Recherchieren zu einem Wissenschaftlichen (Informatik-)Thema

Ulf Leser

Inhalt

- Übersicht
- Wissenschaftliche Publikationen
- Bewertung von Publikationen
- Ressourcen und Techniken
- Literatursammlung erstellen und managen

Wissenschaftliches Thema

- In (Informatik-)Seminaren werden Themen meist gestellt
  - Ein Problem, eine Methode, ein Vergleich
  - Meist Paper und / oder kurze Erklärung
- Granularität: Themen können sehr breit oder sehr eng sein
  - Breit: Übersichtsarbeiten, geringe technische Tiefe
  - Eng: Spezialarbeiten, hohe technische Tiefe
- Themen stehen meist nicht genau fest
  - Recherche ergibt neue Aspekte – Themenanpassung
  - Hier können persönliche Präferenzen eingebracht werden
- Anders: Recherchieren zur Themenfindung
  - Explorativ versus zielgerichtet

Ziel der Recherche

- Im Ergebnis der Recherche kennt man
  - Verschiedene Aspekte des Problems und deren Unterschiede
  - Die verschiedenen Ansätze zur Lösung
  - Die wichtigste Literatur zum Thema
  - Eigener Schwerpunkt und welche Literatur man verwenden wird
- Vollständigkeit
  - IdR ist es unmöglich, alle Literatur zum Thema zu sichten
  - Bewertung der Relevanz und Auswahl notwendig
  - Teilweise macht man das selbst (lesen), teilweise vertraut man anderen (Wikipedia, Buch, Übersichtsartikeln, Datenbanken, ...)
  - Wichtig ist gestufte Bewertung: Titel/Autoren, Zitate, Erscheinungsort, Abstract, querlesen ...
  - Literaturliste mit dem Betreuer besprechen

Allgemeines Vorgehen

- Vor der genauen Recherche muss Thema (intuitiv) verstanden sein
  - ... sonst kann man Relevanz nicht beurteilen
  - Um was geht es in dem Problem?
  - Für was wird eine Methode verwendet?
- Prozesssicht
  - Recherche verläuft iterativ
    - Man sucht, liest, sucht mit anderen Begriffen, liest, folgt Links, ...
  - Management des Prozesses wichtig
    - Verwaltung und Systematisierung des Gefundenen
  - Beim späteren Schreiben ergeben sich neue Aspekte
    - Ergänzende Recherchen
  - Ggf. auch Anpassung des Themas möglich oder notwendig

Eng gefasste Themen

- Manches Seminarthema besteht aus einem Paper
- Das heisst nicht, dass man nicht auch andere Paper suchen und lesen muss!
  - Andere Ansätze kennenlernen
  - Übersicht gewinnen
  - Eigenes Paper besser einschätzen (was ist besser / neu / ...)

Beispiel

- Breit gefasst: Multidimensionale Indexstrukturen
  - Etabliertes Thema; es gibt Bücher und Übersichtsartikel; es gibt 10000+ Originalarbeiten; es gibt andere Seminararbeiten dazu im Web; es gibt Wikipedia Artikel; es gibt kommerzielle Produkte; ...
  - Übersicht gefragt: Was wird indiziert, welche Suchen werden unterstützt, welche grossen Klassen von Ansätzen gibt es, was sind deren Vor- und Nachteile, wo wurden sie schon mal verglichen, ...
- Eng gefasst: kdb-Trees
  - Sehr spezielles Thema; vielleicht eine (kurze) Erwähnung in einem Buch; vielleicht 10-20 wirklich relevante Arbeiten, die man vor allem in Spezialdatenbanken findet; kaum Erwähnung im Web; ...
  - Details sind gefragt: Wie funktioniert Löschen, welche Komplexität haben alle Operationen, wo gibt es empirische Untersuchungen, sind die Abhängig von den Daten, gibt es aktuelle Weiterentwicklungen, ...
- Themenanpassung (MDI)
  - Punkte / Flächen/Körper? Exakte / Ähnlichkeitssuche? Memory / Disk?
  - Was interessiert Sie mehr?

Inhalt

- Übersicht
- Wissenschaftliche Publikationen
- Bewertung von Publikationen
- Ressourcen und Techniken
- Literatursammlung erstellen und managen

Arten von Publikationen

- Zentrale Unterscheidung: Peer-review
- Peer-reviewed: Journale, Konferenzen, Workshops
  - In anderen Fächern ist das anders!
  - Informatikspezifisch: Konferenzen fast wichtiger als Journale
- Nicht: Bücher, Buchkapitel, Technical Reports
  - Reports: Früher pro Institut, heute auch internationale, fachspezifische Repositorien (arXiv, corr)
- Nicht wissenschaftlich: Blogs, White Paper, Wikipedia
- Sonderfälle
  - Dissertationen, Diplom- und Seminararbeiten
  - Poster
  - Vortragsfolien, Vorlesungen, auch Video

Peer-Review

- „Gute“ oder „neue“ Wissenschaft nicht objektiv definierbar
- Entscheidung soll durch Community getroffen werden
- Peer-Review: Paper werden von 2-3 Expert_innen bewertet
  - Ablehnung, Annahme, Revision
  - Einschätzung der Relevanz, Neuheit, Qualität des Textes
  - Unterschiedliche Anforderungen je nach Erscheinungsort
    - Nature: Extrem hoch, insb. Neuheit und Allgemeinrelevanz
    - Kleine Workshops: Eher niedrig, oft sehr spezielle Ergebnisse
- Vielfältige Kritik
  - Expert_innen die keine sind; schlampiges Lesen; Unterdrückung von Konkurrenz (single-blind); Trend zu Modethemen und inkrementellen Ergebnissen; idR keine Ergebnisvalidierung; ...
- Aber bestes bekanntes Verfahren

WikiPedia

- Zur frühen Recherche sehr nützlich
- Erfahrungsgemäß (in Informatik) meistens korrekt, aber nur geringe Themenabdeckung oder Tiefe
- Als Referenz allgemein nicht akzeptiert; zum Finden von Referenzen schon

Übersichtsartikel

- Zu nahezu allen Themen erscheinen regelmäßig Übersichtsartikel (surveys, reviews)
  - Vor allem als Buchkapitel / in Journalen
- Unterschiedliche Qualität und Tiefe
- Gute Surveys sind ein Segen und die halbe Miete
- Meistens sehr lange und nützliche Literaturlisten
- Für praktische Themen sehr relevant: Empirische Übersichts- und Vergleichsarbeiten (Benchmarks)

Lexika

- Lange Zeit bedeutungslos
- In den letzten Jahren aber wieder etwas populärer
  - Encyclopedia of ... Database Systems (Springer): „Comprehensive reference to about 1,400 entries, covering key concepts and terms in the broad field of database systems.”
  - Synthesis Lectures on ... Data Management (Morgan Claypool)
  - ...

Englisch, Deutsch, ...

- Alles wichtige, aktuelle ist Englisch
- Sehr gute deutsche Lehrbücher als Einstieg

Inhalt

- Übersicht
- Wissenschaftliche Publikationen
- Bewertung von Publikationen
- Ressourcen und Techniken
- Literatursammlung erstellen und managen

Impact Factor, Zitationen

- Qualität ist nicht binär
  - Erhebliche Unterschiede zwischen peer-reviewed Papern
- Wichtiges Indiz: Erscheinungsort und Autor
  - Welches Journal, welche Konferenz
  - Was keinen eindeutigen Autor hat, zählt nicht
- Typische Qualitätsmaße
  - Impact Factor: Durchschnittliche Zahl Zitierungen eines Papers in diesem Journal nach 5 Jahren; Wertebereich 0-32
  - Kritik: Nicht alle Paper eines Journals gleich gut; Themen mit kleinere Communities haben weniger Zitate; Fächerunterschiede; ...
  - Alternative: Zitationszahlen pro Paper
    - Google Scholar, Microsoft Academic Search, CiteSeer
  - Früher: Verlag (heute nicht mehr)

Wem trauen?

- Zentral: Zahl der Zitate
  - Gewichtet nach Jahr – 10 pro Jahr ist ganz ordentlich
  - Mit etwas Exploration ein Gefühl für das Thema bekommen
    - Viele Paper mit vielen Zitierungen?
    - Spezialgebiet mit insgesamt wenig Zitierungen?
  - Bringt einen Bias: „The rich get richer“
- Auch wichtig: Erscheinungsort
  - Suche nach „Erschienen in“ und sehen, wie oft Paper typischerweise zitiert werden
  - Vorsicht vor Abkürzungen und Schreibweisen
- Auch wichtig: Autor
  - Verlangt viel Erfahrung, starker Bias gegen Newcomer

Spezifika eines Fachs

	Informatik	**Life Sciences**	**Humanities**
Wo wird publiziert	Konferenzen, Work-shops, Journals, TRs	Nur Journals	Bücher
Wie wird bewertet	Zitierungen, Ort	Impact Factor	Autor
Wo wird gesucht	Google Scholar (ResearchGate?)	World of Science	Bibliothek
Was ist „viel“	100 pro Jahr	1000 pro Jahr	10 pro Jahr
Autorenlisten	2-5	Oft >20	1
Struktur von Papern	Frei	Fest (Intro, Methods, Results, Diskussion)	Frei

Open Access

- Idee: Artikel lesen umsonst, Publizieren kostet Geld
  - Traditionell: Publizieren ist umsonst, Lesen kostet Geld
- Das ist unabhängig von Peer-Review
  - Die meisten Open Access Journale sind peer-reviewed
- Das ist unabhängig von kommerziellen Interessen
  - Die meisten Open Access Journale sind kommerziell orientiert
- Erheblicher politischer Druck zu Open Access
- In der Informatik gibt es bisher keine OA-Journale
  - Aber Autoren stellen traditionell alle Artikel frei zur Verfügung
  - Bedeutung von TR-Repositorien

Alter – Wann ist Literatur nicht mehr aktuell?

- So alt ist die Informatik nicht
  - Viele grundlegende Arbeiten aus den 60ziger – 80ziger
- Theoretische Resultate (Beweise) halten „ewig“
- Methodische Resultate halten lange
  - Geschäftsmodelle, Best Practise, Software Engineering, ...
- Empirische Resultate können schnell veralten
  - Besser nicht älter als maximal 10 Jahre
  - Hauptspeicher - andere Indexverfahren
  - Cachelines – Scan statt Index
  - Breitbandinternet – Berechnungen verteilen
  - ...

Inhalt

- Übersicht
- Wissenschaftliche Publikationen
- Bewertung von Publikationen
- Ressourcen und Techniken
- Literatursammlung erstellen und managen

Recherche-Datenbanken

- Google Scholar / Microsoft Academic Search / Citeseer
  - Fulltextindexierung wissenschaftlicher Veröffentlichungen
  - Ziemlich vollständig und aktuell
  - Wichtiges Feature: Zitierende Arbeiten
  - Voll automatisch; Viele Qualitätsmängel vorhanden (Precision)
  - Zugriff auf Volltexte über Links
    - Innerhalb der HU weit mehr Content als ausserhalb
  - Bibliographische Informationen oft unvollständig
- DBLP
  - Manuell gepflegt, Fokus auf DB+LP
  - Unvollständig, keine Volltexte, Links auf PDFs
  - Sehr gut bzgl. bibliographischer Daten

Weitere Datenbanken

- ResearchGate
  - Closed, eventuell mehr PDF, Zugriffsregeln unklar, Zitierungszahlen unzuverlässig, manuell kuriert (Autoren), unklare Scores für Autoren
- PubAnnotator, Mendely, ...
- ACM-DL, IEEE-Explore, Springer Link, ...

Bibliotheken

- Zur Recherche wenig nutzbringend
- Sehr gut als Arbeitsort
- Sehr wichtig zum Zugriff auf lizensierten Content
  - eBooks, online-subscriptions, ...

Recherche Techniken (GS)

- Suchfeature benutzen
  - Volltextsuche, Phrasen, Negation
  - Suche nach Autoren
  - Suche nach Erscheinungsort
  - Suche nach zitierenden Arbeiten
  - Suche nach Jahren
- Auch wichtig: Wichtige Referenzen in guten Papern

Gestufte Bewertung

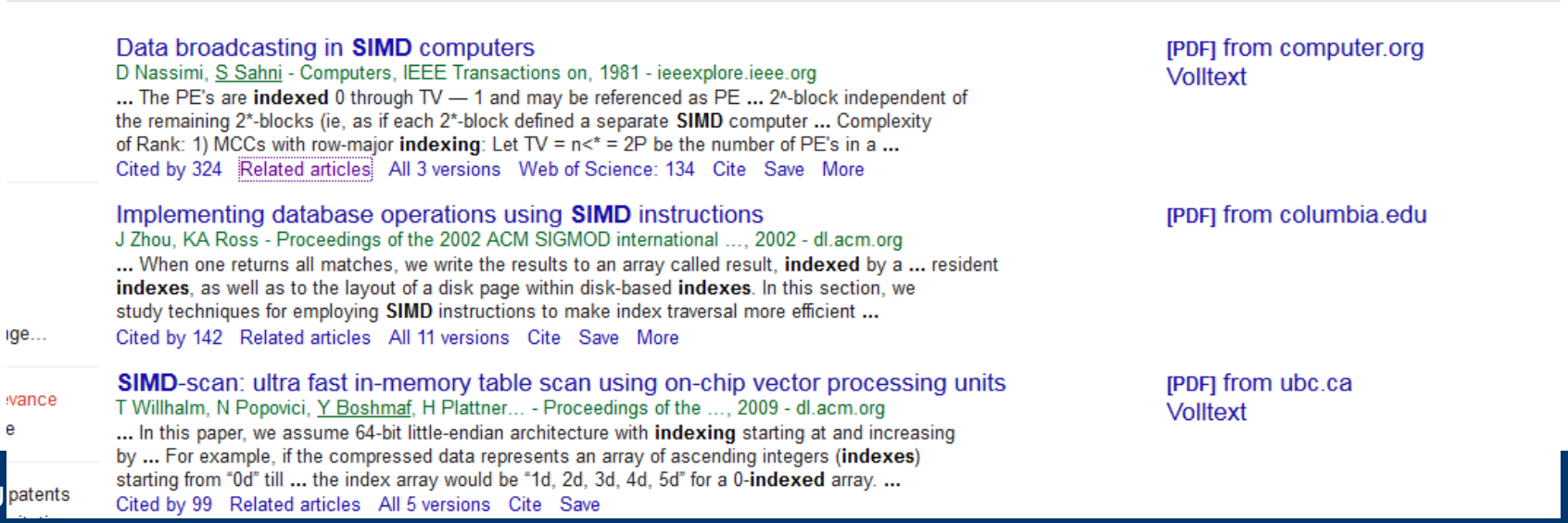
- Man muss vieles ansehen, aber das meisten nur sehr kurz
- Erste Einschätzung: Titel, Ort, Zitierungen (C)
  - Journal oder Konferenz; Bei Zahl Zitierungen das Jahr beachten
- Zweite Einschätzung: Abstract (C)
  - Ggf schnell aufhören, wenn erkennbar irrelevant
- Dritte Einschätzung: Paper querlesen (B)
  - Welches Problem, welcher Ansatz Daten, welche Ergebnisse?
  - Algorithmen, Beweise, Formeln überspringen
  - Paper mit Stichworten in Liste aufnehmen („Cite“ – kopieren)
- Vierte Einschätzung: Paper genau lesen (A)
  - Eigene Zusammenfassung in Liste aufnehmen

Iterativer Prozess

- Explorationsphase: Liste B-Paper erstellen
  - Keywortsuche, Surveys suchen, Zitationszahlen wichtig
  - Man findet die „seminal paper“ und erhält Übersicht
  - Paper nur schnell beurteilen
- Komplettierungsphase: In B-Paperliste die A-Paper finden
  - Gezielte weitere Suchen
    - Spezielle Keywords und Phrasen
    - Aktuelle Arbeiten, die A/B-Paper zitieren
    - Wichtige Arbeiten aus Referenzlisten von B/A-Papern
  - Führt zu weiteren A und B Papern (ständig bewerten)
- Verwendungsphase
  - Kann zu gezielten weiteren Recherchen führen
  - Aber nicht verlieren! |A|=10 ist viel; ggf. Betreuer fragen

Beispiel

- Autorensuche
- Erscheinungsjahr
- Zitierende Arbeiten
- Versionen
- Bibliographische Informationen
- [Related Articles, Web of Science]



Data broadcasting in **SIMD** computers
D Nassimi, S Sahni - Computers, IEEE Transactions on, 1981 - ieeexplore.ieee.org
... The PE's are **indexed** 0 through TV — 1 and may be referenced as PE ... 2^-block independent of the remaining 2*-blocks (ie, as if each 2*-block defined a separate **SIMD** computer ... Complexity of Rank: 1) MCCs with row-major **indexing**: Let TV = n<* = 2P be the number of PE's in a ...
Cited by 324 Related articles All 3 versions Web of Science: 134 Cite Save More

[PDF] from computer.org
Volltext

Implementing database operations using **SIMD** instructions
J Zhou, KA Ross - Proceedings of the 2002 ACM SIGMOD international ..., 2002 - dl.acm.org
... When one returns all matches, we write the results to an array called result, **indexed** by a ... resident **indexes**, as well as to the layout of a disk page within disk-based **indexes**. In this section, we study techniques for employing **SIMD** instructions to make index traversal more efficient ...
Cited by 142 Related articles All 11 versions Cite Save More

[PDF] from columbia.edu

SIMD-scan: ultra fast in-memory table scan using on-chip vector processing units
T Willhalm, N Popovici, Y Boshmaf, H Plattner... - Proceedings of the ..., 2009 - dl.acm.org
... In this paper, we assume 64-bit little-endian architecture with **indexing** starting at and increasing by ... For example, if the compressed data represents an array of ascending integers (**indexes**) starting from "0d" till ... the index array would be "1d, 2d, 3d, 4d, 5d" for a 0-**indexed** array. ...
Cited by 99 Related articles All 5 versions Cite Save

[PDF] from ubc.ca
Volltext

Inhalt

- Übersicht
- Wissenschaftliche Publikationen
- Bewertung von Publikationen
- Ressourcen und Techniken
- Literaturliste managen

Aufbereiten einer Textsammlung

- Notizen und Zusammenfassungen zu allen A und B Papern machen und verwalten
- Bibliographische Informationen gleich in hoher Qualität sammeln
- Strukturieren: Nach Problemvarianten, nach verwendeten Datensätzen, nach Architektur, nach Denkschule, ...
- Individuelle Schlagworthierarchie sinnvoll
  - Und notwendig bei Dissertationen
- Spezielle Software benutzen

Exzerpte erstellen

- Eigene Worte verwenden
- Texte drucken und markieren, dann zusammenfassen
- Systematik erstellen (Farben) – Kernaussage, Kritikpunkt, experimentelle Ergebnisse, ...
- A Paper ausführlich beschreiben, B Paper nur kurz
- Gute Zusammenfassungen kann man ggf. in die Seminararbeit übernehmen
  - Meist sind die aber zu kurz

Zum Zitieren

- Alle: Autoren, Titel, Jahr
- Konferenzen/Workshops: Name der Konferenz, Ort
- Journal: Ausgabe (volume, issue), Seitenzahl
- Report (auch Dissertation, DA ...): Institution, Nummer

- Möglich: DOI
- Nicht in Referenzliste: URL
- Nützlich zur Suche: Abstract
- Nützlich zum Nachsehen: PDF

Management von Literatur

- Spezielle Software zum
  - Datenbank für Literatur (Bibliographie, Link, Verschlagwortung, ...)
  - Ggf mit PDF Verwaltung und Volltextindexen
  - Ggf. automatische Extraktion bibliographischer Daten aus Webseiten
  - Formatierung von Referenzen in Texten nach Vorlagen
    - Word Plug-In, BibRef
  - Sortieren, suchen, filtern
  - Zugriff auf Datenbanken (PubMed)
- Endnote (Campuslizenz), Bibliographix, Citavi, jab-ref, ...
- Online-Systeme: Mendely, zetero, ...

Live-Beispiel

- Suchen
- Zugriff PubMed
- Word-Formatierung
- Referenzstile
- Formatierter Export
- BibTex-Export