Datenbank Spektrum (2017) 17:1–4
DOI 10.1007/s13222-017-0251-z

EDITORIAL

Editorial

Andreas Henrich[1] · Daniel Blank[1] · Dirk Lewandowski[2] · Theo Härder[3]

1 Schwerpunktthema: Innovative vertikale Suchlösungen: Konzeption, Umsetzung und Einsatz

Die populären Suchdienste für das Web dominieren die allgemeine Wahrnehmung von Suchmaschinen. Dabei gibt es ein breites Spektrum unterschiedlicher Suchsysteme und viele davon nutzen wir auch regelmäßig. Dies gilt beispielsweise für die lokale Suche nach Dateien auf unserem Rechner, die Suche in Online-Shops, die Suche in Bibliothekskatalogen oder die spezifische Suchfunktion für den Webauftritt eines Unternehmens oder einer Organisation – oft als Site-Search bezeichnet.

Eine allgemein anerkannte Definition von „vertikalen Suchlösungen“ existiert dabei wohl nicht. Als eine der ersten Verwendungen des Begriffs wird bisweilen der Blogpost „The Vortals Are Coming! The Vortals Are Coming!“ von Danny Sullivan aus dem April 2000 in *Search Engine Watch* genannt. Dabei addressiert er Suchfunktionen für sogenannte Vortals (Vertical Portals). Bezieht man die Definition primär auf die Suche im Web, so ergibt sich eine klare Kontrastierung zu horizontalen Suchlösungen, die versuchen, das Web in seiner ganzen Breite durchsuchbar zu machen. Demgegenüber beschränken sich vertikale Suchlösungen beispielsweise auf eine fachliche Domäne oder auch auf bestimmte Medientypen. Diese Beschränkung führt dazu, dass die Suche im Idealfall tiefer und semantisch reicher gestaltet werden kann, indem beispielsweise auf entsprechendes Domänenwissen zurückgegriffen wird.

Um die Spezifika vertikaler Suchlösungen besser einschätzen zu können, ist es hilfreich, sich die verschiedenen Aufgabenbereiche einer Suchmaschine grob zu verdeutlichen:

- Festlegung und Erfassung der zu indexierenden Dokumente
- Aufbereitung und Anreicherung der Dokumente zur Indexierung
- Indexierung und Suche (in der Regel mit invertierten Listen)
- Bereitstellung einer Nutzerschnittstelle zur Eingabe der Suchanfragen und zur Präsentation der Ergebnisse

Bei der *Festlegung und Erfassung der zu indexierenden Dokumente* ist zunächst die Auswahl der Datenquellen wichtig. Durch die entsprechende Auswahl erfolgt hier natürlich einerseits eine Beschränkung, aber eben auch eine Qualitätssicherung. Dabei spielt die inhaltliche Qualität eine große Rolle, aber auch die technische Zugreifbarkeit und die Verfügbarkeit von Metadaten ist zu berücksichtigen. Hinzu kommen Fragen zur Aktualisierung des Bestandes und ggf. zu Zugriffsrechten, die geprüft werden müssen. Zu erwähnen ist auch die von vielen vertikalen Suchlösungen erbrachte Integrationsleistung, da sie eine übergreifende

Andreas Henrich
andreas.henrich@uni-bamberg.de

Daniel Blank
daniel.blank@uni-bamberg.de

Dirk Lewandowski
dirk.lewandowski@haw-hamburg.de

✉ Theo Härder
haerder@cs.uni-kl.de

[1] Fakultät WIAI, Medieninformatik, Otto-Friedrich-Universität Bamberg, An der Weberei 5, 96047 Bamberg, Deutschland

[2] Fakultät Design, Medien und Information, Department Information, Hochschule für Angewandte Wissenschaften Hamburg, Finkenau 35, 22081 Hamburg, Deutschland

[3] AG Datenbanken und Informationssysteme, TU Kaiserslautern, 67663 Kaiserslautern, Deutschland

Suche über zum Teil recht heterogenen Datenbeständen einer Domäne erlauben. Dazu sind oft auch spezielle Adapter oder Konverter zu entwickeln, um die Bestände sinnvoll an die Suchlösung anbinden zu können.

Gerade die Einschränkung auf eine Domäne ermöglicht auch die *Aufbereitung und Anreicherung der Dokumente zur Indexierung*. Als einfaches Beispiel seien medizinische Texte genannt, die sich durch ein entsprechendes Fachvokabular und eine intensive Nutzung von Abkürzungen auszeichnen. Hier kann man versuchen, über ein Fachlexikon eine Anreicherung bzw. Normierung der indexierten Begriffe vorzunehmen.

Interessant ist bei den Systemen, die in den einzelnen Beiträgen dieses Schwerpunktheftes vorgestellt werden, dass für die eigentliche *Indexierung* und zur *Suche* jeweils Standardbibliotheken verwendet werden. Die Nutzung der Programmbibliothek Apache Lucene[1] für den Retrieval-Kern ist dabei weit verbreitet. Oft kommen auch die auf Apache Lucene basierenden Bibliotheken bzw. Plattformen Solr[2] oder Elasticsearch[3] zum Einsatz. Dies macht letztlich zwei Dinge deutlich: Zum einen sind die in den Bibliotheken implementierten Retrieval-Modelle offensichtlich so leistungsfähig und anpassbar, dass Speziallösungen keinen substantiellen Mehrwert versprechen. Hieran knüpft unmittelbar die zweite Schlussfolgerung an. Die Effektivität einer Suchlösung – also letztlich die Fähigkeit, vage Informationsbedürfnisse zu befriedigen – wird von den vier in der obigen Auflistung genannten Teilen gemeinsam beeinflusst. Alle Komponenten müssen passend umgesetzt sein, um eine effektive Lösung zu realisieren. Dabei sind Anpassungen und Optimierungen an der Dokumentenerfassung und -aufbereitung ebenso zwingend wie eine entsprechende Gestaltung der Nutzerschnittstelle. Aufwände, die in diese Systemkomponenten fließen, sind offensichtlich zielführender als der Versuch, den Retrieval-Kern über eine Optimierung der Parameter hinaus anzupassen.

Als letzter Teil der Suchlösung verbleibt damit die *Nutzerschnittstelle zur Eingabe der Suchanfragen und zur Präsentation der Ergebnisse*. Hier bietet es sich unter anderem an, die gezielte Suche in einzelnen Feldern der Metadaten, die häufig in domänenspezifischen Schemata definiert sind, zu unterstützen. Auch fachbezogene Suchtermvorschläge und andere Unterstützungsformen bei der Anfrageformulierung finden sich häufig. Analog ist eine entsprechende Aufbereitung der Ergebnisse auf der SERP (Search Engine Result Page) sinnvoll. Ein Ansatz könnte die Kategorisierung der Ergebnisse gemäß einem fachbezogenen Klassifikationsschema sein. Auch Verlinkungen zu weiterführenden Informationen (z. B. Webseiten der Autoren) sind denkbar.

Während die bisherigen Überlegungen sehr stark den Aspekt einer domänenspezifischen Suche in den Vordergrund gestellt haben, kommen weitere interessante Kriterien für eine mögliche Klassifikation vertikaler Suchlösungen hinzu:

- adressierte Domäne(n)
- betrachtete Medientyp(en)
- Art der Suchergebnisse
- gewählte Architektur

Die Medientypen ergeben sich dabei zum Teil aus der Domäne. Bei einer Suchlösung für Kulturschätze ist beispielsweise eine Bildsuche naheliegend. Damit treten neben den üblichen (Text-)Dokumenten als Suchergebnisse auch andere Medienobjekte auf. In anderen Suchlösungen sind die erwarteten Ergebnisse zum Beispiel Produkte (Shopsuche), Personen (Expertensuche) oder Antworten bzw. Fakten (Question Answering). Im Hinblick auf die Architektur sind auf einer groben Ebene Metasuchmaschinen und eigenständige Suchlösungen zu unterscheiden. Auf der detaillierteren Ebene stellt sich dann z. B. die Frage, wie auf die zu indexierenden Dokumente zugegriffen wird (Crawling, Harvesting, ...) und ob ein integriertes Ergebnisranking oder einzelne Teilrankings geliefert werden.

Ein Problem haben dabei letztlich alle vertikalen Suchlösungen: Sie werden oft mit den populären horizontalen Suchmaschinen für das Web verglichen. Dabei wird schnell übersehen, dass diese Suchmaschinen – neben den finanziellen Möglichkeiten – stark von einigen Charakteristika des Web profitieren. Chris Crawford, Managing Director für Accenture's internal IT, hat hierzu in einem Blog 2012 die bedeutendsten Faktoren hervorgehoben: Die horizontalen Suchlösungen profitieren von der Verlinkung im Web, von der schieren Größe des Web, die dazu führt, dass es für die meisten Anfragen eine große Zahl potentiell relevanter Seiten gibt, und von der Tatsache, dass viele Betreiber von Webauftritten viel Aufwand in die sogenannte Suchmaschinenoptimierung stecken. Die Nutzer der Suchmaschinen sind dadurch verwöhnt und erwarten auch von vertikalen Suchlösungen schnelle Antwortzeiten, eine hohe Aktualität und relevante Ergebnisse zu (fast) allen Anfragen. Mit dieser Erwartungshaltung kämpfen die Anbieter vertikaler Suchlösungen. Hier gilt es durch die Optimierung auf die fokussierte Domäne entsprechende Vorteile zu erzielen. Dazu kann auch beitragen, dass Inhalte gefunden werden können, die von horizontalen Suchmaschinen aufgrund von rechtlichen oder institutionellen Vorgaben nicht indexiert werden können.

Im vorliegenden Schwerpunktheft haben Sie die Möglichkeit, an neun Beispielen die Chancen und Herausforde-

[1] http://lucene.apache.org/, letzter Abruf 04.01.2017.

[2] http://lucene.apache.org/solr/, letzter Abruf 04.01.2017.

[3] http://www.elastic.co/de/products/elasticsearch, letzter Abruf 04.01.2017.

rungen bei der Entwicklung und dem Betrieb von vertikalen Suchlösungen zu identifizieren.

Der erste Beitrag behandelt die Metasuchmaschine BASE (Bielefeld Academic Search Engine) als Ansatz zur Indexierung wissenschaftlicher Metadaten. Daran schließt sich ein Beitrag zur Optimierung der Onsite-Suche bei otto.de als vertikale Suchlösung im E-Commerce an.

Einen Block zur domänenspezifischen Suche in Medizin, Lebenswissenschaften und Psychologie bilden die Aufsätze zu SABIO-RK, LIVIVO und PubPsych. Anschließend folgt ein Block mit Suchmaschinen und Informationsdiensten, die im weitesten Sinne den Digital Humanities zuzurechnen sind. Dabei handelt es sich um die Suchmaschine zum europäischen Kulturportal Europeana, einen spezialisierten Fachinformationsdienst für Darstellende Kunst sowie die Bildsuche im bayerischen Kulturportal bavarikon.

Ein Beitrag zu den spezifischen Chancen und Herausforderungen bei Suchmaschinen für Kinder rundet das Themenheft ab.

2 Community-Beiträge in diesem Heft

Die Rubrik „Datenbankgruppen vorgestellt“ enthält den Beitrag *Das Fachgebiet „Informationssysteme“ am Hasso-Plattner-Institut* von Felix Naumann und Ralf Krestel. Der Artikel skizziert die Ergebnisse wichtiger Forschungsprojekte der letzten Jahre und stellt die aktuellen Forschungsthemen vor. Weiterhin gibt er einen Überblick über die Lehre am Fachgebiet.

In der Rubrik „Kurz erklärt“ erscheint *Machine Learning Meets Databases* von Stephan Günnemann (TU München). Der Beitrag führt kurz in das momentane Hype-Gebiet *Machine Learning* (ML) ein und diskutiert sein Zusammenspiel mit anderen Bereichen wie *Data Mining* und *Datenbanken*. Weiterhin gibt er einen Überblick über spezielle ML-Funktionalität, die in jüngerer Zeit bereits in bestimmte Datenbanksysteme und große, verteilte Datensysteme integriert wurde.

Erfreulich umfangreich ist wiederum die Rubrik „Dissertationen“. Sie enthält in diesem Heft 12 Kurzfassungen von Dissertationen aus der deutschsprachigen DBIS-Community.

Schließlich berichtet die Rubrik „Community“ unter *News* über weitere aktuelle Informationen, welche die DBIS-Gemeinde betreffen.

3 Künftige Schwerpunktthemen

3.1 Big Graph Data Management

A graph is an intuitive mathematical abstraction to capture how things are connected. In the past decade, the focal point in many data management applications has shifted from individual entities and aggregations thereof toward the connection between entities. Hence today, the graph abstraction is appealing as a natural data model foundation for an increasing range of use cases in interactive as well as analytical graph data management scenarios. Graph-specific use cases can be found in various domains, such as social network analysis, product recommendations, and knowledge graphs. Graph-oriented scenarios also emerge in more traditional enterprise scenarios, such as supply chain management or business process analysis. Therefore, the database community reacts to this newly sparked interest in graph data management with a vast number of projects in research as well as in industry.

Graph management use cases pose novel and unique challenges to data management systems. On the operational side, typical interactive queries involve transitive closure computation along paths. Common analytical measures, such as page rank and other vertex centrality measures are also significantly more complex than traditional group-by/aggregate queries. From a data structure perspective, the irregular and skewed structure of graphs makes it challenging to achieve a good distribution over non-uniform memory access or cluster nodes for efficient parallelization – particularly, if the graph is large and changing over time. Further challenges among others are declarative graph analytics abstractions for static as well as for dynamic graphs, graph-query-aware optimization strategies, topology indexing, temporal topology indexing, topology estimation, materialized view usage, and maintenance for graph analytical measures.

Graph data management is an exciting research field, now and for the years to come. This special issue aims at exhibiting our community’s current work in the field. We therefore welcome contributions from research and industry that provide original research on the problems mentioned above or that are generally related to big graph data management and processing. We also welcome case studies that showcase the challenges of graph management and graph query processing from a practical perspective, point out particular research questions, and potentially outline novel research directions.

We are looking for contributions from researchers and practitioners in the above described context, which may be submitted in German or in English.

Important dates:

- Deadline for submissions: February 1st, 2017
- Issue delivery: DASP-2-2017 (July 2017)

Paper format: 8–10 pages, double column (cf. the author guidelines at www.datenbank-spektrum.de).

Guest editors:
Hannes Voigt, TU Dresden
hannes.voigt@tu-dresden.de
Marcus Paradies, SAP
m.paradies@sap.com

3.2 Best Workshop Papers of BTW 2017

This special issue of the "Datenbank-Spektrum" is dedicated to the Best Papers of the Workshops running at the BTW 2017 at the University of Stuttgart. The selected Workshop contributions should be extended to match the format of regular DASP papers.

Paper format: 8–10 pages, double column

Selection of the Best Papers by the Workshop chairs and the guest editor: April 15th, 2017

Guest editor:
Theo Härder, University of Kaiserslautern
haerder@cs.uni-kl.de

Deadline for submissions: June 1st, 2017
Issue delivery: DASP-3-2017 (November 2017)

3.3 Data Processing in Industrie 4.0

The Data Processing paradigm undergoes several changes recently. In the vision of Industrie 4.0 assets become smart and more autonomous. This leads to new application areas for analytics and data processing in general. We invite submissions on original research as well as overview articles covering topics from the following non-exclusive list:

- Industrie 4.0 Reference Architectures
- Sensor Data Streaming
- Sensor Data Management
- Digital Twin Technology
- Analytics in Industrie 4.0
- Edge Analytics/Fog Computing
- Sensor Data Analytics
- Advanced Analytics

Expected size of the paper: 8–10 pages (double-column). Contributions either in German or in English are welcome.

Deadline for submissions: Oct. 1st, 2017
Issue delivery: DASP-1-2018 (March 2018)

Guest editors:
Bernhard Mitschang, Universität Stuttgart
Bernhard.Mitschang@ipvs.uni-stuttgart.de
NN.