl für Bildungs- wie für Forschungszwecke eingesetzt werden können. Weiter gibt es eine Reihe von geometrischen Softwareprodukten mit überwiegend didaktischen Bezügen. Die verlinkten Wikipedia Listen zu mathematischer Software umfassen jeweils nur eine kleine Anzahl von Softwareprodukten, werden manuell gepflegt und unterscheiden sich in Inhalt und Form. Dazu kommt, dass sich die englischen und die entsprechenden deutschen Seiten substanziell unterscheiden. Die deutschsprachigen Seiten von Wikipedia sind hierbei in der Regel nicht so umfassend.

Die Informationen in den Wikipedia Listen für Software sind für den Nutzer oft sehr nützlich, etwa die Strukturierung der List_of interactive_geometry_software[3] nach verschiedenen Kriterien wie 2D oder 3D oder die zugrundeliegenden Beweistechniken. Aber auch hier gilt, dass diese Listen meist nicht vollständig sind. Da sich Entwicklung und Nutzung mathematischer Software in den letzten Jahren stark intensiviert hat, liegt der Schluss nahe, dass sich die manuelle Erstellung und Pflege von Softwarelisten und -portalen als unzureichend erweist. Man muss andere organisatorische Modelle und automatisierte Werkzeuge entwickeln, um den Zugang zur wissenschaftlichen Software zu vereinfachen.

Eine mögliche Alternative stellen Software Repositorien dar, wie etwa das CRAN[4] für die Statistik. Diese Repositorien setzen auf einer einheitlichen Programmiersprache auf, für CRAN etwa R, und ermöglichen das Einstellen und die Verwaltung der in R entwickelten Softwarepakete. Dieser Ansatz zielt auf eine spezielle Nutzergruppe.

Publikationsbasierter Ansatz

Einen anderen Weg geht das swMATH Projekt. Mathematische Software wird immer häufiger in wissenschaftlichen Zeitschriften und Bücher zitiert, oft benutzte mathematische Softwareprodukte kommen auf Tausende von Zitationen. Das ist der Ausgangspunkt für den publikationsbasierten Ansatz.

Der publikationsbasierte Ansatz startet mit der automatisierten Identifizierung der Softwareprodukte in wissenschaftlichen Publikationen. Hierzu werden die Informationen aus der bibliografischen Datenbank zbMATH[5] ausgewertet, insbesondere Titel, Review oder Abstract und Zitationen. Die Datenbank zbMATH, umfasst heute über 4 Millionen bibliografische Einträge zur Mathematik, zu denen auch Reviews und Abstracts gehören. zbMATH ist eine der relevantesten Review und Abstract Sammlungen in der Mathematik. Weltweit über 7.000 aktive Reviewer tragen zur Aktualität bei. Die Einträge reichen durch die Integration des ‚Jahrbuch über die Fortschritte der Mathematik' bis ins Jahr 1868 zurück.

Obwohl Software auf sehr unterschiedliche Art zitiert wird und ein Standard für Software Zitationen noch fehlt, lassen sich mit heuristischen Methoden, wie z.B. die Suche nach bestimmten Textmustern, Softwareprodukte erstaunlich gut identifizieren. Die Informationen aus den Publikationen liefern ein Profil der Softwareprodukte, das auf die mathematischen Herkunft der Software, deren Nutzung (Anwendungsgebiete) und deren Akzeptanz (Anzahl der Zitationen in bestimmten Zeiträumen) verweist. Als Seiteneffekt erhält man auch die Softwareprodukte, die benachbart sind (related software), also die Software, die über die Literaturreferenzen verbunden sind. Für jedes

Softwareprodukt wird in swMATH eine eigene Webpage erzeugt, die über einen persistenten Identifier aufrufbar ist.

Websites der Software

Da die Zitationen oft nur aus dem Namen des Softwareproduktes bestehen, erhält man hieraus keine Informationen über die benutzten Versionen sowie über die organisatorischen, technischen und rechtlichen Details zur Nutzung. Das indirekte Wissen über die Softwareprodukte muss also ergänzt werden. Dazu wird mittels Internetsuche die Website der Software lokalisiert (die URL der Website selbst ist die erste wesentliche Zusatzinformation) und diese analysiert. Wesentliche Unterschiede zwischen Publikationen und Software bestehen darin, dass Software nicht auf den Code reduziert werden kann, sondern aus verschiedenen Objekten (Code, APIs, Dokumentationen, Installationshinweisen, Tutorials, etc.) besteht und dass in der Regel bestimmte Versionen einer Software angeboten werden. Daraus resultieren die große Heterogenität der Websites in Inhalt und Form, wodurch die inhaltliche Analyse anspruchsvoll wird. Bisher werden nur einfache heuristische Verfahren zur inhaltlichen Analyse eingesetzt.

Versionierung der Software

Die Gestaltung der Websites folgt der Entwicklung der Software, verändert sich also mit den Versionen. Um trotzdem persistente Informationen nicht nur für die Softwareprodukte sondern auch für die Versionen anzubieten, werden Informationen aus dem Internet Archive[6] benutzt.

Die URLs der Software Websites werden vom Internet Archive gescannt und dort dauerhaft archiviert. Im Moment beschränkt sich die archivierte Information aber meist nur auf die Homepage der Software. Durch eine Spezifizierung der Knoten der Homepage, die sich durch eine Strukturanalyse erreichen lässt, ist aber eine persistente und vollständige Archivierung der Software Websites möglich. Da die Software Citation Principles[7], die den Ausgangspunkt für die Entwicklung eines Zitationsstandards für Software bilden, die Versionierung der benutzten Software fordern, wird dann, bei entsprechenden technischen und rechtlichen Voraussetzungen der benutzten Daten, sogar die spätere Überprüfung von Rechenergebnissen möglich.

Status Quo

Derzeit (März 2018) weist swMATH ca. 21.000 Objekte (Softwareprodukte, Dienste, Benchmarks, Software Journale, Begleitbüchern, Manuals, etc.) nach, die (in ca. 160.000 Publikationen) ca. 280.000 mal zitiert werden. Für das Retrieval stehen eine einfach und eine erweiterte Suche bereit. Zusätzlich

steht eine Browsing Funktionalität zur Verfügung, die ein gezieltes Navigieren der vorhandenen Software Produkte nach Name, Keywords, Softwaretyp und der MSC-Klassifikation ermöglicht. Die Collection zu Education Software in swMATH enthält zur Zeit etwa 50 Softwareprodukte.

Ausblick

Der swMATH Ansatz bietet mehrere wesentliche Vorteile. Der erste liegt in der weitgehenden Automatisierung der Methoden zur Erstellung und Pflege der Software Informationen. Das lässt sich weiter ausbauen, wenn man etwa auf Entwicklungsplattformen wie github zurückgreift, die mittlerweile von vielen Softwareentwicklern genutzt werden. Zur von uns angestrebten vollständigen Erfassung aller relevanten Softwareprodukte kommen auch Repositorien wie z.B. zenodo[8] oder Software Heritage[9] in Betracht. Ein zweiter Vorteil besteht in der Vollständigkeit und Persistenz der Information über eine Software, dass also nicht nur Informationen über sie selbst sondern auch über deren Nutzung und Akzeptanz.

Interessant ist auch die Verallgemeinerbarkeit des zugrundeliegenden Ansatzes. Mit der publikationsbasierten Methode lassen sich prinzipiell auch für andere Forschungsfelder spezielle Informationsdienste aufbauen und deren Daten zugänglich machen, etwa bei mathematische Modellen.

Zu guter Letzt bekommen die Softwareentwickler mehr Sichtbarkeit und Anerkennung. Im Wissenschaftsbereich zählen bekanntermaßen in erster Linie Veröffentlichungen. Wir glauben, dass mit swMATH hier eine neue Plattform entstanden ist, die auch den Entwicklern von Software größere Resonanz und Geltung verschafft.

Literatur

[1] http://www.swmath.org

[2] https://en.wikipedia.org/wiki/List_of_educational_software

[3] https://en.wikipedia.org/wiki/List_of_interactive_geometry_software

[4] https://cran.r-project.org/

[5] https://zbmath.org/

[6] https://archive.org/

[7] https://www.force11.org/software-citation-principles

[8] https://zenodo.org

[9] https://www.softwareheritage.org/

The work for this article has been conducted within the Research Campus MODAL funded by the German Federal Ministry of Education and Research (BMBF grant number 05M14ZAM).